Advances in Game-Based Learning

Christian Sebastian Loh
Yanyan Sheng
Dirk Ifenthaler *Editors*

Serious Games Analytics

Methodologies for Performance Measurement, Assessment, and Improvement

Springer

Advances in Game-Based Learning

More information about this series at http://www.springer.com/series/13094

Christian Sebastian Loh • Yanyan Sheng
Dirk Ifenthaler
Editors

Serious Games Analytics

Methodologies for Performance Measurement, Assessment, and Improvement

Springer

Editors
Christian Sebastian Loh
Virtual Environment Lab (V-Lab)
Southern Illinois University
Carbondale, IL, USA

Dirk Ifenthaler
University of Mannheim
Mannheim, Germany

Curtin University
Perth, Australia

Deakin University
Melbourne, Australia

Yanyan Sheng
Department of Counseling, Quantitative
Methods, and Special Education
Southern Illinois University
Carbondale, IL, USA

Advances in Game-Based Learning
ISBN 978-3-319-05833-7 ISBN 978-3-319-05834-4 (eBook)
DOI 10.1007/978-3-319-05834-4

Library of Congress Control Number: 2015940118

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

In 2013, Springer published the edited volume on *Game Analytics*: *Maximizing the Value of Player Data* (Seif El-Nasr, Drachen, & Canossa, 2013). On the surface, it would appear that game analytics is applicable to serious games also. However, this is not true because the motivation for game analytics is monetization (hence, maximizing monetary value of data), but the purpose of *Serious Games Analytics* is to measure the performance of play-learners for assessment and improvement.

Serious games is an emerging field where the games are supposed to be using sound learning theories and instructional design principles to maximize learning and training success. But why should stakeholders believe serious games to be effective, if they have no reference as to what actions performed in the serious games constitute newly acquired skills, abilities, or knowledge? Are players simply having a fun time, really learning something (that may or may not relate to said skills/abilities), or gaming the system (i.e., finding loopholes to fake that they are making progress)?

The purpose of this edited volume is to collect in one place how gameplay data in serious games may be turned into valuable analytics (or actionable intelligence) for performance measurement, assessment, and improvement, using existing or emerging empirical research methodologies from various fields, including: computer science, software engineering, educational data mining, educational sciences, statistics, and information visualization.

Besides being the companion book to *Game Analytics*: *Maximizing the Value of Player Data*, this volume is also the first book in the *Advances in Game-Based Learning* (AGBL) series (Ifenthaler, Warren, & Eseryel; www.springer.com/series/13094)—both by Springer. Despite what some may feel to be a buzzword-loaded title, our intention in publishing *Serious Games Analytics* is three folds:

(a) To identify with the growing serious games industry
(b) To recognize the existing market need for actionable insights and analytics
(c) To present, in one place, advanced research related to serious games and analytics from both academia and the industrial sectors

It should be clear that the book points to a clear and present need for serious games analytics, and that researchers and industry leaders are already taking active parts in working out the issues surrounding serious games analytics. A total of 67 authors put their thoughts and efforts behind these chapters, describing problems faced and solutions found, as well as highlighting issues currently discussed and debated within the serious games communities.

The 19 chapters in this book represent the first step in defining what serious games analytics are—at least, for this point in time, and what they can become in the near future. The chapters in this edited volume are divided into six parts:

- In *Part I*, *Foundations of Serious Games Analytics*: the two chapters review the history and the rise of serious games as training/learning and policy-forming tools, discuss the movement towards analytics, and differentiate among game analytics, learning analytics, and serious games analytics. A meta-analysis of serious games data collection methods reveals not only the trends but also the lack of standardized and better-validated methods for research in serious games analytics.
- In *Part II*, *Measurement of Data in Serious Games Analytics*: the four chapters examine the design issues of serious games. Instead of gameplay design, serious games are more concerned with the design of in situ interaction data collection (via telemetry or *Information Trails*), and the design of analysis to yield actionable insights. The many areas of discussion include the recommendation for in situ data collection, the types and quality of interaction data (log files, online database, psychophysiological data), and innovative methodologies (e.g., data mining, statistical/machine learning, similarity measures, pattern recognitions) to obtain analytics and insights for performance improvement.
- In *Part III*, *Visualizations of Data for Serious Games Analytics*: the two chapters discuss the importance of data visualizations and their applications in serious games analytics. More than just pretty graphics, visualization of information should become a pertinent feature in serious games because it helps communicate to stakeholders the analytics and insights obtained from the in situ user-generated interaction data.
- In *Part IV*, *Serious Games Analytics for Medical Learning*: market forecast informs us that the next wave of serious games applications would be in the fields of medical learning and mobile applications. The three chapters in this section examine the applications of serious games for medical use—e.g., medical education, rehabilitation, and patient care. Serious games researchers would do well to take note of this upcoming, but largely unexplored area of research.
- In *Part V*, *Serious Games Analytics for Learning and Education*: the four chapters in this section reflect the current trends of “assessment” in educative serious games. Although the Evidence-Centered Design (ECD) framework has its origin in the measurement and testing industry, it has since been applied to stealth assessment for game-based learning, psychometric testing, and serious game design.
- In *Part VI*, *Serious Games Analytics Design Showcases*: we have included several showcases of serious games research projects with innovative designs

and/or interesting applications. They include: psychological profiles generation, replay analysis in game design, startle reflex in affective computing, and gameplay assessment through pattern matching.

We hope the chapters included in this volume will serve as launch pads or blueprints for future research and development projects and provide the serious games industry with the empirical evidence it has been seeking. Serious games publishers, developers, researchers, and consumers need to come together to dialog and create the foundation for *serious games analytics* research for future collaboration and to further advance the field.

Without the assistance of experts—in the field of serious games and game-based learning (two related, but different, groups), and their contributions in writing the chapters, this book project would not exist, at all. We must also thank the series editor of AGBL and Springer for believing in this book project. Last but not least, we would like to thank all the reviewers for their tremendous help in providing constructive and editorial comments for the chapters. We would like to extend a big handshake (virtually) and "Thank You" to all of those who have made this book journey a pleasant one. Kudos to all and we now know who to contact for our next book project!

Sebastian would like to thank his family for the mental supports. Working with Springer (and Dirk) on this first edited book project has been a true blessing because they have made the process a breeze. He would like to thank Yanyan for the many lunch meetings and discussions sessions about the book chapters. In addition, he would like to extend special appreciations to Dirk for being a friend when he came calling in 2010 and for providing him and his wife with fond memories of the Black Forest.

Carbondale, IL, USA Christian Sebastian Loh
Carbondale, IL, USA Yanyan Sheng
Mannheim, Germany Dirk Ifenthaler

References

Ifenthaler, D., Warren, S., & Eseryel, D. (Eds.). (2013). *Advances in game-based learning* series. New York: Springer.

Seif El-Nasr, M., Drachen, A., & Canossa, A. (Eds.). (2013). *Game analytics*: *Maximizing the value of player data*. London: Springer.

Contents

Part I Foundations of Serious Games Analytics

Part II Measurement of Data in Serious Games Analytics

Part III Visualizations of Data for Serious Games Analytics

Contributors

About the Editors

Dirk Ifenthaler (University of Mannheim, Germany; dirk@ifenthaler.info) is Professor of Instructional Design and Technology at the University of Mannheim, Germany, Adjunct Professor at Deakin University, Australia, Affiliate Research Scholar at the University of Oklahoma, USA, as well as an Adjunct Associate Professor at Curtin University, Australia. His previous roles include Professor and Director, Centre for Research in Digital Learning at Deakin University, Australia, Manager of Applied Research and Learning Analytics at Open Universities Australia, and Professor for Applied Teaching and Learning Research at the University of Potsdam, Germany. Dirk was a 2012 Fulbright Scholar-in-Residence at the Jeannine Rainbolt College of Education, at the University of Oklahoma, USA. Professor Ifenthaler's research focuses on the intersection of cognitive psychology, educational technology, learning science, data analytics, and computer science. He developed automated and computer-based methodologies for the assessment, analysis, and feedback of graphical and natural language representations, as well as simulation and game environments for teacher education. His research outcomes include numerous coauthored books, book series, book chapters, journal articles, and international conference papers, as well as successful grant funding in Australia, Germany, and the United States—see Dirk's website for a full list of scholarly outcomes at www.ifenthaler.info. Professor Ifenthaler is the Editor-in-Chief of the Springer journal *Technology*, *Knowledge and Learning* (www.springer.com/10758). Dirk is the Past-President for the AECT Design and Development Division, 2013–2015 Chair for the AERA Special Interest Group Technology, Instruction, Cognition and Learning and Co-Program Chair for the international conference on Cognition and Exploratory Learning in the Digital Age (CELDA).

Christian Sebastian Loh (Virtual Environment Lab, Southern Illinois University; csloh@siu.edu) is Director of Virtual Environment Lab (V-Lab) and Associate Professor of Learning Systems Design & Technology at Southern Illinois University.

He has designed serious games for research, pioneered new metrics for expert-novice differentiation, developed tools for in situ data collection (Information Trails) and real-time visualization of performance (Performance Tracing Report Assistant, PeTRA). His research interests are about the measurement, assessment, and improvement of human performance with serious games and other virtual environments. Dr. Loh was a recipient of the Defense University Research Instrumentation Program (DURIP) grant awarded by the U.S. Army Research Office (ARO) and a Past-President of the Multimedia Production Division of the Association for Educational Communications and Technology (AECT). He serves on the panel of judges at the annual Serious Games Showcase & Challenge (SGS&C) competition hosted by Interservice/Industry Training, Simulation and Education Conference (I/ITSEC), and the editorial board of a number of international journals, including *Technology, Knowledge and Learning (TKL)*, and *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS).*

Yanyan Sheng (Southern Illinois University; ysheng@siu.edu) is Associate Professor of Quantitative Methods at Southern Illinois University. Her primary research interests focus on modeling dichotomous responses in educational and psychological measurement using advanced modern statistics, and specifically, on developing and applying complex yet efficient Bayesian hierarchical item response models. She developed complex Bayesian multidimensional models using Gibbs sampling with various latent dimensional structures and has written and published computer programs for these models. Dr. Sheng received the 2006 Distinguished Dissertation Award by American Psychological Association (APA) Division 5 (Measurement, Evaluation & Statistics), and the 2014 Outstanding Scholar Award by College of Education and Human Services at Southern Illinois University Carbondale. She has published over 30 peer-reviewed journal articles and is currently the Associate Editor of *International Journal of Quantitative Research in Education*.

About the Authors

Marc T.P. Adam (The University of Newcastle, Australia; marc.adam@newcastle.edu.au) is a Lecturer in IT at the University of Newcastle, Australia. He received a Diploma in Computer Science from the University of Applied Sciences Würzburg, Germany, and a Ph.D. in Economics of Information Systems from the Karlsruhe Institute of Technology, Germany. His work has been published in *Journal of Management Information Systems (JMIS)*, *Economics Letters (ECOLET)*, *International Journal of Electronic Commerce (IJEC)*, *Electronic Markets (EM)*, *Journal of Neuroscience, Psychology, and Economics (JNPE)*, *International Conference on Information Systems (ICIS)*, *European Conference on Information Systems (ECIS)*, and others. His research interests include business information systems, consumer behavior, design of information systems, electronic markets, gamification, human–computer interaction, NeuroIS, serious games, and user experience.

Vincent Aleven (Carnegie Mellon University, aleven@cs.cmu.edu), an Associate Professor in Carnegie Mellon's Human–Computer Interaction Institute (HCII), has over 20 years of experience in research and development of advanced learning technologies grounded in cognitive science theory and theory of self-regulated learning. He has built intelligent tutoring systems (e.g., mathtutor.web.cmu.edu) and educational games, as well as authoring tools to facilitate their creation (e.g., ctat.pact.cs.cmu.edu). He has used these systems as platforms to investigate student learning in classrooms and online courses. He led a team that created games for science learning for children in K-3 in collaboration with Carnegie Mellon's Entertainment Technology Center. Aleven and his students have done research on educational games for policy reasoning and intercultural negotiation skills. He is a cofounder of Carnegie Learning, Inc., a Pittsburgh-based company that markets Cognitive Tutor™ math courses. He is a member of the Executive Committee of the Pittsburgh Science of Learning Center (LearnLab). He is or has been PI on seven major research grants and co-PI on 10 others. He has over 180 publications to his name. He is Co-Editor-in-Chief of the *International Journal of Artificial Intelligence in Education.*

Laura K. Allen (Arizona State University, laurakallen@asu.edu) is a graduate student in the Department of Psychology and the Learning Sciences Institute at Arizona State University. Her academic background includes a B.A. in English Literature and Foreign Languages (2010) and an M.A. in Cognitive Psychology (2014). She is currently pursuing a doctoral degree in the area of Cognitive Science. Her current research investigates the cognitive individual differences that contribute to proficiency in reading comprehension and writing, as well as the application of certain cognitive principles to educational practice. Ms. Allen is particularly interested in examining how natural language processing techniques can enhance the automated detection of these processes.

Katherine August (ETH Zurich, kit.august@gmail.com) Kit's current research at Stevens Institute of Technology focuses upon the design, development, and evaluation of innovations for smart cloud connected systems and devices to improve human–robot collaborations and assistive devices. Kit was a Visiting Research Fellow at Queens University of Belfast (2011–2012), Whitaker International Scholar in Biomedical Engineering and Academic Guest at INI University of Zurich/ETH Zurich and SMS/IRIS ETH Zurich (2009–2011), Research Assistant at NJIT Department of Biomedical Engineering (2003–2009). Kit was a Member of the Technical Staff at Bell Laboratories (1991–2002) in New Service Concepts and Advanced Communications Technologies; she is a BAM Technology in the Arts Award Winner, and earned 19 US and 50 international patents for technology innovations: signal processing for intelligent wireless; intelligent adaptive sensors and antenna arrays; Virtual Assistant Coach; Language Tutor; Intelligent Search Algorithms; Steganography; etc. Kit has a BFA from Parsons The New School for Design, an MSCS-MIS from Marist College, and a Ph.D. in Biomedical Engineering from New Jersey Institute of Technology. She is a Senior Member of IEEE, Chair of IEEE Young Professionals; Member of: EMBS, Robotics and Automation, Communications, Computational Intelligence, WIE, EdSoc; Sigma Xi.

Jodi Asbell-Clarke (EdGE at TERC, jodi_asbell-clarke@terc.edu) is the director of the Educational Gaming Environments Group (EdGE) at TERC in Cambridge, MA, USA. TERC is a not-for-profit research and development organization that has been focusing on innovative, technology-based math and science education for nearly 50 years. As the director of EdGE, Jodi leads a team of game designers, educators, and researchers who are designing and studying social digital games as learning environments that span home, school, and community. Jodi's background includes M.A. in Math, an M.Sc. in Astrophysics and a Ph.D. in Education. She started her career at IBM working on the first 25 missions of the space shuttle as an onboard software verification analyst. After teaching at the laboratory school at the University of Illinois, she joined TERC and has spent the past 20+ years developing science education programs and researching new ways to promote science learning. In 2009, she cofounded EdGE at TERC.

Ryan S. Baker (Columbia University, baker2@exchange.tc.columbia.edu) is Associate Professor of Cognitive Studies at Teachers College, Columbia University. He earned his Ph.D. in Human–Computer Interaction from Carnegie Mellon University. Dr. Baker served as the first technical director of the Pittsburgh Science of Learning Center DataShop, the largest public repository for data on the interaction between learners and educational software. He is currently serving as the founding president of the International Educational Data Mining Society, and as associate editor of the *Journal of Educational Data Mining*. His research combines educational data mining and quantitative field observation methods to better understand how students respond to educational software, and how these responses impact their learning. He studies these issues within intelligent tutors, simulations, multiuser virtual environments, and educational games.

Malcolm Bauer (Educational Testing Service, mbauer@ets.org) is a Managing Senior Scientist in ETS's Cognitive and Learning Sciences Group. Malcolm's research focuses on characterizing patterns of learning in STEM disciplines and applying those patterns in the development of innovative assessment and learning systems. He currently has an IES grant for creating and assessing learning trajectories for algebra for use in formative assessment. He has a Ph.D. in cognitive psychology and a B.S.E. in Electrical Engineering and Computer Science, both from Princeton University. Before ETS, he was a systems scientist at the Human–Computer Interaction Institute at Carnegie Mellon University.

Maria Bertling (Educational Testing Service, mbertling@ets.org) is a senior research assistant at the Educational Testing Service (ETS). Her current interests lie in the area of educational measurement and using technology-based assessments in conjunction with advanced psychometric models to measure student learning. She is particularly interested in evidence identification and accumulation from complex interactive environments (e.g., games, conversation-based assessments, auto-tutor systems) that inform student and measurement models and enable teachers and educators shape their instructional practices and decisions. She holds a Master's degree in cognitive psychology.

Karen Blackmore (The University of Newcastle, Australia; karen.blackmore@newcastle.edu.au) is a Lecturer in Information Technology at the School of Design, Communication and IT, the University of Newcastle, Australia. She received her BIT (Spatial Science) With Distinction and Ph.D. (2007) from Charles Sturt University, Australia. Dr. Blackmore is a spatial scientist with research expertise in the modeling and simulation of complex social and environmental systems. Her research interests cover the use of agent-based models for simulation of socio-spatial interactions and the use of simulation and games for serious purposes. Her research is cross-disciplinary and empirical in nature and extends to exploration of the ways that humans engage and interact with models and simulations. Before joining the University of Newcastle, Dr. Blackmore was a Research Fellow in the Department of Environment and Geography at Macquarie University, Australia and a Lecturer in the School of Information Technology, Computing and Mathematics at Charles Sturt University. She is a member of the Association for Computing Machinery (ACM).

Benjamin Cawrse (SimIS Inc., benjamin.cawrse@simisinc.com) Benjamin Cawrse is a Senior Software Engineer for SimIS, Inc. He obtained his Master's degree in Computer Science and minor's in Modeling and Simulation and Computer Engineering from Old Dominion University, Norfolk, VA. He has been a core member of the AIMS development team and was responsible for the software development in this project. He contributed to conference papers and journal contributions featuring the project.

Gregory K.W.K Chung (National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California; greg@ucla.edu) is Assistant Director for Research Innovation at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). His current work at CRESST involves developing math learning games for middle school students and computer-based assessments to measure problem-solving and content knowledge in military and engineering domains. He has experience in developing Web-based assessment tools for diagnostic and embedded assessment purposes using Bayesian networks, domain ontologies, and other advanced computational tools.

Benjamin A. Clegg (Colorado State University; benjamin.clegg@colostate.edu) is an Associate Professor of Psychology in the Department of Psychology at Colorado State University. His central research focus is on skill acquisition, including training with systems like microworlds and simulators. His past research has been funded under the DARPA AugCog program, and as a member of a MURI grant from the Army Research Office. He has been a coauthor on a total of over 100 publications, reports, and conference and scientific meeting presentations. Clegg is currently a consulting editor for the *American Journal of Psychology*, and has also served as an expert reviewer for 22 different major psychology journals, and for the National Science Foundation and Army Research Office.

David J. Cornforth (The University of Newcastle, Australia; david.cornforth@newcastle.edu.au) holds the B.Sc. in Electrical and Electronic Engineering from Nottingham Trent University, UK (1982), and the Ph.D. in Computer Science from the University of Nottingham, UK (1994). He has been an educator and researcher at Charles Sturt University, the University of New South Wales, and currently at the University of Newcastle. He has also been a research scientist at the Commonwealth Scientific and Industrial Research Organization (CSIRO), Newcastle, Australia. He has published over 120 peer-reviewed journal and conference articles. His research interests include data mining, artificial life, health information systems, pattern recognition, artificial intelligence, multi-agent simulation, and optimization. He is convener of the Applied Informatics Research Group, University of Newcastle.

Kristen E. DiCerbo (Pearson, kristen.dicerbo@pearson.com) is a Principal Research Scientist at Pearson. Her research program centers on the use of interactive technologies in learning and assessment. Dr. DiCerbo investigates how to use evidence from learner activity in games and simulations to understand what learners know and can do. She has also engaged with teachers to understand how to best communicate information about student performance to inform instructional decisions. She has a Master's degree and Ph.D. in Educational Psychology from Arizona State University. Prior to coming to Pearson she worked as a school psychologist and a researcher with the Cisco Networking Academy.

James E. Folkestad (Colorado State University; james.folkestad@colostate.edu) is an Associate Professor in the School of Education. He teaches a graduate level course on technology-enhanced learning (TEL) and a course on technology-enhanced instruction to preservice teachers. He is also the director of the Colorado State University (CSU) Analytics for Learning and Teaching (ALT) center. His research is focused on TEL applications and the use of analytics to understand teaching and learning.

Theodore W. Frick (Indiana University, frick@indiana.edu) is a Professor Emeritus in the Department of Instructional Systems Technology, School of Education, Indiana University Bloomington. His current research interests include improvement of teaching and learning, simulations and games for understanding educational systems, and predicting patterns in educational systems.

Gerald R. Gendron (SimIS Inc., gerald.gendron@simisinc.com) Gerald "Jay" Gendron is the Director of Programs and Chief Data Scientist at SimIS, Inc. He has written on multiple perspectives of learning and sociological impacts in technology-centric training systems. He has also led data analysis studies to identify trends in modeling and simulation as well as education. Serving as the SimIS Chief Data Scientist, Jay has conducted research and analysis on special interest items for the Joint Staff. He is especially interested in and writes about the impact generation has on training and business interpersonal dynamics. He is also an award-winning writer and speaker who has presented at international conferences and symposia. He has worked in a variety of defense assignments from acquisition to analysis and from live training to Advanced Distributed Learning (ADL) within the special operations community.

Judy Goldsmith (University of Kentucky; goldsmit@cs.uky.edu) has degrees from Princeton University and the University of Wisconsin-Madison. She post-doc'ed at Dartmouth College and Boston University, taught at the University of Manitoba, and joined the Computer Science Department of the University of Kentucky in 1993. She is a full professor. Goldsmith's work has received The First Annual IJCAI-JAIR Best Paper Prize, Honorable Mention, 2003 for her 1998 paper, and honors for student's papers at FLAIRS '12 and CGAMES '13. Her research focuses on decision-making, including decision-making under uncertainty, computational social choice, preference handling, and computational complexity. In 1998, Goldsmith received an AAAS mentoring award. She has helped organize and/or participated in several conferences for women in STEM disciplines. In spring 2013, she received a Provost's Award for Outstanding, the Outstanding Teaching Award from her department, and the UK College of Engineering Henry Mason Lutes Award for Excellence in Engineering Education.

Emmanuel Guardiola (Conservatoire National des Arts et Métiers, France; emmanuelguardiola@gmail.com) is game director and expert in game design methodology. He contributes to more than 30 major titles in the game industry, for independent studios and publishers. He has a Ph.D. in Computer Science and drives research on player psychological profiling through gameplay at the CNAM computer science laboratory (Paris). He was one of the key creators of the game design training at the French Graduate School on Games and Interactive Media (ENJMIN).

Erik Harpstead (Carnegie Mellon University, eharpste@cs.cmu.edu) is a Ph.D. student in the Human–Computer Interaction Institute at Carnegie Mellon University and a fellow in the Program for Interdisciplinary Research (PIER) program. His research interest is the design and evaluation of educational games. To this end, he works on developing novel tools and techniques for evaluating educational games in terms of their stated goals. He is also interested in better understanding ways in which game designers and learning scientists can collaboratively evaluate games in service of redesign.

Geoff Hookham (The University of Newcastle, Australia; geoffrey.hookham@newcastle.edu.au) completed his Bachelor's degree in Information Technology in 2007, a Graduate Certificate in Digital Media in 2010. His interest in the capacity of games to be entertainment and learning mediums has led Geoff to pursue a Ph.D. at the University of Newcastle studying engagement in serious games. From 2008 to the present, Geoff has tutored and taught in the areas of animation and computer games, as well as interactive and narrative design.

G. Tanner Jackson (Educational Testing Service, gtjackson@ets.org) is a research scientist in the Research and Development division at Educational Testing Service in Princeton, NJ. Tanner received a Ph.D. degree in cognitive psychology in 2007 and an M.S. degree in cognitive psychology in 2004—both from the University of Memphis. Tanner received a B.A. degree in psychology from Rhodes College in 2001. After completing a Postdoctoral Fellowship at the University of Memphis (2008–2011), Tanner continued his research as an Assistant Research Professor within the Learning Sciences Institute at Arizona State University (2011–2013). Tanner's current work at ETS focuses on innovative assessments and student process data. His main efforts involve the development and evaluation of conversation-based formative assessments (through ETS strategic initiatives) and game-based assessments (working in collaboration with GlassLab). Additionally, Tanner is interested in how users interact with complex systems and he leverages these environments to examine and interpret continuous and complex data streams, including user interactions across time within an adaptive assessment system.

Nathan Jacobs (University of Kentucky, jacobs@cs.uky.edu) graduated from the University of Missouri in 1999 with a B.S. in Computer Science and completed his Ph.D. in Computer Science at Washington University in St. Louis in 2010. He is currently an Assistant Professor of Computer Science at the University of Kentucky. His research area is computer vision, with a focus on algorithms for widely distributed cameras, object tracking, environmental monitoring, and surveillance.

Herbert F. Jelinek (Charles Sturt University, hjelinek@csu.edu.au) holds the B.Sc. (Hons.) in human genetics from the University of New South Wales, Australia (1984), Graduate Diploma in Neuroscience from the Australian National University (1986) and Ph.D. in medicine from the University of Sydney (1996). He is Clinical Associate Professor with the Australian School of Advanced Medicine, Macquarie University, and a member of the Centre for Research in Complex Systems, Charles Sturt University, Australia. Dr. Jelinek is currently visiting Associate Professor at Khalifa University of Science, Technology and Research, Abu Dhabi, UAE. He is a member of the IEEE Biomedical Engineering Society and the Australian Diabetes Association.

Yue Jia (Educational Testing Service, yjia@ets.org) is a senior Psychometrician in the Research and Development Division at Educational Testing Service (ETS) in Princeton, NJ. She is also the associate project director for Psychometrics and Research under the ETS contract of the National Assessment of Educational Progress (NAEP). She joined ETS in September 2006. She received her M.A. and Ph.D. in statistical science from Southern Methodist University in 2004 and in 2007, respectively.

Jina Kang (The University of Texas at Austin, jina.kang@austin.utexas.edu) is a doctoral student in the Learning Technologies Program at the University of Texas at Austin. She has been working as a teaching assistant for the classes related to design strategies and interactive multimedia production. Her research interests are in learning analytics, in which she is currently focusing on visualizing students' learning behavior in serious games and providing just-in-time feedback to help teachers track and understand the learning paths.

Chandan Karmakar (The University of Melbourne, karmakar@unimelb.edu.au) received his B.E. from Shahjalal University of Science and Technology, Bangladesh in 1999 and Ph.D. from the University of Melbourne, Australia in 2012. He is currently a postdoctoral research fellow at the University of Melbourne, Australia, where he is involved in research in the area of low cost healthcare devices, nonlinear signal processing, and pattern recognition in biomedical signals. As an early career researcher graduating in 2011, he has published over 55 refereed research papers in highly reputed journals and conferences. He has also received UniMelb early career researcher grant for 2013 to develop novel methods of heart rate variability analysis based on complex network theory. He is a regular reviewer in major international biomedical journals and features on technical program committee for several major international conferences. His current research interests include nonlinear signal processing, physiological modeling, biomedical instrumentation, and pattern recognition techniques.

Frances Kay-Lambkin (University of New South Wales, Australia; f.kaylambkin@unsw.edu.au). Over the past 10 years, Frances Kay-Lambkin has worked in a clinical research capacity with people experiencing psychotic disorders, depression, personality disorders, and alcohol/other drug use problems, with specific experience in the use of cognitive behavior therapy, motivational interviewing, and mindfulness-based stress reduction techniques among people with co-occurring mental health and alcohol/other drug problems. Her main research activity has been on the development and clinical trial of computer- and Internet-delivered treatments for people with co-occurring mental health and alcohol/other drug use problems. She has led several large-scale randomized controlled clinical trials of face-to-face, phone-based, and computerized psychological treatments, and translated these treatments into clinical practice. Associate Professor Kay-Lambkin has also developed tobacco-focused psychological treatments incorporating a multiple behavior change focus, and in clinical treatment trials evaluating the efficacy of such treatments among people with mental health problems. Her vision is to bring high quality, evidence-based treatment for multiple health problems to the point-of-care for people experiencing mental health and addictive disorders to ensure that the right person receives the right intervention at the right time.

Ahsan H. Khandoker (The University of Melbourne, ahsank@unimelb.edu.au) received the Doctor of Engineering degree in physiological engineering from the Muroran Institute of Technology, Muroran, Japan, in 2004. He is currently working as an Assistant Professor in the Department of Biomedical Engineering, Khalifa University, Abu Dhabi, UAE. He is also working as a Senior Research Fellow for Australian Research Council Research Networks on Intelligent Sensors, Sensor

Networks and Information Processing, University of Melbourne, Melbourne, Australia. He has published 38 peer reviewed journal articles and more than 75 conference papers the research field of physiological signal processing and modeling in fetal cardiac disorders, sleep disordered breathing, diabetic autonomic neuropathy, and human gait dysfunction, and is passionate about research helping clinicians to noninvasively diagnose diseases at early stage. He has also worked with several Australian Medical device manufacturing industries, as well as hospitals as a research consultant focusing on integration of technology in clinical settings.

Fengfeng Ke (Florida State University, fke@fsu.edu) is an associate professor of education in the Department of Educational Psychology and Learning Systems at the Florida State University where she works in the areas of game-based learning, virtual reality, computer-supported collaborative learning, and inclusive design of computer-assisted learning. She has published widely in the fields of innovative learning technologies and inclusive pedagogy for e-learning.

Kate Kenski (University of Arizona, kkenski@email.arizona.edu) is an Associate Professor of Communication and Government & Public Policy at the University of Arizona where she teaches political communication, public opinion, and research methods. She is a former editor of the *International Journal for Public Opinion Research* and former associate editor of *Public Opinion Quarterly*. Her book The Obama Victory: How Media, Money, and Message Shaped the 2008 Election (coauthored with Bruce W. Hardy and Kathleen Hall Jamieson; 2010, Oxford University Press) has won several awards including the 2011 ICA Outstanding Book Award, the 2012 NCA Diamond Anniversary Book Award, the 2012 NCA Political Communication Division Roderick P. Hart Outstanding Book Award, and The PROSE Award for 2010 Best Book in Government and Politics. Kenski is also coauthor of the book Capturing Campaign Dynamics: The National Annenberg Election Survey (2004, Oxford University Press). She has published research in journals such as the *American Behavioral Scientist*, *Communication Research*, *The International Journal of Public Opinion Research*, *The Journal of Applied Social Psychology*, *The Journal of Broadcasting & Electronic Media*, *Presidential Studies Quarterly*, and *Public Opinion Quarterly*. Her current research focuses on incivility in online forums and multimedia teaching strategies to mitigate cognitive biases.

Alexander Koenig (Sensory-Motor Systems Lab, ETH Zurich, Switzerland; alexander.c.koenig@gmx.de) holds an M.S. from Georgia Institute of Technology, Atlanta, USA in Electrical Engineering (2006), and a Ph.D. from ETH Zurich, Switzerland in Mechanical Engineering (2011). From 2011 to 2013, he was a Postdoctoral Research Fellow at the Motion Analysis Lab at Harvard University, Cambridge, USA. His research involves general principles of motor adaptation and motor learning, and the use of psychophysiological measurements in healthy subjects and neurological patients. Alexander is a serial entrepreneur and author of over 30 publications, patents, and book chapters on neurorehabilitation. He is currently a senior researcher at BMW Group Research and Technology, Munich, Germany, where he works on transferring psychophysiological methods from academia to industry for automatic driver state assessment.

Simone Kriglstein (Vienna University of Technology, Vienna; simone.kriglstein@igw.tuwien.ac.at) studied Computer Science at the Vienna University of Technology and graduated in 2005. Her diploma thesis focused on "Visual Perception and Interface Design." She received her doctorate degree from the University of Vienna in 2011. In her doctoral thesis, she described a development process for ontology visualizations based on the human-centered design approach. From 2005 to 2007 she worked as usability consultant/engineer and user interface designer. From 2007 until 2011 she was a research assistant and teaching staff at the University of Vienna, Faculty of Computer Science. Since 2011, she works for several projects at the University of Vienna (e.g., OCIS, playthenet) and the Vienna University of Technology (e.g., CVAST). From 2012 to 2014 she also worked as postdoctoral researcher at SBA Research. Her research interests are interface and interaction design, usability, information visualization, and games.

Jaejin Lee (The University of Texas at Austin, jaejinlee@utexas.edu) is a Ph.D. candidate in Learning Technologies program at the University of Texas at Austin. His research interests center on educational uses of educational games, multimedia development and 3D graphics in authentic learning environments, and emerging technologies. He has worked on various instructional design projects in higher education and is a GRA in the Office of Instructional Innovation in College of Education responsible for multimedia design and visualization laboratory in the college. He has participated in Alien Rescue Project over 5 years as a 3D modeler. Currently, he is working on his dissertation with a topic about the effect of 3D fantasy on academic achievement and game engagement in educational games.

Min Liu (The University of Texas at Austin, mliu@austin.utexas.edu) is a Professor of Learning Technologies at the University of Texas at Austin. Her teaching and research interests center on educational uses of new media and other emerging technologies, particularly the impact of such technologies on teaching and learning; and the design of new media-enriched interactive learning environments for learners at all age levels. She has published over 60 research articles in leading peer-reviewed educational technology journals, eight peer-reviewed book chapters, and presents regularly at national and international technology conferences. She also serves on a number of editorial boards for research journals in educational technology. Her current R&D projects include studying the design and effectiveness of immersive, rich media environments on learning and motivation; analytics in serious game environments; examining the affordances and constraints of using mobile technologies in teaching and learning; understanding MOOCs as an emerging online learning tool; and use of Web 2.0 tools to facilitate instruction.

Sa Liu (The University of Texas at Austin, liusa@utexas.edu) is a third-year doctoral student in the Learning Technologies Program at the University of Texas at Austin.

She is a board member of *Texas Education Review* journal and Associate Editor for Gaming & Education section. Her research interests include learning analytics for serious games, technology-promoted teacher development, and computer-supported language learning.

Kristine Lohr (University of Kentucky, kmlohr2@email.uky.edu) received her M.D. from the University of Rochester School of Medicine and Dentistry in 1975. She completed an internal medicine residency at Ohio State University Hospital in 1978 and a rheumatology fellowship at Duke University Medical Center in 1981. She served on the faculty at the Medical College of Wisconsin in Milwaukee (1981–1987) and the University of Tennessee Health Sciences Center (1987–2007) before joining the faculty at the University of Kentucky in 2007. At UTHSC, she served 13 years as Course Director for two medical school courses. Currently, she is Interim Chief of the Division of Rheumatology, Professor of Medicine, and Director of the Rheumatology Training Program at the University of Kentucky. Dr. Lohr is a past recipient of the American College of Rheumatology Research and Education Foundation Clinical Scholar Educator, and has served on several ACR committees. Current ACR responsibilities include the Audiovisual Aids Subcommittee, Annual Review Course, Annual Meeting Program Committee, and Committee on Workforce and Training. Currently, she serves as a member of the American Board of Internal Medicine Rheumatology Board, after serving on the ABIM Rheumatology Exam Writing Committee. Her current research interests include medical decision-making and patient safety.

Rosa Mikeal Martey (Colorado State University, rosa.martey@colostate.edu) brings a background in studying online activity using multi-methodological approaches, including survey research, computer log analyses, experiments, and interviews. Her research focuses on social interaction in games, game design, and game principles in learning. She currently serves as key personnel on the IARPA-funded Reynard project for which she was the lead designer and programmer of a multiplayer game in Second Life used as an experimental setting for data collection (SCRIBE). Other funded research includes a project that incorporates game design principles into university instruction.

Brian McKernan (University at Albany, SUNY, brian.mckernan@gmail.com) is postdoctoral associate at the Institute for Informatics, Logics, & Security Studies, University at Albany, SUNY. Brian received his Ph.D. in Sociology from the University at Albany, SUNY. His research adopts a cultural sociology framework to examine the roles of media and popular culture in civil society.

Christopher J. MacLellan (Carnegie Mellon University, cmaclell@ cs.cmu.edu) is a Ph.D. student in the Human–Computer Interaction Institute at Carnegie Mellon University and a fellow in the Program for Interdisciplinary Research (PIER) program. His work centers on the applying artificial intelligence and machine learning techniques to construct models of how humans perform open-ended tasks. Using this approach, he studies human learning, problem-solving, and design in the context of intelligent tutoring systems and educational technologies.

Danielle S. McNamara (Arizona State University, danielle.mcnamara@asu.edu) is a Professor in the Psychology Department at Arizona State University. She focuses on educational technologies and discovering new methods to improve students' ability to understand challenging text, learn new information, and convey their thoughts and ideas in writing. Her work integrates various approaches and methodologies including the development of game-based, intelligent tutoring systems (e.g., iSTART, Writing Pal), the development of natural language processing tools (e.g., iSTART, Writing Pal, Coh-Metrix, the Writing Assessment Tool), basic research to better understand cognitive and motivational processes involved in comprehension and writing, and the use of learning analytics across multiple contexts. She has published over 300 papers (see soletlab.com) and secured approximately 19 million in funding from federal agencies such as IES and NSF. More information about her research and access to her publications are available at soletlab.com.

Radu P. Mihail (Valdosta State University, r.p.mihail@valdosta.edu) graduated from Eastern Kentucky in 2009 with a B.S. in Computer Science and completed his Ph.D. in Computer Science at the University of Kentucky in Lexington, KY in 2014. He is currently an Assistant Professor of Computer Science at Valdosta State University. His research area is in computer vision, with a focus on medical image processing, outdoor image analysis, and teaching.

Geoffrey T. Miller (Eastern Virginia Medical School, millergt@evms.edu) Geoffrey Tobias Miller is an Assistant Professor, School of Health Sciences, and Director of Simulation, Research and Technology at the Sentara Center for Simulation and Immersive Learning at Eastern Virginia Medical School (EVMS) in Norfolk Virginia. Geoff joined EVMS in January of 2011, and is overseeing the expansion of simulation-based educational activities, curriculum development and educational outcomes and translational analysis, with an emphasis on the creation and improvement of operational and clinical competence assessment using advanced educational technology, modeling and simulation, specializing in immersive virtual environments, serious gaming and innovative educational technology development. Previously, Geoff was the Associate Director of Research and Curriculum Development for the Division of Prehospital and Emergency Healthcare at the Michael S. Gordon Center for Research in Medical Education (GCRME), University of Miami Miller School of Medicine.

Robert J. Mislevy (Educational Testing Service, rmislevy@ets.org) occupies the Frederic M. Lord Chair in Measurement and Statistics at Educational Testing Service. He is Professor Emeritus of Measurement, Statistics, and Evaluation at the University of Maryland. He earned his Ph.D. at the University of Chicago in 1981. Dr. Mislevy's research applies developments in statistics, technology, and cognitive science to practical problems in educational assessment. His work includes a multiple-imputation approach for integrating sampling and test-theory models in the National Assessment of Educational Progress (NAEP), an evidence-centered framework for assessment design, and Bayesian inference network scoring methods for simulation- and game-based assessments. Dr. Mislevy has won NCME's Award for Technical Contributions to Educational Measurement three times. He has received NCME's Award for Career Contributions, AERA's E. F. Lindquist Award for contributions to educational assessment, the International Language Testing Association's Messick Lecture Award, and AERA Division D's Robert L. Linn Distinguished Address Award. He is a member of the National Academy of Education and a past president of the Psychometric Society. His publications include Automated Scoring of Complex Performances in Computer Based Testing, Bayesian Networks in Educational Assessment, and the chapter "Cognitive Psychology and Educational Assessment" in Educational Measurement (Fourth Edition).

Brad A. Myers (Carnegie Melon University, bam@cs.cmu.edu) is a Professor in the Human–Computer Interaction Institute in the School of Computer Science at Carnegie Mellon University. He is an IEEE Fellow, ACM Fellow, winner of nine best paper type awards and three Most Influential Paper Awards. He is also a member of the CHI Academy, an honor bestowed on the principal leaders of the field. He is the principal investigator for the Natural Programming Project and the Pebbles Handheld Computer Project, and previously led the Amulet and Garnet projects. He is the author or editor of over 430 publications, and he has been on the editorial board of five journals. He has been a consultant on user interface design and implementation to over 75 companies, and regularly teaches courses on user interface design and software. Myers received a Ph.D. in Computer Science at the University of Toronto, and the M.S. and B.Sc. degrees from the Massachusetts Institute of Technology during which time he was a research intern at Xerox PARC. From 1980 until 1983, he worked at PERQ Systems Corporation. His research interests include user interfaces, programming environments, programming by example, and interaction techniques.

Rodney D. Myers (independent scholar, rod@webgrok.com) is an independent scholar who teaches courses in instructional design and technology. His research is broadly oriented towards exploring how to design and use emerging technologies to create meaningful and memorable learning experiences. His current research focuses on how online learning experiences—games and simulations in particular—can be designed so that they effectively promote learning while remaining engaging and motivating.

Stéphane Natkin (Conservatoire National des Arts et Métiers, France; stephane.natkin@cnam.fr) is chair professor at the Department of Computer Sciences of the Conservatoire National des Arts et Métiers (CNAM) in Paris. He is the Director of the Graduate School on Games and Interactive Media (ENJMIN), major French school delivering a Master in video games. At the Computer Research Laboratory CEDRIC of the CNAM he leads Interaction and Game department. He has worked during the last 20 years in the field of multimedia systems, video games, and critical computer system (safety and security), both from the research and the industrial point of view. He is the author of numerous computer science publications and communications to international congresses in these fields. He acts, as a scientific advisor, for France Telecom R&D for the research programs related to entertainments and games, distributed architecture and software engineering. From 1992 to 1995 he managed "La Galerie Natkin-Berta," an art gallery situated in the center of Paris which presented modern paintings, sculpture, and electronic art. He is the author of the book "Internet Security Protocols" DUNOD 2001 and "Computer Games and Media in the XXI century" Vuibert 2004; and "Video Games and Interactive Media, A Glimpse at New Digital Entertainment," AK Peters Ed, 2006.

Keith Nesbitt (The University of Newcastle, Australia; keith.nesbitt@newcastle.edu.au) completed his Bachelor's degree in Mathematics at Newcastle University in 1988 and his Masters in Computing in 1993. From 1989 to 1999, Keith worked on applied computer research for BHP Research investigating business applications of Virtual Reality and Intelligent Agents. His Ph.D. examined the design of multi-sensory displays and was completed at Sydney University in 2003. Outcomes from this work have received international recognition in the "Places and Spaces" exhibit and consequently exhibited at a number of international locations and reviewed in the prestigious journal *Science*. In 2007, he completed a postdoctoral year in Boston working at the New England Complex Systems Institute visualizing health-related data. He has extensive experience in the field of Human Interface Design as it relates to issues of Perception and Cognition in Computer Games. Keith currently works at the University of Newcastle teaching mainly in areas related to programming and game design and production. Despite his scientific background, Keith's interests also extend to more creative areas. He has 11 painting exhibitions to his credit as well as collaborations with musicians that have produced three CDs and a musical. You can find more about his art and science at www.knesbitt.com.

Marimuthu Palaniswami (The University of Melbourne, palani@unimelb.edu.au) received his M.E. from the Indian Institute of science, India, M.Eng.Sci. from the University of Melbourne and Ph.D. from the University of Newcastle, Australia. He served as a codirector of the Centre of Expertise on Networked Decision & Sensor Systems (2002–2005). Presently, he is running the largest funded ARC Research Network on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP, http://www.issnip.unimelb.edu.au/) program with $4.75 million ARC funding over 5 years. He has published more than 400 refereed papers (including 116 journals) and a huge proportion of them appeared in prestigious IEEE Journals and Conferences. He has won the University of Melbourne Knowledge Transfer Award in 2007 and 2008. He was given a Foreign Specialist Award by the Ministry of Education, Japan in recognition of his contributions to the field of Machine Learning. His research interests include SVMs, Sensors and Sensor Networks, Machine Learning, Neural Network, Pattern Recognition, Signal Processing, and Control. He is the codirector of the Centre of Expertise on Networked Decision & Sensor Systems. He holds several large Australian Research Council Discovery and Linkage grants with a successful industry outreach program.

Mathew G. Rhodes (Colorado State University, matthew.rhodes@colostate.edu) is an Associate Professor in the Department of Psychology at Colorado State University. He teaches graduate and undergraduate courses on cognition, with a focus on learning and memory. His research focuses on how people control memory so as to optimize learning.

Robert Riener (ETH Zurich, robert.riener@hest.ethz.ch) studied Mechanical Engineering at TU München, Germany, and the University of Maryland, USA. He received a Dr.-Ing. degree in Engineering from the TU München in 1997. After postdoctoral work from 1998 to 1999 at the Centro di Bioingegneria, Politecnico di Milano, he returned to TU München, where he completed his Habilitation in the field of Biomechatronics in 2003. In 2003, he became assistant professor at ETH Zurich and the University of Zurich, medical faculty ("double-professorship"); since 2010 he has been full professor for Sensory-Motor Systems, ETH Zurich. Since 2012, Riener belongs to the Department of Health Sciences and Technology. Riener has published more than 400 peer-reviewed journal and conference articles, 20 books, and book chapters and filed 20 patents. He has received 18 personal distinctions and awards including the Swiss Technology Award in 2006, the IEEE TNSRE Best Paper Award 2010, and the euRobotics Technology Transfer Awards 2011 and 2012. Riener's research focuses on the investigation of the sensory-motor actions in and interactions between humans and machines. This includes the design of novel user-cooperative robotic devices and virtual reality technologies applied to neurorehabilitation. Riener is the inventor and organizer of the Cybathlon 2016.

Daniel H. Robinson (Colorado State University, dan.robinson@colostate.edu) is Editor of Educational Psychology Review and Associate Editor of the *Journal of Educational Psychology*. He has also served as an editorial board member of nine refereed international journals including: *American Educational Research Journal*, *Contemporary Educational Psychology*, *Educational Technology, Research, & Development*, *Journal of Behavioral Education*, and the *Journal of Educational Psychology*. He has published over 100 articles, books, and book chapters, presented over 100 papers at research conferences, and taught over 100 college courses. His research interests include educational technology innovations that may facilitate learning, team-based approaches to learning, and examining trends in articles published in various educational journals and societies. Dr. Robinson was a Visiting Fulbright Scholar, Victoria University, Wellington, New Zealand.

Elizabeth Rowe (EdGE at TERC, elizabeth_rowe@terc.edu) is the Director of Research for the Educational Gaming Environments (EdGE) group at TERC, responsible for data collection, analysis, and interpretation for all EdGE projects. In her 14 years at TERC, Dr. Rowe has studied and developed innovative uses of technology in and out of school including several NSF-funded projects such as *Kids'*

Survey Network, *InspireData* software for K-12 students and the *Learning Science Online* study of 40 online science courses for teachers. Dr. Rowe has led formative and summative evaluations of several technology professional development programs. Prior to joining TERC, Dr. Rowe was a research analyst at the American Institutes for Research where she analyzed national survey data for the National Center for Education Statistics. She holds a Bachelor's degree in mathematics and a Ph.D. in human development and family studies.

Adrienne Shaw (Temple University, adrienne.shaw@temple.edu) is an assistant professor in the Department of Media Studies and Production at Temple University and a Media and Communications Ph.D. program faculty member. Her primary areas of interest are video games, gaming culture, the politics of representation, and qualitative audience research. Her forthcoming book is titled Gaming at the Edge: Sexuality and Gender at the Margins of Gamer Culture (University of Minnesota Press, 2015).

Valerie Shute (Florida State University, FL; vshute@fsu.edu) is the Mack & Effie Campbell Tyner Endowed Professor in Education in the Department of Educational Psychology and Learning Systems at Florida State University. Her current research involves using games with stealth assessment to support learning—of cognitive and noncognitive knowledge, skills, and dispositions. Her research has resulted in numerous grants, journal articles, books, chapters in edited books, a patent, and a couple of recent books (e.g., Shute & Ventura (2013), Measuring and supporting learning in games: Stealth assessment, The MIT Press; and Shute & Becker (Eds.) (2010), Innovative assessment for the 21st century: Supporting educational needs, Springer-Verlag).

Shamus P. Smith (The University of Newcastle, Australia; shamus.smith@ newcastle.edu.au) is a Senior Lecturer in Computer Science in the School of Electrical Engineering and Computer Science at the University of Newcastle (Australia). He received his B.Sc., B.Sc. (Hons), and Ph.D. in Computer Science from Massey University (New Zealand). Dr. Smith is a software engineer and has research expertise in virtual reality and human–computer interaction. His research interests include touch-based technologies (e.g., multi-touch tables and haptic devices), the reuse of gaming technology (e.g., eHealth applications and virtual training systems), and technology-enhanced learning (e.g., m-learning and smartphone apps). His research is interdisciplinary and empirical in nature and focuses on the use of advanced interfaces and the complex interactions that result when they are deployed with human users. Before joining the University of Newcastle in 2013, Dr. Smith was a Lecturer in Computer Science (2004–2013) in the School of Engineering and Computing Sciences, Durham University (UK) and a postdoctoral research associate (1998–2004) in the Department of Computer Science, the University of York (UK). He is a member of the Institute of Electrical and Electronics Engineers (IEEE) and the IEEE Computer Society and associate of the Australian Computer Society (ACS).

Erica L. Snow (Arizona State University, erica.l.snow@asu.edu) is a graduate student in the Department of Psychology and the Learning Sciences Institute at Arizona State University. Her academic background includes a Psychology B.S. (2007) and a Cognitive Psychology M.A. (2014). She is currently pursuing a doctoral degree in the area of Cognitive Science. Her current research explores the interplay of students' learning outcomes, learning behaviors, and individual differences within intelligent tutoring systems and educational games. Ms. Snow is particularly interested in how methodologies from artificial intelligence, educational data mining, and learning analytics can be applied to discover patterns in students' logged interactions with computer-based learning environments.

Shonté Stephenson (GlassLab Games, shonte.berkeley@gmail.com) earned her Ph.D. in Education with an emphasis in quantitative measurement from the University of California at Davis, and she has 2 years in postdoctorate training specializing in quantitative methods and evaluation from UC Berkeley. Her research primarily focused on developing viable modeling approaches for investigating differential item functioning (DIF) using IRT and factor analytic methods. Her particular interest centered on the development of a procedural framework for identifying test items that were susceptible to observed school context effects using multilevel modeling. As a visiting assessment fellow at GlassLab for the past year, Dr. Stephenson helped to produce innovative analyses of gameplay patterns from educational video games. Applying both exploratory analyses and mining techniques to data, she worked to make inferences about user behavior in the virtual environment.

Jennifer Stromer-Galley (Syracuse University, jstromer@syr.edu) studies the effects of new communication technology by presidential campaigns and is currently writing a book for Oxford University Press on the subject. She also examines uses and effects of new communication technologies in small groups, focused on political deliberation and on the communication of leadership and conflict. Current research includes serving as key personnel on two IARPA-funded projects (SCIL/DSARMD and Reynard/SCRIBE) and coprincipal investigator on the Deliberative E-Rulemaking Project, an NSF-funded project to apply natural language processing and multilevel deliberation to federal agency online rulemaking.

Tomek Strzalkowski (University at Albany, SUNY; tomek@albany.edu) is Professor of Computer Science and Director of ILS Institute at the University at Albany. Before coming to Albany, he was a Natural Language Group Leader and a Principal Scientist at GE R&D, and a faculty at the Courant Institute of New York University. His research interests include computational linguistics and sociolinguistics, human–machine interaction, and online and educational games. He is the principal investigator on several large federally funded projects, including IARPA's Sirius and Metaphor programs.

Andreas Tolk (SimIS Inc., andreas.tolk@simisinc.com) Andreas Tolk is Chief Scientist at SimIS Inc. in Portsmouth, Virginia. He is responsible for the evaluation of emerging technologies regarding their applicability for Modeling and Simulation applications, in particular in the domains of medical simulation, defense simulations, and architectures of complex systems. He is an adjunct professor at Old Dominion University. Dr. Tolk edited seven text books on systems engineering and modeling and simulation. He published more than 250 articles and papers in journals and conferences. He received over 30 best paper awards for his contributions. He received the Excellence in Research award from the Frank Batten College of Engineering and Technology in 2008, the first Technical Merit Award of the Simulation Interoperability Standards Organization (SISO) in 2010, the Outstanding Professional Contribution award of the Society for Modeling and Simulation (SCS) in 2012, and the Distinguished Professional Achievement award of SCS in 2014.

Peter Walla (Webster Vienna Private University, Austria; peter.walla@webster.ac.at). After a solid training in Zoological Neurophysiology on the single neuron level (spider eye photoreceptors), Peter started to focus on human memory functions by utilizing brain imaging methods with high temporal resolution. Various visiting research positions (Japan, Scotland, Australia) at renowned Universities strengthened his education in Cognitive and Affective Neuroscience and gave him the essential skills and expertise to obtain his Ph.D. and two further postdoctoral degrees, one in Cognitive Neurobiology (Habilitation at the Vienna Medical University) and the other in Biological Psychology (Habilitation at the Vienna University). In 1998, in the frame of his dissertation, he published a Nature article demonstrating neurophysiological correlates of nonconscious word memory. Since then he has focused on nonconscious processes in general (cognitive and affective). Peter is a Full Professor at the Webster University in Vienna (Austria), conjoint Professor at Newcastle University in Australia and Senior Research Fellow at Vienna University. He is not only an active Neuroscientist, but also offers his expertise as a Neuro-consultant (www.neuroconsult.com.au).

Günter Wallner (University of Applied Arts Vienna, Vienna; guenter.wallner@uni-ak.ac.at) is senior scientist at the University of Applied Arts Vienna where he received his doctorate degree in natural sciences for his thesis about GPU Radiosity in 2009. Before that he studied Computer Science at Vienna University of Technology from which he received his diploma degree for his thesis on game design in 2005. He also lectured on scientific and information visualization at the University of Vienna for 4 years. His research interests include the design, development, and evaluation of digital games as well as computer graphics, rendering, and visualization. His current research focus lies on analysis and visualization of game telemetry data. His work has been published in international journals and conferences, such as Computers & Graphics, Entertainment Computing, CHI, FDG, and ACE.

Elena Winzeler (The University of Texas at Austin, emwinzeler@utexas.edu) is an M.Ed. candidate in the Learning Technologies Program at the University of Texas at Austin and works as a Learning Experience Designer at Six Red Marbles.

A member of the Alien Rescue team since 2013, her involvement has centered on enhancing teachers' experiences by improving the teacher's manual and creating screencasts to guide teachers in using the program. In collaboration with others, she has recently designed an interactive dashboard mockup that aims to provide students' gameplay data to teachers for assessment and monitoring.

Reviewers

Name	Institution	Email
Alex Beaujean	Baylor University, USA	Alex_beaujean@baylor.edu
Anthony Betrus	SUNY Potsdam, USA	betrusak@potsdam.edu
Lingguo Bu	Southern Illinois University, USA	lgbu@siu.edu
Jae Hwan Byun	Wichita State University, USA	jh1016@gmail.com
David Gibson	Curtin University, Australia	david.c.gibson@curtin.edu.au
Dirk Ifenthaler	University of Mannheim, Germany	dirk@ifenthaler.info
Christian S. Loh	Southern Illinois University, USA	csloh@siu.edu
Jun Lu	American University, USA	lu@american.edu
Yanyan Sheng	Southern Illinois University, USA	ysheng@siu.edu

Part I
Foundations of Serious Games Analytics

Chapter 1
Serious Games Analytics: Theoretical Framework

Christian Sebastian Loh, Yanyan Sheng, and Dirk Ifenthaler

Abstract "Serious Games" is a unique industry that is concerned with the training/learning performance assessment of its clients. It is one of three digital technology industries (along with digital games, and online learning) that are rapidly advancing into the arena of analytics. The analytics from these industries all came from the tracing of user-generated data as they interacted with the systems, but differed from one another in the primary purposes for such analytics. For example, the purpose of game analytics is to support the growth of digital (entertainment) games, while that of learning analytics is to support the online learning industries. Although some game and learning analytics can indeed be used in serious games, they lack specific metrics and methods that outline the effectiveness of serious games—an important feature to stakeholders. Serious Games Analytics need to provide (*actionable*) *insights* that are of values to the stakeholders—specific strategies/policies to improve the serious games, and to (re)train or remediate play-learners for performance improvement. Since the performance metrics from one industry are unlikely to transfer well into another industry, those that are optimal for use in the Serious Games industry must be properly identified as *Serious Games Analytics*—to properly measure, assess, and improve performance with serious games.

Keywords Serious games analytics • Game analytics • Learning analytics • Definition • Theoretical foundation • Human performance assessment

C.S. Loh (✉)
Virtual Environment Lab (V-Lab), Southern Illinois University, 625 Wham Drive,
Mailcode 4610, Carbondale, IL 62901-4610, USA
e-mail: csloh@siu.edu

Y. Sheng
Department of Counseling, Quantitative Methods, and Special Education,
Southern Illinois University, 625 Wham Drive, Mailcode 4618,
Carbondale, IL 62901-4618, USA
e-mail: ysheng@siu.edu

D. Ifenthaler
University of Mannheim, L4, 1, Mannheim 68131, Germany

Curtin University, Perth, Australia

Deakin University, Melbourne, Australia
e-mail: dirk@ifenthaler.info

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based
Learning, DOI 10.1007/978-3-319-05834-4_1

1 From Edu-Games to Serious Games

Throughout history, games have always had a special place in the minds of ancient thinkers and scholars, for the sharpening of minds, and for the mediation of learning and training (e.g., military strategy with Chess). Given the (serious games) focus of this book, we will limit our discussions to digital games only, using the release of *Pong* in 1972 as the first landmark of digital games in modern history. Despite the sporadic use of digital games for learning and training (much like its non-digital predecessors), the term *serious games* was not known until many years later.

In fact, there are several accounts about the origin of the term (Crookall, 2010; Laamarti, Eid, & El Saddik, 2014). For instance, the term could be an oxymoron dating back to the Renaissance (i.e., *serio ludere*), or the title of a book, such as the Swedish novel published in 1912, *Den allvarsamma liken* (translated as *The Serious Games*, Söderber, 1977), or the book by Clark Abt in 1970 named *Serious Games* (Djaouti, Alvarez, Jessel, & Rampnoux, 2011).

In this chapter, we used the term *serious games* loosely to refer to: the Serious Games industry, the field of serious game research, and the digital games created for serious play. We will maintain the term in its plural form to reference various types and titles of *serious games*, and in singular form for its entirety, as in *serious games* analytics, or *serious games* industry.

1.1 Early-Days Digital Games for Learning

Although *The Oregon Trail* first debuted in 1971 to some schools in Minneapolis, the game was only made available to the public much later in 1985, on the Apple II platform. According to the official website (www.oregontrail.com), nearly 65 million copies of the game have been sold over the last 40 years, making this the most popular *educational game* (or *edu-game*) in digital game history. Although game score was implemented in that game, its purpose was to increase the challenge of the gameplay and thus, the entertainment value; and not for assessment of performance. Players received no additional bonus if playing as a banker. But they could double their scores if playing as carpenters, or triple it as farmers. This is equivalent to playing *The Oregon Trail* at a respective setting of Easy, Normal, or Hard.

As computer-based instruction became popular in the late 1980s, the advent of authoring software (e.g., Authorware, Director, and Flash) made it possible for educators to begin creating their own games for instruction. Sometimes, these games were used to teach specific subjects or skills, while other times, they could be used to illustrate difficult concepts or procedures. As long as the intentions of these educational games were for "show-and-tell," learners' performance assessment was never really a concern for educators—especially when they have other means of assessing students' learning in the classrooms (Ifenthaler, Eseryel, & Ge, 2012).

The success of early computer-based educational games (such as *The Oregon Trails*) enticed many publishers, and even teachers, to venture into the educational games market. An industry for *edutainment* soon arose from the 1990s as computers and educational technology became commonplace in the classrooms. *Edutainment* was a portmanteau created to denote the marriage of *edu*cation and enter*tainment*. While the intention was to make learning more entertaining and motivating (at least to school-aged children) by injecting game elements (e.g., animations, wacky sounds, bright colors, challenges) into boring learning materials, the quality of edutainment soon plummeted as more and more publishers rushed to release poorly designed games into the system for quick profits. Once the low quality edutainment began to fill the market, they were chided by pundits as "drill-and-kill games" (Prensky, 2001) and the edutainment industry was doomed for failure (Van Eck, 2006).

By the turn of the century, the term *digital game-based learning* (DGBL) was popularized through the writings of Marc Prensky (2001) and James P. Gee (2003). Gee even listed 36 learning principles where good commercial-off-the-shelf (COTS) video games can affect how people learn and, therefore, potentially revolutionize education. Such writings have since been quoted widely as they not only legitimized digital games for research by the academia, but also for learning and training in the education, business, and healthcare industries (see Aldrich, 2005; Michael & Chen, 2006).

1.2 The Serious Games Industry

The year 2002 became known as the start of the wave of *serious games* because of two major events (Djaouti et al., 2011). The first was the release of the *Serious games*: *Improving public policy through game-based learning and simulation* report (Sawyer & Rejeski, 2002) by the Woodrow Wilson International Center, which later became the impetus for the formation of the Serious Games Initiatives. The second was the public release of *America*'*s Army*, a "war-game" commissioned by the US Army to showcase the military life as an "engaging, informative, and entertaining" experience (McLeroy, 2008b). The game went on to receive many accolades for its design and become the most successful recruitment tool for the US Army (Turse, 2003). As of August 2008, the game was downloaded 42.6 million times and accumulated 9.7 million registered users from over 60 countries, as well as 230 million hours of playing time.

Two new organizations soon arose to take the place of Serious Games Initiatives: Games for Health (GFH) and Games for Change (GFC). As global situations intensified in the past few years, governmental agencies (such as the FBI and Homeland Security) have looked towards serious games to facilitate training and public awareness in areas such as cybersecurity, homeland safety, and disaster preparation. Laamarti et al. (2014) also reported rising numbers of research publications using *serious games*, *serious gaming*, and *serious play* as keywords.

Every December in Orlando, FL., an event called *Serious Games Showcase and Challenge* (SGS&C) takes place alongside the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC), which happens to be the world's largest military simulation convention. This annual event is an international serious games competition that draws professional, independent, and student developers from all over the world to submit serious games in four categories covering the business, academic, government, and mobile sectors.

According to BankersLab (2013), about 25 % of the Global Fortune 500 companies have already adopted serious games for training—particularly from the United States, Britain, and Germany. Recent market research report (Ambient Insight, 2013) forecasted the serious games market to reach $2.3 billion, in addition to the $6.6 billion in the simulation-learning market in 2017. The latter is not only limited to military training, but also include new areas of medical and surgical training using simulation (and serious games). Recent trends revealed the future to be on the side of mobile gaming. The latest market report from Ambient Insight reveals the mobile serious games market has a 5-year compound annual growth rate of 12.5 % and is expected to reach $410 million by 2018. Without a shadow of a doubt, *Serious Games* has come of age.

2 Serious Games: Not for Entertainment

People carry different mental images about what *digital games* should be. Such mental images are typically formed at an earlier age: either by observing the gameplay of others, or through their own experiences in interacting with digital games. This is why a discussion about (digital) game-based learning (DGBL) can be as meaningless as arguing about what food is good for you: what is considered to be food to one person may not even be edible to another. It is preferable to use *serious games* instead.

There have been several attempts to define the term "serious games", and among them are:

1. Abt's definition (1987)—serious games "have an explicit and carefully thought-out educational purpose and are not intended to be played primarily for amusement" (p. 9),
2. Zyda's definition (2005)—serious games are "mental contests played with a computer in accordance with specific rules that uses entertainment to further government or corporate training, education, health, public policy, and strategic communication objectives" (p. 26), and
3. Sawyer's definition (2009)—serious games include "any meaningful use of computerized game/game industry resources whose chief mission is not entertainment." We will compare this definition with his original definition (Sawyer & Rejeski, 2002) in Sect. 2.

In summary, serious games are "digital games created not with the primary purpose of pure entertainment, but with the intention of serious use as in training,

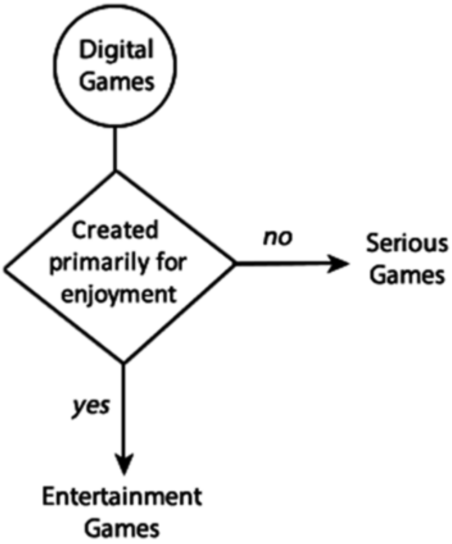


Fig. 1.1 Entertainment games vs. serious games

education, and health care." This is probably the most widely accepted definition for serious games at the moment (Fig. 1.1).

In our view, referring to *serious games* as "any digital game that is not created with the primary purpose of entertainment" is far too broad and simplistic. One reason is that non-entertainment games have long existed before the Serious Games Initiatives. What then, are the differences between serious games and DGBL? This definition will only muddle the situation and make it harder to differentiate the "real" serious games (post 2002) from the early *educational games* (such as *The Oregon Trail*) and the failed *edutainment*.

For instance, a recent survey on GameClassification (http://www.gameclassification.com) even included non-entertainment games from as early as the 1950s to 2000s (Alvarez, Djaouti, Rampnoux, & Alvarez, 2011). Their categories for Serious Games included: games for storytelling, advertisement, and propaganda created for informative, subjective, educative, marketing, and communicative *message broadcasting* (Fig. 1.2).

2.1 Message Broadcasters Are Not Serious Games

Alvarez et al. (2011) found that up to 90 % of serious games consisted of *message broadcasters*: non-entertainment (or serious) games created with the purpose of broadcasting a certain message through one-way communication. Only about 10 % of non-entertainment games were made with skill improvement or training as their primary purpose. One simple explanation is that message broadcasters are easier

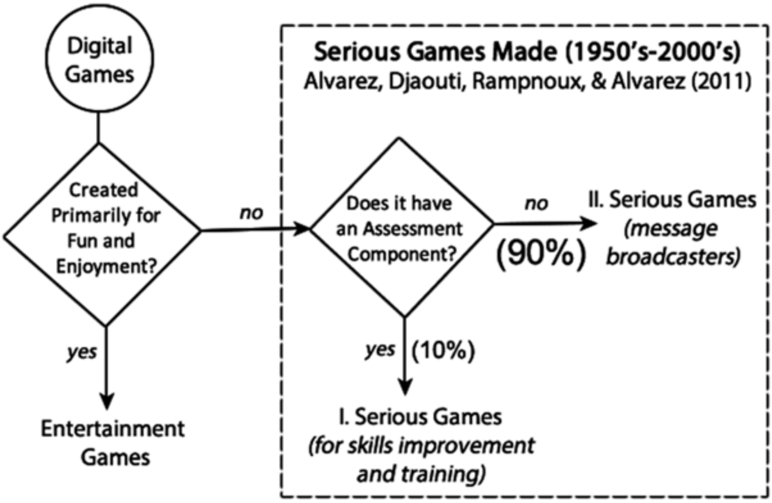

Digital Games
Created Primarily for Fun and Enjoyment?
no
yes
Entertainment Games
Serious Games Made (1950's-2000's)
Alvarez, Djaouti, Rampnoux, & Alvarez (2011)
Does it have an Assessment Component?
no
(90%)
II. Serious Games (message broadcasters)
yes
(10%)
I. Serious Games (for skills improvement and training)

Fig. 1.2 Difference between entertainment and serious games

and cheaper to make than those made for training because there are less work involved—programmers need not create an assessment component for message broadcasters. There is also no need to present overly comprehensive or accurate information. We think this survey further muddles serious games for what they really are because it adds many immature (early educational games) or poor examples (from the failed edutainment era) into the mix. It would make more sense to create a serious games database based on serious games created after the serious games report (Sawyer & Rejeski, 2002).

Message broadcasters are closer in nature to entertainment games because they are assumed to have "done their job" by virtue of design. Assessment is almost never a first priority, but is often only added as an afterthought. Many educational games and edutainment that were created to teach subject contents without appropriate assessment can all be labeled as message broadcasters. For example, *America's Army* was created to showcase the army (with the intention to recruit).

A more recent example is the *3M Wind Energy Virtual Lab* (available at http://scienceofeverydaylife.com/innovation/), which claimed to be an inquiry-based learning lab created to challenge school children to find the best renewable energy available to support 400 households with the lowest cost per year (Schaffhauser, 2014). However, upon closer inspection, the game turned out to be nothing more than a one-sided promoter for 3M's products. Students never had any opportunity to learn why wind energy was chosen (above other forms of energy, be it environmental friendly, or not). Instead, the game highlighted how 3M chemical coatings can greatly improve the efficiency of wind tower turbine blades.

What kind of characteristics should we expect from serious games? The National Summit on Educational Games (Foundation of American Scientists, 2006) identified the following attributes that are important for (game-based) learning: clear goals, repeatable tasks (to build mastery), monitoring of learners' progress, encouraging increased time on task (through motivation), and adjusting the learning difficulty level to match learners' level of mastery (i.e., personalization of learning). Play-learners acquire skills and processes that are not easily taught in classrooms, including strategic and analytical thinking, problem-solving, planning and execution, decision-making, and adaptation to rapid change.

Moreover, serious games are able to contextualize play-learners' experience and support situation cognition (Watkins, Leigh, Foshay, & Kaufman, 1998). It is obvious that the strength of serious games lies in the improvement of skills, performance, and decision-making processes, rather than in message broadcasting or information dissemination. Djaouti et al. (2011) simply named this category of serious games as *training games*, in addition to message broadcasters and data exchange games.

3 Gamification, Game-Based Learning, and Serious Games

When an article with the title "Gamification, game-based learning, and serious games, what's the difference?" (Drager, 2014) is reposted to the game-based learning community in LinkedIn, you know something is amiss. If those who are interested in GBL do not understand the differences among the terms, then who else would? As more and more "experts" opine on what each of these terms should mean, the discussion related to analytics quickly becomes very muddled.

3.1 Gamification Is Not Games

Gamification is not a game at all! Instead, it borrows from the concept of game mechanics to motivate people to continue certain behaviors—such as posting photos on Facebook, booking hotels using mobile apps, and encouraging the sales force to work harder, through point systems, badges, or monetary awards. Gamification can be used in conjunction with digital games (Domínguez et al., 2013), but it is not a new type of digital game in and of itself.

Although we do not see gamification to be anything like serious games (because they are not games), we are somewhat concerned that we may soon witness the resurrection of the (failed) edutainment under the guise of gamification, given the recent chatter regarding "the use of game thinking and game mechanics to engage audiences and solve problems" (Zichermann & Cunningham, 2011).

Thus far, most gamification examples are from industries where the administration has made use of award points and badges to motivate their sales/work teams. [Readers who are interested in exploring more about gamification should refer to the book, *Gamification by Design* (Zichermann & Cunningham, 2011) for concrete examples of gamification in the industries.]

But as the idea gathered steam, even educators became excited about how they might begin to gamify their e-learning classrooms (especially in higher education). However, not everyone is on the same page as to what gamification really entails in the e-learning classrooms. As Gerber (2012) observed, "often it seems that the spaces of edutainment and games-based learning get mixed up with gamification" (p. 1). Although gamification is not the same as games, those who are from the outside looking in can easily confuse the term to mean "turning something into games."

There are two disturbing trends with gamification that we have observed. The first is researchers passing off edutainment projects as gamification. We have come across a GBL project where the *Temple Run* game is to be sectioned up with multiple-choice questions for nursing education. Despite our advice to the project leader to avoid turning this into another edutainment game, he or she decided to forge ahead because the project would be "easily funded" when presented as gamification. In the end, the *Temple Run* game did receive funding as a gamification project. Perhaps the reviewers did not really know the differences between edutainment and gamification, or perhaps they simply did not care.

The second trend is the attempt to gamify e-learning with *games* (i.e., Flash-based animations and message broadcasters). Notice how the term 'gamify' seemed to imply "enhancing e-learning with the addition of games", instead of using game mechanics to motivate e-learners. You have probably heard of other projects where people try to enhance online classes using animations and games, and claiming gamification. Over time, such actions will cause the line between edutainment and gamification to blur. After all, who are we to say if what you are doing is (is not) gamification? People are entitled to their own opinions, aren't they? But if more edutainment is being passed off as gamification, we believe it will eventually fail—because the edutainment of the 1990s was a failed attempt to make learning more game-like.

When that happens, it will likely result in another round of lost confidence among stakeholders. Being none the wiser, these stakeholders may lump serious games with edutainment and gamification, and declare this movement to be another failed attempt to mediate learning through games. This is why it is extremely important for the serious games community to be careful about how they approach and handle the subject of gamification.

3.2 Problems with Game-Based Learning: Media Comparison

Educators have lamented about the lack of *evidence of benefits* for DGBL (e.g., Kirriemuir & McFarlane, 2003; Sandford & Williamson, 2005) from the start of the serious games movement. Research studies based on subjective data obtained from surveys, self-reports, and pen-and-paper tests are rather unconvincing. As Van Eck (2006) pointed out, "we are not likely to see widespread development of these games … until we can point to persuasive examples that show games are being used effectively in education."

But how do you go about measuring the effectiveness of game-based learning (or serious games) in a study? This question is important to many serious games publishers because if they can get their hands on data to prove that their products is more effective than, say, traditional classroom teaching, it could boost sales! Comparing technology-enhanced instruction against a traditional classroom taught face-to-face by a teacher seems to be an easy comparison, especially if you were to position the latter as an older method, in contrast to the newer, more technologically advanced method.

This approach was used extensively in the early days of computer-based instructions by instructional designers to study the effectiveness of instructional delivery media. After many years of cookie-cutter research with inconclusive findings to support teacher or technology, the *Media Comparison Studies* (MCS) method was severely criticized and debunked by Salomon and Clark (1974).

3.2.1 Media Comparison

Discussions about MCS can sometimes be confusing to someone from outside the field of instructional design because teachers are considered an instructional medium. When placed in that light, Clark (1985b) was able to explain why meta-analysis of years of MCS often showed "no statistical significance." After all, how do you begin to compare technology-enhanced instruction against a master teacher, or an inept teacher? In the former, the learning outcome will likely favor the master teacher, while in the latter, findings will likely support the technology. Serious games researchers should avoid introducing confounds into their studies (e.g., comparing methods of teaching/delivery of instruction, be it computer-based instruction, online learning, or serious games).

Even though Clark has been proven right for 40 years, his writings are less well circulated outside the field of instructional design, save medical education (Clark, 1992). It is not surprising that Hastings and Tracey (2004) reported a resurgence of MCS because younger researchers who were not familiar with Clark's perspective were once again falling into the same trap to measure the effectiveness of technology via media comparison. Journal editors outside the field of instructional technology are encouraged to familiarize themselves with the "Clark vs. Kozma debates" to maintain a high level of rigor in research reporting (see Clark, 1983, 1985a, 1985b, 1994; Kozma, 1991, 1994).

An MCS is easily identified by its *comparison* between two media, or methods, of instruction (e.g., the Internet, a teacher, or any emerging technology). Interestingly, Clark's observations about MCS are technology independent and can be applied to almost any kind of technology used in learning and instruction. Examples of media comparison design involving game-based learning can be seen in the study by Moshirnia (2007) and the study with the Maritime Warfare School (Caspian Learning, 2010). As mentioned previously, this research method is flawed and should be avoided at all costs in serious games research.

3.2.2 Pretest–Posttest Validity

Pretest and Posttest design is by far the most common methodology used in serious games research (Bellotti, Kapralos, Lee, Moreno-Ger, & Berta, 2013). Pretest–Posttest studies have external validity issues because players cannot be sufficiently quarantined throughout the play period of serious games, which often last up to 20 or 40 h per game, over several days or weeks. In other words, the inquiry method of Pretest–Posttest would not ensure that changes in learning performance are only attributed to serious games. In addition, maturation can be another issue as players share information with one another about how to overcome certain game levels. Cheat sheets and walkthroughs created and posted to the Internet by other players can further exacerbate the problem.

In the real world, players can spend days, if not months, completing a serious game. In comparison, many serious games studies in the literature employ only single session gameplay lasting 5–30 min as their research conditions (Byun & Loh, 2015; Grimshaw, Lindley, & Nacke, 2008; IJsselsteijn, de Kort, Poels, Jurgelionis, & Bellotti, 2007). This is a far cry from the real-world experience and can severely limit the ability to generalize findings from these studies.

3.2.3 Talk Aloud and Self-Reports

A large portion of serious games research employs self-reports and perception questionnaires to collect data about users' beliefs and additional feedback. Because self-reports and talk-aloud methods produce subjective data, they are often wrought with bias—participants tend to report what they think the researchers want them to say (Donaldson & Grant-Vallone, 2002). Such actions can contaminate the data and could easily threaten the validity of a study.

We understand that think-/talk-aloud protocol is a staple method for Human–Computer Interactions (HCI) research and has been used extensively for usability studies of serious games. Although think-/talk-aloud is an acceptable and popular research method for User eXperience (UX) studies, we do not think they are suitable for *serious games analytics* research. Another staple of HCI studies is the A/B test. Researchers need to be cautious as to how they design the A/B test because it can easily fall into the trap of media comparison.

Without proof of effectiveness, high production costs and an unknown Return of Investment become important factors that deter decision-makers with strong business acumen from adopting serious games. We need better ways to evaluate and assess the learning performance of serious games, and strong evidence to convince stakeholders that serious games can indeed improve play-learners' performance (Nickols, 2005). But before we talk about analytics, a redefinition of serious games may be in order.

4 Serious Games as Tools

America's Army was a highly visible showcase and successful recruitment tool for the US Army. It was envisioned by Colonel Casey Wardynski to take advantage of the computer gaming technology to provide the public with "a virtual Soldier experience that was engaging, informative and entertaining" (McLeroy, 2008a). Interestingly, this project was not the first digital game worked on by the US military.

The first group to try their hand at game modification (or game modding) was the US Marines. The commission came from General Charles Krulak, a Commandant of the US Marine Corps. His directive (MCO 1500.55, Krulak, 1997) was to explore the possibility of using PC-based war-games as a tool to *improve* military thinking and decision-making skills due to increasingly limited time and opportunities for live training exercises.

Marine DOOM (so named because the game was modified using id Software's *DOOM II*) was a networked multiplayer game, playable by four soldiers (with different responsibility), to promote the importance of teamwork. The MCO 1500.55 directive further expressed the need to "continue development of new training tools" for training and skill improvement using games. Readers should note the shift in objectives from human performance improvement (*Marine DOOM* created in 1996) to message broadcasting—recruitment (*America's Army* created in 2002). The games obviously served different needs and played different roles at different times. As with any technology, serious games can be used as tools, both to improve skills/performance and to broadcast messages.

The original intent of serious games was to make use of the advanced digital (gaming) technology to create *tools* for skills and performance improvement. The same sentiment was echoed in the report, *Serious Games*, by Sawyer and Rejeski (2002). They said, "Many organizations are turning to games… for help in improving the evaluation, prediction, monitoring, and educational processes surrounding their policy development" (p. 5). They went on to note that games are becoming "extremely effective training tools" (p. 11). Similarly, Michael and Chen (2005, para. 3–4) asserted that "it's not enough to declare that *game teach* and leave it at that… Serious games, like every other tool of education, must be able to show that the necessary learning has occurred."

4.1 Games for Skills and Human Performance Improvement

Going forward, we believe it is important to return to the root of serious games by placing *performance improvement* back into the definition. We submit that "serious games are digital games and simulation *tools* that are created for non-entertainment use, but with the primary purpose *to improve skills and performance* of play-learners through training and instruction." The term *play-learner* is a homage to Johan Huizinga, who said, "Let my playing be my learning, and my learning be my playing."

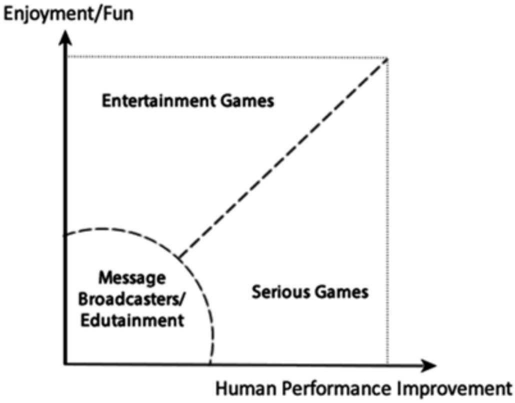

Enjoyment/Fun
Entertainment Games
Message Broadcasters/ Edutainment
Serious Games
Human Performance Improvement

Fig. 1.3 Differences among entertainment games, message broadcasters, and serious games

A play-learner who trains and learns with serious games will "play as they learn, and learn as they play." This means, the end result of play-learning with serious games should be to *improve skills and performance* of its users, and not stop short at information or knowledge acquisition.

Although some researchers are beginning to use serious games as a testing and assessment tool (Herold, 2013), the term *performance* should include human (work/ learning) performance also. When Krulak (1997) first commissioned serious games for military training, his intention was to improve the users' decision-making skills and work (combat) performance. While there is nothing wrong with assessing play-learners with "games as tests" (e.g., stealth assessment), it is a balancing act between testing and training. Researchers would do well to remember that the strength and attraction of *serious games* are in "learning by doing" (Aldrich, 2005)—bringing the learning contexts into a game environment, and not "learning by testing"—because this reminisces the "drill-and-kill" edutainment approach.

Figure 1.3 depicts how we view entertainment games, message broadcasters, and serious games. Each is defined by how much they score along the axis of *enjoyment/fun* and *human performance improvement*. We understand that by inserting human performance improvement back into the definition of *serious games*, we may have created more problems: How do we differentiate these types of games from the message broadcaster? Should there be a new name for this kind of performance improving games? (We will leave it to the Serious Games industry to figure this one out.)

We think the name, *serious games*, has great appeal (Crookall, 2010), and is appropriate, if only we could separate message broadcaster games from the mix. An alternative, *immersive games*, may serve our purpose, but do we really want another term? For serious games to be useful for learning, additional cognitive support, specifically debriefing—whether in-game or after game, as in After Action Review (AAR) for the military—is absolutely necessary (Crookall, 2010). By casting serious games as a *tool to improve skills and performance*, the grounds are thereby provided to begin the discussion of *serious games analytics*.

Furthermore, once we concede that the primary purpose of serious games is to improve skills and performance, the need for a *stealth*-like approach to assessment (DeRosier, Craig, & Sanchez, 2012) would be eliminated. Besides giving people the wrong impression about assessment, the word *stealth* also implicates serious games assessment to be some kind of a covert operation. While it is true that game-based learning assessment was indeed measured covertly a few years ago (i.e., people were not fully informed about what kind and how much *gameplay data* were being collected), this issue is now moot as the floodgate to user-generated data collection has long since been flung open by the proliferation of mobile apps and *datafication*.

4.2 Gameplay Data

By *gameplay data*, we refer to players' in-game actions and behaviors that are digitally traced through numerical variables within the game environments; particularly actions or behaviors stemming from key decision points or game events. *Players*' *behaviors* are simply the course of actions taken by players during the process of problem-solving within the (serious) games. A single action is an isolated event, but repeated actions (when players faced with similar scenarios) constitute behavior.

As Medler (2011) observed, "a prevalent feature in many game and platform systems" is to record players' *gameplay data* for (market) research and analysis. The data mining process described (i.e., data recording, data cleaning, data analysis, and data visualization) can even occur near real-time in situ, meaning, occurring within the game environments as the game is still in process—via advanced methods such as game telemetry (Zoeller, 2013), or *Information Trails* (Loh, Anantachai, Byun, & Lenox, 2007).

Collecting players' *gameplay data* for analysis was once both costly to implement (e.g., paying each player for their time) and difficult to execute, "requiring strong analytical skills and experience" (Wallner & Kriglstein, 2012). Privacy laws of different countries may also limit the collection and sharing of data to ensure the protection of peoples' personal data. But the advent of mobile technology, social networking, sharing of information, datafication, and self-quantification has changed how people view their data. Collection of gameplay data and sharing what used to be private information with friends (and friends of friends) are now viewed as being sociable.

4.3 Datafication

According to Cukier and Mayer-Schoenberger (2013), *datafication* is the conversion of all aspects of life (or life's activities) into data. Advances in mobile technology, particularly the advent of fitness/activity trackers (e.g., Fitbit Flex, Jawbone Ups, and Nike Fuel Band), have fueled a new fad for self-quantification.

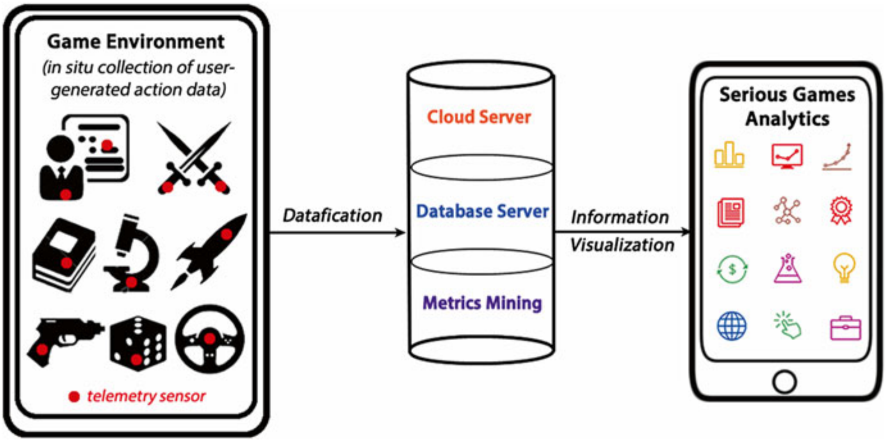


Fig. 1.4 From datafication of in-game actions to analytics

Many people have begun to take notice of how many steps they take, how much, and what food/drink they consume, how many hours they spent sitting, sleeping, exercising, etc.

Humanity finally has the capability to datafy activities into information (via various monitoring devices) that they can manipulate and understand. People seem highly interested in making sense of the activities that are happening to and around them (see Fig. 1.4). The iOS8 even have a healthkit function built-in ready to capture and share data with these activities trackers (Preimesberger, 2014). Why is there such a fascination of datafying human activities? Cukier and Mayer-Schoenberger (2013, para 25) explained, "Once we datafy things, we can transform their purpose and turn the information into new forms of value."

In serious games, we are doing the same thing, i.e., *datafying* user-generated data of play-learners' within serious games and turning them into "new and valued information" for skills and performance improvement, as *Serious Games Analytics*. For example, researchers are now looking into innovative means to datafy game consoles (e.g., X-Box Kinect and psychophysiological headbands) to co-opt human gaits and emotions as analytics:

- Rehabilitation of stroke patients (Chap. 10 in this volume)
- Teaching and evaluation of medical procedures (Chaps. 9 and 11 of this volume)

4.4 In Situ vs. Ex Situ Data Collection

There are two ways to collect play-learners' gameplay data: ex situ, or in situ method. Ex situ data are collected from "outside the system" where the object or event under observation lives. User survey data (demographics, feedback), pretest/posttest, talk aloud, and focus-group interviews all fall under this category because they are typically collected in the real world, not from within serious games environments. Research data are often collected ex situ out of convenience or constraints. Constraints can include imminent danger to the researchers (e.g., measuring the temperature of the sun), costs (e.g., sending researchers to Mars), size restrictions (e.g., observing a red blood cell in a human body), or Black box conditions that make it impossible to access the innards of a system (in the case of serious games, ex situ data can be seen as an indirect data collection method).

- Meta-analysis of data collection methods in serious games (Chap. 2)

In comparison, in situ data are collected from "within the system," in which an investigated event occurs. Since serious games are nothing more than software application, it is easy for computer scientists and software engineers to directly manipulate the variables and functions in serious games to track what play-learners actually do in the game environment. In situ data collection can be made possible via: (a) data dump as log data, (b) game *telemetry* (popularized in the entertainment game development circles, by Zoeller, 2013), and (c) *Information Trails* (originated from the instructional design circle, by Loh, 2006, 2012b).

Knowing about in situ data collection is not enough because we have yet to address the issue of assessment. More specifically, is it possible to think of assessment using similar terms, as in situ or ex situ? What advantage does in situ assessment (within the game habitat) offer over that of ex situ (outside the game system)? Using *log data* as an example, although log data is a kind of in situ data, generated during the game via in-game user actions, it is typically collected for analysis post hoc—after the gameplay has completed. Moreover, log data are usually analyzed apart from the game session by analysts who are located elsewhere. As such, it involves an ex situ assessment process.

In comparison, *Information Trails* and *telemetry* comprise in situ data collection processes, alongside in situ assessment systems. The greatest advantage of an in situ analysis algorithm built (to tap) into the game engine is that it allows for ad hoc (formative) assessment. This means that stakeholders can access the assessment report as the game is still in progress, without waiting for play-learners to complete the entire game.

- Log data (Chaps. 4 and 13 of this volume)
- Telemetry (Chaps. 3, 5 and 7 of this volume)

4.5 Actionable Insight: Using Analytics to Improve Skills and Human Performance

Amazon went to great lengths to create a data analytics system to trace and analyze the online purchasing habits of their customers. LinkedIn wanted to understand how each registered user connected with one another and to discover the obscure pattern of human interactions and social networking linkages. Similarly, stakeholders of serious games are interested in understanding what play-learners might do in certain training scenarios to improve skills and performance, increase return of investment, reduce human errors, and mitigate retraining or remediation (Loh, 2012a).

When serious games are finally recognized to be much more than edutainment "on steroid," we can expect researchers to start using them as "tools" for data collection and research (Herold, 2013). Serious games have come full circle in that they are finally becoming what they were originally set out to be—as *tools* for skills and performance improvement, and additionally, turning the information obtained into new insights and policies of value to stakeholders (Cukier & Mayer-Schoenberger, 2013).

Whether this information is being known as *actionable intelligence* (Rao, 2003), *actionable insights* (LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011), or *analytics* is mere semantics. The crucial point here is this information can be extremely valuable to various stakeholders (Canossa, Seif El-Nasr, & Drachen, 2013; Nickols, 2005) in assisting in future decision-making processes, enhancing the training systems through (re)design, improving the skills and performance of trainees, lowering human error rates through retraining, and creating new revenues through monetization.

- Data-driven game design (Chaps. 12 and 17 this book)
- Expertise Performance Index (Chap. 5 of this volume)
- Monetization (*Game Analytics*: Canossa et al., 2013)

5 Types of Analytics

There is a fallacy in the era of Big Data that the resulting sheer amount of data from collecting everything available in the system would provide the answer to every problem under the sun (Cukier & Mayer-Schoenberger, 2013). While storing exabytes of Big Data may be advantageous, or even necessary, for Google and Facebook (Miller, 2013), trying to collect all that data is not a wise move for most other (smaller) organizations. Besides, since there is yet to be any massive multiplayer online serious game, where would one go to gather that kind of Big Data?

Despite the falling cost of data storage, keeping huge amounts of data around can still be a pricey affair. There is also the question as to how one goes about analyzing the data. Furthermore, the idea to collect *all* gameplay data of play-learners indiscriminately is both inefficient and asinine. Since data collection would need to occur online, too much network traffic in addition to the large amount of gaming graphics that need to be transmitted can result in severe game lag and detrimental gameplay experience—directly affecting the performance of the tool (i.e., serious games).

Online collection of gameplay data for game-based analytics necessitates careful planning due to the simultaneous transmission of gaming (graphical) data, along with the gameplay data. In comparison, online collection of user-data for web-based analytics is much simpler because the information is largely text based and analysts do not need to worry about transmitting gaming (graphical) data.

Also related to *analytics*, is the issue of (performance) metrics. Some metrics are so common that they can be found across many different industries, covering the same grounds in learning, gaming, and business analytics. Some of these are: socio-demographics information, purchasing habits, and likely items to put in the same *bin* (a data mining terminology). Others include metrics from UX research in the field of HCI. For example:

- Time required to complete the lesson or game session
- Number of mistakes made during the lesson or game session
- Number of self-corrections made
- Time of access (login to logout time)
- Amount of learning/gaming contents accessed
- Specific types of learning/gaming contents accessed, and others

The serious games market is quickly becoming muddled over what constitutes "analytics" and which metrics to include because there is no proper taxonomy available (Ifenthaler, 2015). The unhelpful addition of voices from various "experts" and pundits only serve to further confuse the matter (Hughes, 2014). For instance, equivalence has been drawn between analytics and assessment; there are even suggestions to use SCORM/xAPI (meant for Learning Management Systems, or LMS) to track user-activities in serious games. Such confusions have provided us with the urgency to clear up some of the gobbledygook. In the following sections, we will examine the differences between *learning analytics* and *game analytics*, and the reasons to establish an independent *serious games analytics*.

5.1 Learning Analytics

Several concepts closely linked to processing educational data are educational data mining (EDM), academic analytics, and learning analytics. However, these concepts are often confused and lack universal agreements or applied definitions (Ifenthaler, 2015).

- *Educational data mining* (EDM) refers to the process of extracting useful information out of a large collection of complex educational data sets (Berland, Baker, & Blikstein, 2014).
- *Academic analytics* (AA) is the identification of meaningful patterns in educational data in order to inform academic issues (e.g., retention, success rates) and produce actionable "strategies" in budgeting, human resources, etc. (Long & Siemens, 2011).
- *Learning analytics* (LA) emphasizes on insights and responses to real-time learning processes based on educational information from digital learning environments, administrative systems, and social platforms. Such dynamic educational information is used for real-time interpretation, modeling, prediction, and

> optimization of learning processes, learning environments, and educational decision-making in near real time (Ifenthaler & Widanapathirana, 2014).

Applications of learning analytics presuppose a seamless and system-inherent analysis of learner's progression in order to continuously adapt the learning environment. Learning analytics provides the pedagogical and technological background for producing real-time interventions at all times during the learning process. It is expected that the availability of such personalized, dynamic, and timely feedback supports the learner's self-regulated learning, as well as increases their motivation and success. However, such automated systems may also hinder the development of competences, such as critical thinking and autonomous learning (Ifenthaler, 2015).

5.1.1 Metrics for Learning Analytics

Metrics for learning analytics include the learners' individual characteristics, such as socio-demographic information, personal preferences and interests, responses to standardized inventories (e.g., learning strategies, achievement motivation, personality), skills and competencies (e.g., computer literacy), prior knowledge and academic performance, as well as institutional transcript data (e.g., pass rates, enrollment, dropout status, special needs).

Other metrics included in learning analytics frameworks are snippets or streams from the social web, which may highlight preferences of social media tools (e.g., Twitter, Facebook, LinkedIn) and social network activities (e.g., linked resources, friendships, peer groups, Web identity). In addition, physical data about the learner's location, sensor data (e.g., movement), affective states (e.g., motivation, emotion), and current conditions (e.g., health, stress, commitments) may be used as learning analytics metrics, if available.

Another important metric of learning analytics is the rich information available from learners' activities in the online learning environment (i.e., LMS, personal learning environment, learning blog). These, mostly numeric data, refer to logging on and off, viewing and/or posting discussions, navigation patterns, learning paths, content retrieval (i.e., learner-produced data trails), result in assessment tasks, responses to ratings and surveys. More importantly, rich semantic and context-specific information are available from various sources, including discussion forums, complex learning tasks (e.g., written essays, wikis, blogs), interactions between facilitators and students, online learning environment, and others (Ifenthaler & Widanapathirana, 2014).

5.2 Game Analytics

A much more detailed treatise of game analytics is already available in our companion book, *Game analytics*: *Maximizing the value of player data* (Canossa et al., 2013) and will not be repeated here.

It suffices to understand that since most digital games are created to entertain its customers and for profit, game analytics are likely to comprise metrics that are aimed to improve gameplay. Examples include: correct game balance, better game design, detect hidden problems, relieve bottlenecks, categorize game contents by players' like and dislike, differentiate types of players (see Thawonmas & Iizuka, 2008; Williams, Yee, & Caplan, 2008), and identify new opportunities for post-sales revenues (i.e., monetization, see Canossa et al., 2013).

Although digital games and serious games are both based on business models and must ultimately lead to monetary profits in order to survive, the two industries are created to meet the needs of different markets. Digital games are created with the primary purpose to entertain, whereas serious games are primarily designed "to support knowledge acquisition and/or skill development" (Bellotti et al., 2013).

Besides the usual business and user-experience-related metrics, game analytics produces insights that are reflective of players' enjoyment in gameplay. Metrics that are specific to game analytics include hours of continuous play, frequency of return to the game server, length of subscription for subscription-based massive multiplayer online games (MMOGs), in-game/in-app purchases, and so on. These metrics are useful in determining how captivated a player is to the game contents allowing game developers to understand how game design directly affects players' gameplay/enjoyment, and determine the kind of contents that players are willing to pay for in the future.

5.3 Does Game Analytics + Learning Analytics = Serious Games Analytics?

If serious games are games/tools created *not* for the primary purpose of entertainment, but for skills and human performance improvement of play-learners, then a logical question to ask would be: "Does Game analytics + Learning analytics = Serious games analytics?" Borrowing from the Pareto principle (i.e., the 20–80 rule), the answer is "20 % Yes" and "80 % No."

The "20 % Yes" comes from the fact that some metrics may be commonly found in both or all three industries and can yield some general analytics (e.g., time of completion, length of access, and others). But because learning analytics and game analytics are from *distinctly* different industries, there really ought to be a different set of metrics that are specifically tailored for serious games. Although the serious games industry can indeed repurpose learning analytics and game analytics metrics to obtain analytics, such insights are incomplete because these metrics are not conceptualized optimally for serious games; hence, the "80 % No."

Consider this: Why should game developers and publishers make use of learning analytics to improve their game design or improve sales? Put it the other way: Will researchers and educators from the learning analytics, or EDM, community consider game analytics for learning improvement? This will never happen because the tools are not designed to fit the tasks at hand, nor are they designed by experts of that field.

Serious games communities have their own ways to solve problems, and these methods can be quite different from those favored by the learning analytics and game analytics communities. In the LinkedIn social network, for example, members are often asked about evidences of serious games effectiveness. Such questions are never asked within the entertainment gaming community as they make little sense there.

A peruse of the Society for Learning Analytics Research (SoLAR) website revealed a small section regarding the potential to use learning analytics as "serious gaming analytics" (www.solaresearch.org/events/lasi-2/lasi/lasi-program-wednesday/serious-gaming-analytics/). The website claims that "*games are quickly gaining momentum as a tool for learning and assessment… However, the methods by which to harness this data to understand learners and improve our games are less than clear.*"

One explanation could be: even though game analytics is a "hot topic" and related to learning analytics, there is no critical mass in the Learning Analytics community to make this into a first priority. Both man power and resources are needed to clarify and research the methods that are useful in obtaining serious games analytics from games. Barring any unforeseen new developments, it is likely that (serious) gaming analytics will remain a peripheral interest for the group of learning analysts. We have also observed a growing interest among the psychological/educational testing and measurement researchers to adapt digital games for classroom assessment with interests and supports from third-party "testing companies," as can be seen in:

- Evidenced-Centered Design and Stealth Assessment (Chaps. 4, 12, 13, 14 and 15 of this volume)

5.4 Why Serious Games Analytics?

Similar to how serious games differ from entertainment games by their primary purposes, serious games analytics also differs from game analytics by the primary purpose of *skills and performance improvement*, as that is the primary objective in training.

The primary purpose for game analytics is to (a) improve gameplay and make the games more enjoyable to the players, and (b) improve game design and create content that players like in order to increase post-sale revenues. The entertainment gaming industry has little need to improve the skills and performance of game players; therefore, there is no inherent "value" to pursue performance assessment. In comparison, the primary purpose of serious games is to improve the skills and performance of play-learners, so the primary purpose of serious games analytics would be to (a) obtain valuable actionable insights to better the game or learning design, and (b) improve skills and performance of the play-learners to better convince stakeholders of the games' effectiveness. Although profits are also important to the serious games industry, it must play second fiddle to improving skills and

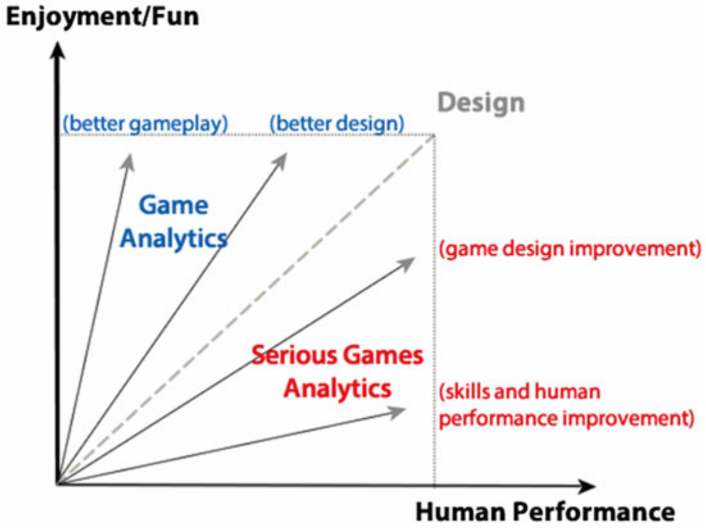


Fig. 1.5 The axes of emphasis in game analytics and serious games analytics: enjoyment, design, and human performance

human performance. Obtaining actionable insights from data is central to serious games analytics, as seen in the following:

- Converting user-generated data into actionable insights (Chap. 5, 9, 10, and 11 of this volume)

Figure 1.5 shows the 3-axis emphasis between game and serious games analytics, it would be a juggling act trying to meet the needs of all three axes, namely enjoyment, design, and human performance. This means that serious games must ask what *changes in behavior* can be affected by the game—i.e., a positive change in behavior will lead to human performance improvement (indicating the serious game's effectiveness). By focusing on making a well-designed serious game with evidence of improved skills and performance, there may not be a need for the serious games industry to monetize (like entertainment games).

In this chapter, we define serious games analytics as the "actionable metrics developed through problem definition in training/learning scenarios and the application of statistical models, metrics, and analysis for skills and human performance improvement and assessment, using serious games as the primary tools for training." Like GA, serious games analytics can be derived from tracing players' gameplay and the visualization of their actions, behaviors, and play-paths within virtual/gaming environments. Unlike game analytics, the primary purpose of serious games analytics is to improve skills and performance of play-learners, which means that gameplay is being relegated to a lower priority.

5.5 *Analytics Differ by Origins and Purposes*

It is extremely important to recognize that the purpose of serious games analytics is very different from that of the game analytics (see Canossa et al., 2013) or learning analytics (see Siemens et al., 2011) because each of these analytics originated from a distinctly different industry and have been created to aid in the business decision-making of the serious fames, entertainment gaming, and online learning industry, respectively.

"Big Data" from learning analytics come from decidedly educational sources, such as online courseware, LMSs, and various Massive Open Online Courses (or MOOCs) because their purpose is "understanding and optimizing learning and the environments in which it occurs" (www.learninganalytics.net). At least for the moment, these sources have yet to include serious games, possibly because it does not (yet) produce Big Data [See also the related field of Education Data Mining (or EDM); Romero, Ventura, Pechenizkiy, & Baker, 2010]. Big Data from game analytics come from the many MMOGs and a myriad of mobile games. Until MMO serious games become available, Big Data for serious games analytics will take a while to arrive (Readers should check out the Kickstart project, called Tyto Online, by Immersed Games at www.kickstarter.com/projects/immersed/tyto-online-learning-mmorpg).

Since all three analytics are based on different metrics, criteria, and purposes, there is no reason to assume the performance metrics for one industry would work or transfer well into another, unless more evidence becomes available. This means more research is necessary to identify new performance metrics, verify existing ones, and clarify the methods for serious games analytics. As the industry moves towards mobile games and MMO serious games, it is important for new performance assessment methods to be scalable to accept Big Data. We would, therefore, caution against methods that are not scalable, or are unable to accept spatial–temporal data found in many of today's serious games (see Loh & Sheng, 2015, Chap. 5 in this volume, for more details). Besides being scalable, these metric and methods must also support skills and human performance measurement, assessment, and improvement.

Conclusion

Serious games that fit the bill as "tools to improve skills and human performance" will not find adoption to be a problem because stakeholders are only wary of the *unproven technologies*. If they can be convinced of the ability of serious games to meet their training needs, as well as a good Return of Investment, serious games would become worthwhile. This means that the (high) production costs of serious games would become a much small factor in the equation.

In his article, *Serious games*, *debriefing*, *and simulation/gaming as a discipline*, the editor of Simulation & Gaming, David Crookall (2010) explained that we need to take *debriefing* seriously if we wish for educational authorities (or stakeholders) to accept serious games as a legitimate source of learning. Through post-training debriefing, instructors can discuss with trainees what they did right/wrong during the simulation or training game, similar to the AAR practice of the military. He affirmed that "the debriefing should be a design consideration right from the start." Because serious game is a computer application, Crookall believed it can "easily include tools and modules of various kinds to collect data transparently during play. The data can then be processed to provide material for feedback during play, as in-game debriefing, and also made available as part of the end-of-game debriefing" (p. 908).

What Crookall has suggested is highly relevant to this book, in terms of (a) *performance measurement* using serious games via in-game (in situ) data collection, (b) *performance assessment* through in-game (ad hoc) and post-game (post hoc) debriefing tools, and *performance improvement* by identifying the good habits and actions that should be retained and cultivated. Play-learners can use the real-time and/or post-training actionable insights/reports from these serious games analytics for (self-) assessment and improvement. Besides retraining and remediation of poor work habits to reduce risks or workplace errors, these analytics and actionable insights can also be used as "feedback" to the serious games developers, for design improvement to create even better games in the future.

Crookall suggested that because debriefing tools (back-end) can appear "less sexy than the flashy game ware" (front-end), this explains why funders, who do not understand "learning comes from processing the game experience" (p. 908), would refuse to pay for debriefing tools (i.e., human performance assessment) because they see them as irrelevant or useless code. What Crookall observed are issues that plagued serious games. This is why *Serious Games Analytics* is needed and the reason for us to highlight the importance of skills and human performance improvement with serious games.

The latest serious games market report (Ambient Insight, 2013) revealed that the future of serious games lies in the mobile sector. The report further spoke of new market opportunities in: (1) location-based learning games, (2) mobile augmented reality games, and (3) mobile learning value-added services. It should be obvious to the readers that the mobile sector thrives on telemetry and user-generated data! With new data, there will be a renewed need to gather, combine, understand, and predict what is to come—moving gradually from data-driven assessment to predictive assessment (Kay & van Harmelen, 2012).

Changes are coming our way in the form of medical simulation/serious games, MMO serious games, and mobile serious games. Serious games researchers and developers can expect a new deluge of demand for serious games analytics. The need for innovative methodologies that can produce actionable insights for human performance measurement, assessment, and improvement is just beginning.

References

Abt, C. C. (1987). *Serious games*. Lanham, MD: University Press of America (Reprint).

Advances in game-based learning. Geneva, Switzerland: Springer. doi:10.1007/978-3-319-05834-4_5.

Aldrich, C. (2005). *Learning by doing: A comprehensive guide to simulations, computer games, and pedagogy in e-learning and other educational experiences*. San Francisco: Pfeiffer.

Alvarez, J., Djaouti, D., Rampnoux, O., & Alvarez, V. (2011). *Serious games market: Some key figures* (*from 1950's to 2000's*). Retrieved October 22, 2014, from http://serious.gameclassification.com/files/articles/sgc_report_03-11.pdf

Ambient Insight. (2013). *2013-2018 North America mobile edugame market*, Monroe, WA. Retrieved December 12, 2014, from http://www.ambientinsight.com/Resources/Documents/Ambient-Insight-2013-2018-North-America-Mobile-Edugame-Market-Abstract.pdf

BankersLab. (2013). *A smart guide to serious gaming*. Retrieved December 22, 2014, from http://bankerslab.com/blogposts/a-smart-guide-to-serious-gaming-part-1/

Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., & Berta, R. (2013). Assessment in and of serious games: An overview. *Advances in Human-Computer Interaction, 2013*, 11. doi:10.1155/2013/136864.

Berland, M., Baker, R. S., & Blikstein, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning, 19*(1–2), 205–220. doi:10.1007/s10758-014-9223-7.

Byun, J. H., & Loh, C. S. (2015). Audial engagement: Effects of game sound on learner engagement in digital game-based learning environments. *Computers in Human Behavior, 46*, 129–138. doi:10.1016/j.chb.2014.12.052.

Canossa, A., Seif El-Nasr, M., & Drachen, A. (2013). Benefits of game analytics: Stakeholders, contexts and domains. In M. Seif El-Nasr, A. Drachen, & A. Canossa (Eds.), *Game analytics: Maximizing the value of player data* (pp. 41–52). London: Springer. doi:10.1007/978-1-4471-4769-5.

Caspian Learning. (2010). *Improving navy recruits' performance: A serious games study.* Sunderland, England. Retrieved December 12, 2014, from http://www.caspianlearning.co.uk/

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research, 53*(4), 445–459. doi:10.3102/00346543053004445.

Clark, R. E. (1985a). Confounding in educational computing research. *Journal of Educational Computing Research, 1*(2), 137–148.

Clark, R. E. (1985b). Evidence for confounding in computer-based instruction studies: Analyzing the meta-analyses. *Educational Technology Research and Development, 33*(4), 249–262.

Clark, R. E. (1992). Dangers in the evaluation of instructional media. *Academic Medicine, 67*(12), 819–820.

Clark, R. E. (1994). Media will never influence learning. *Educational Technology Research and Development, 42*(2), 21–29. doi:10.1007/BF02299088.

Crookall, D. (2010). Serious games, debriefing, and simulation/gaming as a discipline. *Simulation & Gaming, 41*(6), 898–920. doi:10.1177/1046878110390784.

Cukier, K. N., & Mayer-Schoenberger, V. (2013). The rise of big data: How it's changing the way we think about the World. *Foreign Affairs, 92*(3), 28–40. Retrieved December 12, 2014, from http://www.foreignaffairs.com/articles/139104/kenneth-neil-cukier-and-viktor-mayer-schoenberger/the-rise-of-big-data.

DeRosier, M. E., Craig, A. B., & Sanchez, R. P. (2012). Zoo U: A stealth approach to social skills assessment in schools. *Advances in Human-Computer Interaction*, *2012*. doi:10.1155/2012/654791.

Djaouti, D., Alvarez, J., Jessel, J.-P., & Rampnoux, O. (2011). Origins of serious games. In M. Ma, A. Oikonomou, & L. C. Jain (Eds.), *Serious games and edutainment applications* (pp. 25–43). London: Springer. doi:10.1007/978-1-4471-2161-9_3.

Domínguez, A., Saenz-de-Navarrete, J., de-Marcos, L., Fernández-Sanz, L., Pagés, C., & Martínez-Herráiz, J.-J. (2013). Gamifying learning experiences: Practical implications and outcomes. *Computers & Education, 63*, 380–392. doi:10.1016/j.compedu.2012.12.020.

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology, 17*(2), 245–60. doi:10.1023/A:1019637632584.

Drager, N. (2014). Gamification, game-based learning, and serious games: What's the difference? Retrieved June 09, 2014, from http://www.roninsc.com/blog/2014/05/08/gamification-game-based-learning-and-serious-games-whats-the-difference/

Foundation of American Scientists. (2006). *Summit on educational games: Harnessing the power of video games for learning*. Washington, DC.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy* (1st ed.). New York: Palgrave Macmillan.

Gerber, H. R. (2012). *Can education be gamified?: Examining gamification, education, and the future*. Retrieved June 4, 2014, from http://www.academia.edu/2235680/Can_Education_be_Gamified_Examining_Gamification_Education_and_the_Future

Grimshaw, M., Lindley, C. A., & Nacke, L. (2008). Sound and immersion in the first-person shooter: Mixed measurement of the player's sonic experience. In *Proceedings of the Audio Mostly Conference*. Retrieved December 12, 2014, from http://wlv.openrepository.com/wlv/bitstream/2436/35995/2/Grimshaw_CGAMES07.pdf

Hastings, N. B., & Tracey, M. W. (2004). Does media affect learning: Where are we now? *TechTrends, 49*(2), 28–30. doi:10.1007/BF02773968.

Herold, B. (2013). Researchers see video games as testing, learning tools. *Education Week, 32*(37), 14–15.

Hughes, A. (2014). In serious games, analytics are everything! *Learning Solution Magazine*. Retrieved December 12, 2014, from http://www.learningsolutionsmag.com/articles/1509/in-serious-games-analytics-are-everything

Ifenthaler, D. (2015). Learning analytics. In J. M. Spector (Ed.), *The SAGE encyclopedia of educational technology*. Thousand Oaks, CA: Sage.

Ifenthaler, D., Eseryel, D., & Ge, X. (Eds.). (2012). *Assessment in game-based learning: Foundations, innovations, and perspectives*. New York: Springer. doi:10.1007/978-1-4614-3546-4.

Ifenthaler, D., & Widanapathirana, C. (2014). Development and validation of a learning analytics framework: Two case studies using support vector machines. *Technology, Knowledge and Learning, 19*(1–2), 221–240. doi:10.1007/s10758-014-9226-4.

IJsselsteijn, W., de Kort, Y., Poels, K., Jurgelionis, A., & Bellotti, F. (2007). Characterising and measuring user experiences in digital games. In *Proceedings of the International Conference on Advances in Computer Entertainment Technology* (pp. 27–30).

Kay, D., & van Harmelen, M. (2012). Analytics for the whole institution: Balancing strategy and tactics. *JISC CETIS, 1*(2), 1–31.

Kirriemuir, J., & McFarlane, A. (2003). Use of computer and video games in the classroom. In *Proceedings of the Level up Digital Games Research Conference*. Utrecht, The Netherlands: Universiteit Utrecht.

Kozma, R. B. (1991). Learning with media. *Review of Educational Research, 61*, 179–221. doi:10.3102/00346543061002179.

Kozma, R. B. (1994). Will media influence learning? Reframing the debate. *Educational Technology Research and Development, 42*(2), 7–19.

Krulak, Charles. (1997). *Military thinking and decision making exercises* (No. 1500.55). Washington, DC. Retrieved December 12, 2014, from http://www.marines.mil/Portals/59/Publications/MCO 1500.55.pdf

Laamarti, F., Eid, M., & El Saddik, A. (2014). An overview of serious games. *International Journal of Computer Games Technology, 2014*, 15. doi:10.1155/2014/358152.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review, 52*(2), 21–31.

Loh, C. S. (2006). Designing online games assessment as “information trails”. In D. Gibson, C. Aldrich, & M. Prensky (Eds.), *Games and simulation in online learning: Research and development frameworks* (pp. 323–348). Hershey, PA: Idea Group. doi:10.4018/978-1-59904-304-3.ch016.

Loh, C. S. (2012a). Improving the impact and return of investment of game-based learning. *International Journal of Virtual and Personal Learning Environments, 4*(1), 1–15. doi:10.4018/jvple.2013010101.

Loh, C. S. (2012b). Information trails: In-process assessment of game-based learning. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 123–144). New York: Springer. doi:10.1007/978-1-4614-3546-4.

Loh, C. S., Anantachai, A., Byun, J. H., & Lenox, J. (2007). Assessing what players learned in serious games: In situ data collection, information trails, and quantitative analysis. In Q. Mehdi (Ed.), *Proceedings of the Computer Games: AI, Animation, Mobile, Educational & Serious Games Conference* (*CGAMES 2007*)*, Louisville, KY* (pp. 10–19). Wolverhampton, England: University of Wolverhampton.

Loh, C. S., & Sheng, Y. (2015). Measuring expert-performance for serious games analytics: From data to insights. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: Methodologies for performance measurement, assessment, and improvement*. Advances in Game-Based Learning. Springer International Publishing Switzerland. doi: 10.1007/978-3-319-05834-4_5

Long, P. D., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review, 46*(5), 31–40.

McLeroy, C. (2008a, September). History of military gaming. *Soldiers Magazine,* 4–6. Retrieved December 12, 2014, from http://www.army.mil/article/11936/History_of_Military_gaming/

McLeroy, C. (2008b, September). Improving “America’s army”. *Soldiers Magazine*, 7–9. Retrieved December 12, 2014, from http://www.army.mil/article/11935/

Medler, B. (2011). Player dossiers: Analyzing gameplay data as a reward. *Game Studies, 11*(1). Retrieved December 12, 2014, from http://gamestudies.org/1101/articles/medler

Michael, D., & Chen, S. (2005). *Proof of learning: Assessment in serious games.* Retrieved September 22, 2014, from http://www.gamasutra.com/view/feature/2433/proof_of_learning_assessment_in_.php

Michael, D., & Chen, S. (2006). *Serious games: Games that educate, train, and inform*. Boston: Thomson Course Technology PTR.

Miller, R. (2013). *Facebook builds exabyte data centers for cold storage.* Retrieved June 8, 2014, from http://www.datacenterknowledge.com/archives/2013/01/18/facebook-builds-new-data-centers-for-cold-storage/

Moshirnia, A. (2007). The educational potential of modified video games. In *Proceedings of InSITE 2007: Informing Science and Information Technology Conference* (pp. 511–521), Ljubljana, Slovenia.

Nickols, F. W. (2005). Why a stakeholder approach to evaluating training. *Advances in Developing Human Resources, 7*(1), 121–134. doi:10.1177/1523422304272175.

Preimesberger, C. (2014). *Apple unveils iOS 8 with improved Siri, security, health care features.* Retrieved June 7, 2014, from http://www.eweek.com/mobile/slideshows/apple-unveils-ios-8-with-improved-siri-security-health-care-features.html

Prensky, M. (2001). *Digital game-based learning*. New York: McGraw-Hill.

Rao, R. (2003). From unstructured data to actionable intelligence. *IT Professional, 5*(6), 29–35. doi:10.1109/MITP.2003.1254966.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. J. D. (Eds.). (2010). *The handbook of educational data mining* (1st ed.). Boca Raton, FL: CRC Press.

Salomon, G., & Clark, R. E. (1974). Re-examining the methodology of research on media and technology in education. *Review of Educational Research, 47*, 99–120.

Sandford, R., & Williamson, B. (2005). *Games and learning: A handbook*. Bristol, England: FutureLab.

Sawyer, B., & Rejeski, D. (2002). *Serious games: Improving public policy through game-based learning and simulation*. Washington, DC.

Sawyer, B. (2009). Foreword: From virtual U to serious game to something bigger. In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. xi–xvi). New York: Routledge.

Schaffhauser, D. (2014). *Students blow through wind power in free virtual lab*. Retrieved June 4, 2014, from http://thejournal.com/articles/2014/05/29/students-blow-through-wind-power-in-free-virtual-lab.aspx

Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Shum, S. B., & Ferguson, R., et al. (2011). *Open learning analytics: An integrated & modularized platform.* Retrieved December 12, 2014, from http://solaresearch.org/OpenLearningAnalytics.pdf

Söderber, H. (1977). *Den allvarsamma leken ("The serious game")*. Stockholm, Sweden: LiberForlag (Reprint).

Thawonmas, R., & Iizuka, K. (2008). Visualization of online-game players based on their action behaviors. *International Journal of Computer Games Technology, 2008*, 1–9. doi:10.1155/2008/906931.

Turse, N. (2003). *Zap, zap, you're dead…*. Retrieved June 4, 2014, from http://www.tomdispatch.com/index.mhtml?pid=1012

Van Eck, R. (2006). Digital game-based learning: It's not just the digital natives who are restless…. *EDUCAUSE Review, 41*(2), 16–30. Retrieved Sep 12, 2014 from http://www.educause.edu/ero/article/digital-game-based-learning-its-not-just-digital-natives-who-are-restless.

Wallner, G., & Kriglstein, S. (2012). A spatiotemporal visualization approach for the analysis of gameplay data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (*CHI'12*) (pp. 1115–1124). New York: ACM Press. doi:10.1145/2207676.2208558.

Watkins, R., Leigh, D., Foshay, R., & Kaufman, R. (1998). Kirkpatrick plus: Evaluation and continuous improvement with a community focus. *Educational Technology Research and Development, 46*(4), 90–96. doi:10.1007/BF02299676.

Williams, D., Yee, N., & Caplan, S. (2008). Who plays, how much, and why? A behavioral player census of a virtual World. *Journal of Computer Mediated Communication, 13*(4), 993–1018.

Zichermann, G., & Cunningham, C. (2011). *Gamification by design: Implementing game mechanics in web and mobile apps*. New York: O'Reilly Media.

Zoeller, G. (2013). Game development telemetry in production. In M. Seif-El-Nasr, A. Drachen, & A. Canossa (Eds.), *Game analytics: Maximizing the value of player data* (pp. 111–136). London: Springer.

Zyda, M. (2005). From visual simulation to virtual reality to games. *Computer, 38*(9), 25–32. doi:10.1109/MC.2005.297.

Chapter 2
A Meta-Analysis of Data Collection in Serious Games Research

Shamus P. Smith, Karen Blackmore, and Keith Nesbitt

Abstract Serious game analytics share many of the challenges of data analytics for computer systems involving human activity. Key challenges include how to collect data without influencing its generation, and more fundamentally, how to collect and validate data from humans where a primary emphasis is on what people are thinking and doing. This chapter presents a meta-analysis of data collection activities in serious games research. A systematic review was conducted to consider metrics and measures across the human–computer interaction, gaming, simulation, and virtual reality literature. The review focus was on the temporal aspect of data collection to identify if data is collected before, during, or after gameplay and if so what fundamental processes are used to collect data. The review found that the majority of data collection occurred post-game, then pre-game, and finally during gameplay. This reflects traditional difficulties of capturing gameplay data and highlights opportunities for new data capture approaches oriented towards data analytics. Also we identify how researchers gather data to answer fundamental questions about the efficacy of serious games and the design elements that might underlie their efficacy. We suggest that more standardized and better-validated data collection techniques, that allow comparing and contrasting outcomes between studies, would be beneficial.

Keywords Data collection • Serious games • Meta-review • Data analytics

S.P. Smith (✉)
School of Electrical Engineering and Computer Science, The University of Newcastle, University Drive, Callaghan, NSW 2308, Australia
e-mail: shamus.smith@newcastle.edu.au

K. Blackmore • K. Nesbitt
School of Design, Communication and Information Technology,
The University of Newcastle, University Drive, Callaghan, NSW 2308, Australia
e-mail: karen.blackmore@newcastle.edu.au; keith.nesbitt@newcastle.edu.au

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based Learning, DOI 10.1007/978-3-319-05834-4_2

1 Introduction

This chapter is concerned with serious games and the data collected about serious games. It concerns itself with the process of turning such data into information and, through data analytics, drawing conclusions about that information. Thus, collected data is synthesized as information and ultimately into evidence.

There has been significant growth in research and industry attention, and public awareness, of the collection and analysis of the vast amounts of data that is available electronically. This has characterized the popularization of "Big Data" and the associated knowledge and value to be generated through both automated and perceptual techniques for data analytics. At one end of the data analytics pipeline is the notion of visual analytics, the use of visualization techniques to represent multidimensional data so that people can use their perceptual skills and incomplete heuristic knowledge to find useful patterns in the data. We might term these patterns information, and we note that these approaches rely on human skills that do not necessarily translate well to computers. More automated approaches often referred to as "data-mining" also exist. While the intention is the same, that is to identify useful patterns or information, the approach is complementary, relying on the strengths of computers to perform rapid, repetitive, error-free mathematical tasks that allow large amounts of data to be quickly processed.

Given that neither of these approaches to finding information in large data is discrete, good data analytics might well rely on combining the strengths of both approaches. However, information does not imply evidence. The accuracy of the information very much depends on the quality and validity of the data and the transformations that filter, abstract, and simplify the vast volumes of data to support analysis. If poor quality data is initially collected, then its progress through later stages of the analytics pipeline will be compromised and the validity of any identified patterns weakened.

Serious game analytics share many of the same challenges as data analytics in other computer systems that are focused on human activity. A typical challenge is how to collect data without influencing its generation and more fundamentally, how to collect and validate data from human participants where a primary focus is on what people are thinking and doing.

This chapter will explore data collection issues from serious games as the initial step to any serious gameplay analytics. We use a systematic review process to consider the metrics and measures across the human–computer interaction, gaming, simulation, and virtual reality literature. We identify how researchers gather data to try and answer fundamental questions about the efficacy of serious games and the design elements that might underlie their efficacy.

Data collection is interdisciplinary and a comprehensive literature review over computer science, psychology, and education, for instance, is outside the scope of this chapter. The focus here will be on the temporal aspect of data collection during serious game studies, namely how, and if, data is collected before, during, and after serious gameplay. The chapter uses a framework of traditional data collection methods to identify a core mapping to the serious game literature. The study is broad in that it covers diverse research from numerous disciplines, over a long time frame, that have used a wide range of methods and been driven by different motivations. This diverse body of research is first found by systematically selecting eight relevant literature reviews from 2009 to 2014 related to serious game research. To provide depth to the study, each of the study papers identified in these literature reviews ($n=299$) are examined in terms of the way data is collected to assess the efficacy and usability of the games.

While the enthusiasm for serious games is unquestioned, the business case for serious games still requires more tangible evidence, both qualitative and quantitative. However, a first step to better evidence is a close examination of the data collected from serious games. It is this data that will be processed by any serious game analytics, and ultimately demonstrate the worth of the source serious games. This chapter provides a historical review of data collection as a resource for researchers in serious games, human–computer interaction, and anyone who is concerned about the collection and accuracy of gameplay data for future analytic purposes. Also, in the discussion section of this chapter, we will reflect upon the question of evidence and how well it relates to the two key issues of efficacy and usability in games that are used for serious purposes.

2 Study Method

To perform our study, we designed a systematic process that could be repeated or amended to accommodate both changes in scope and alternative research questions, and extended to incorporate future literature (see Fig. 2.1).

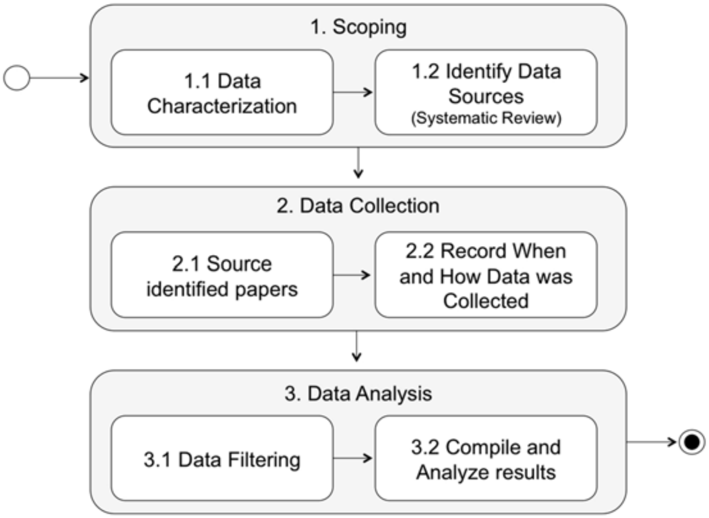

1. Scoping
1.1 Data Characterization
1.2 Identify Data Sources
(Systematic Review)
2. Data Collection
2.1 Source identified papers
2.2 Record When and How Data was Collected
3. Data Analysis
3.1 Data Filtering
3.2 Compile and Analyze results

Fig. 2.1 Overview of the process used in this study

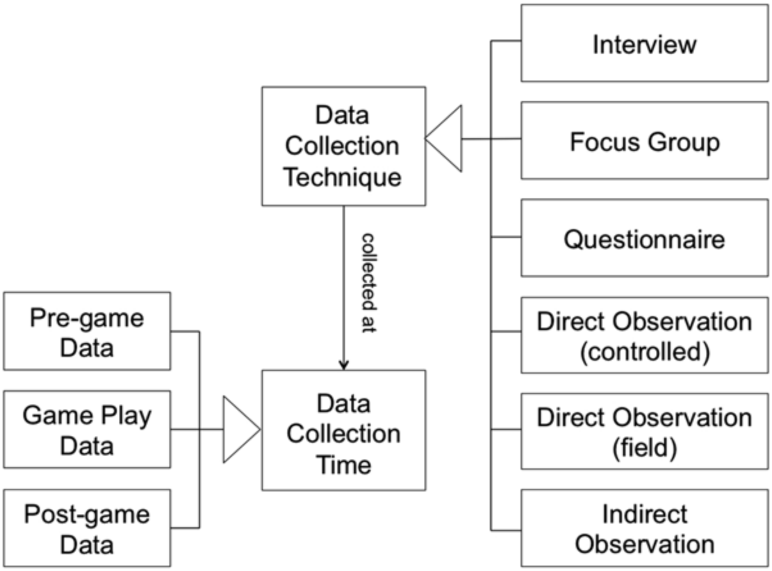


Fig. 2.2 Overview of the data characterization used in this study

2.1 Data Characterization

The process began with a data characterization activity where we identified the type of data we wished to collect about each individual game study (see 1.1 in Fig. 2.1). We focused on the temporal aspect of data collection during serious game studies, namely how, and if, data is collected before, during, and after serious gameplay (see Fig. 2.2). These stages are relatively standard across the range of methodological approaches used to evaluate games.

We decided to categorize the data collected during each of these stages based on common data gathering techniques (see Fig. 2.2). These frequently used techniques are taken from a list provided by Rogers, Sharp, and Preece (2011) in their popular text on human–computer interaction, and include:

- Interviews
- Focus Groups
- Questionnaires
- Direct observation in the field
- Direct observation in a controlled environment
- Indirect observation

Although the first three techniques are self-explanatory, the last three techniques require further clarification. *Direct observation* consists of observing actual user activity and typically involves the collection of qualitative data by capturing the details of what individual or groups are doing with a system, for example, with observers taking notes of user behaviors. This can be conducted *in the field*, where users are interacting in the target environment for a system. Examples could include a normal teaching session in a classroom for an educational system or on a walking tour for a mobile application (Rogers et al., 2011, p. 262). Thus, the system is being used in a real-life situation, e.g., with high ecological validity (McMahan, Ragan, Leal, Beaton, & Bowman, 2011; Smith, Stibric, & Smithson, 2013). Direct observation *in a controlled environment* is typically a laboratory-based environment where conditions can be controlled and standardized between participants and sessions. This allows users to focus on a task without interruption. However, results from studies in such environment may not generalize as the conditions are, by default, artificial.

In contrast to direct observation, where the users can see that they are actively being observed, *indirect observation* involves gathering data where users are not distracted by the collection mechanism. This could include collecting qualitative data, for example, from a user diary, or quantitative data from automated event logging. The latter is particularly attractive for serious games as event logs can be tailored to collect any pertinent information; for example, task sequences, task completion times, and/or percentage of tasks accomplished. Loh (2009) details a number of logging examples including basic game event logs, After-Action Reports as graphical game logs, and biofeedback data to capture physiological reactions. Such *in-process* data collection is by its nature objective and can provide substantial volumes of data for further analytic treatment.

While these techniques cover a good range of the mixed methods used in game research, we also recognize that other categorizations could have been adopted. For example, an alternative and more detailed classification of 16 different data collection techniques used in games studies is provided by Mayer et al. (2014). While there is merit for more complex categorization, we recognized the difficulty of collecting our own data; game studies from various disciplines do not have a standardized approach to describing data collection methods. Since we intended to be reviewing a large and broad range of studies, we sought to keep our data classification as simple as possible. Thus focusing on specific techniques, for example, the use of telemetry or *Information Trails* (Loh, 2012) for indirect observation of gameplay, is outside the scope of the current review. However, we will revisit issues surrounding the data collection process in the discussion section.

2.2 Identify Data Sources (Systematic Review)

We adopted a systematic approach to identifying existing reviews of serious game research across domains (see 1.2 in Fig. 2.1). A systematic review is developed to gather, evaluate, and analyze all the available literature relevant to a particular research question or area of interest, based on a well-defined process (Bearman et al., 2012; González, Rubio, González, & Velthuis, 2010; Kitchenham et al., 2009). The systematic review methodology is extensively used in the healthcare domain (Bearman et al., 2012) and has been widely adopted in other areas including business (González et al., 2010), education (Bearman et al., 2012), and software engineering (Kitchenham et al., 2009; Šmite, Wohlin, Gorschek, & Feldt, 2010).

A systematic review methodology requires the identification of all published works relevant to the requirements. The search strategy adopted covers key term searches in relevant scholarly databases. We included the Web of Science, Scopus, EBSCOhost, and Wiley Interscience bibliographic databases in the search. The search was conducted over article titles to restrict results to primary studies, and includes journal articles, book chapters, and review papers in the results. Search results were restricted to papers published between 2009 and 2014 inclusive.

The objective of our systematic review was to identify all review articles of studies using serious games. To conduct a review that meets our objective, the search term used needs to accommodate two key purposes. The first purpose was to find published works relating to serious games. We expanded the term, *serious games*, to include references to studies of games for *applied*, *learning*, *teaching*, or *educational* purposes (Crookall, 2010). The second purpose is to find review or meta-review articles only as the basis for "drill-down" to individual studies. We therefore included the terms *review*, *meta-review*, or *meta-analysis* in the search term. Several preliminary searches were conducted to refine the individual and combined search terms to develop a search string that located articles of interest without too many false positives. The resulting Boolean search string that we used for the systematic review was:

((*gam* AND* (*serious OR edutainment OR* "*applied gam**" *OR learn* OR game-based learning OR educat* OR teach**) *AND* (*review OR meta-review OR meta-analysis*))

The search initially produced a total of 126 potential papers, of which 73 were found to be unique. These papers were then manually evaluated by title, abstract, and if necessary, by full text, based on the following inclusion criteria:

- Focused on the review of studies:
  - Using randomized control trials, experimental pretest/posttest control group design or quasi-experimental structure
  - Evaluating computer, console or mobile games
  - That was directed at achieving teaching and learning outcomes
- From any country
- Written in English

Papers not meeting the inclusion criteria were excluded from the systematic review. The review process is shown in Fig. 2.3 and identified ten papers for the analysis.

The evaluation process detailed in Fig. 2.3 shows the inclusion of an additional paper that did not appear in the initial 126 papers. Wattanasoontorn, Boada, García, and Sbert's (2013) comprehensive study of serious games for health was not located using the Boolean search string due to the non-inclusion of the term *review* in the article title. It was, however, identified and noted in the preliminary searches that we used to refine the search terms. Wattanasoontorn et al. (2013) include 108 references in the broad health domain in their final review, making this a relevant and comprehensive piece of work for inclusion in our analysis. However, expanding the search string to ensure that this article was located results in an unwieldy number of

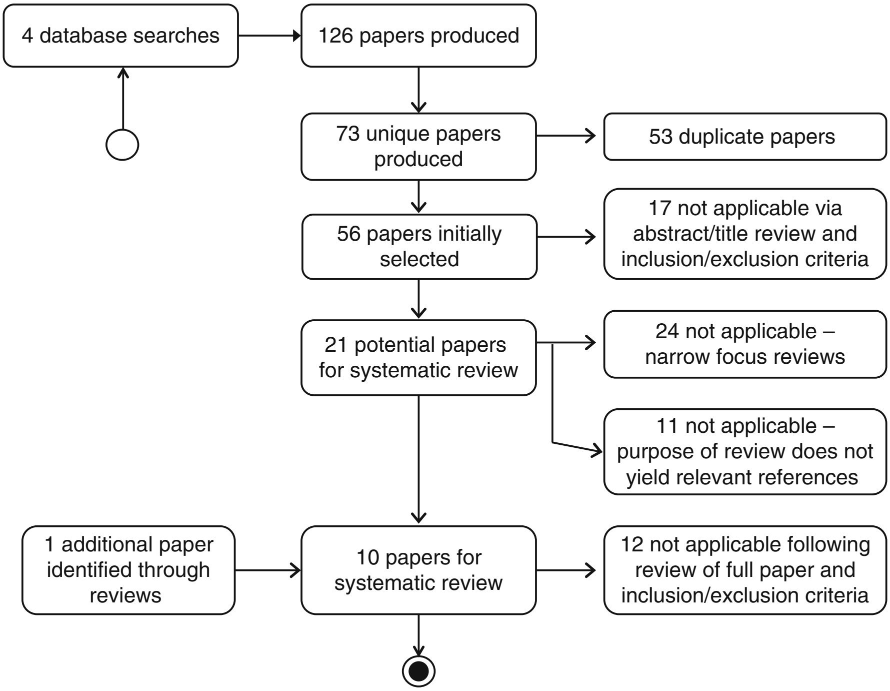

4 database searches
126 papers produced
73 unique papers produced
53 duplicate papers
56 papers initially selected
17 not applicable via abstract/title review and inclusion/exclusion criteria
21 potential papers for systematic review
24 not applicable – narrow focus reviews
11 not applicable – purpose of review does not yield relevant references
1 additional paper identified through reviews
10 papers for systematic review
12 not applicable following review of full paper and inclusion/exclusion criteria

Fig. 2.3 Diagram of the selection process for the systematic review

irrelevant search results. More importantly, it results in a search term that does not meet the objective of the systematic review, which is to identify reviews or meta-reviews of serious games. Thus, the paper was simply added to the end results of the systematic review process.

2.3 Data Collection and Analysis

During the data collection process, we examined in detail the ten high level literature reviews identified in our systematic review. A description of each of these papers is provided in the next section of the chapter.

Two literature review papers, on analysis, were rejected during the data collection process. Hwang and Wu (2012) analyzed the research status and trends in digital game-based learning (DGBL) from 2001 to 2010. Specifically, they explored (1) whether the number of articles in this area is increasing or decreasing, (2) what the primary learning domains related to DGBL are, (3) whether there is a domain focus shift between the first 5 years (2001–2005) and the second 5 years (2006–2010), and (4) which are the major contributing countries of DGBL research. From an initial set of 4,548 papers, Hwang and Wu selected a total of 137 articles for review. However, their paper does not provide details of the specific 137 articles selected, and we were therefore unable to identify data collection methods from these papers. Thus, we have excluded this review from our analysis. The Blakely, Skirton, Cooper, Allum, and Nelmes (2009) systematic review of educational games in the health sciences was also removed. They provide an analysis of the use of games to support classroom learning in the health sciences based on a review of 16 papers. However, it was deemed to be an earlier subset of the latter and more expansive review of serious games for health by Wattanasoontorn et al. (2013).

From the remaining eight literature review papers, we identified 299 referenced studies. Where possible, we then sourced each of the papers and recorded the data collection techniques used pre-game, during gameplay and post-game for each study. In a few cases, papers could not be sourced or the papers were not in English. These papers were excluded, as were all studies that were only reported in non-peer reviewed locations such as websites. We also excluded references to demonstrations or papers that only included a critical analysis of literature. Finally, we excluded duplicate studies so they were only included once in the analysis. This left a total of 188 referenced studies to be included in the final analysis.

3 Systematic Review Papers

The eight review papers identified from our systematic process that were used for data collection covered both general and domain-specific areas. Four of the papers are reviews of the general serious games area and were not focused on any specific area or application. However, two papers are focused on studies in the health domain, while the other two focused on medicine and the humanities. Each of these papers are described below and the number of contributory studies to our review are identified.

Connolly, Boyle, MacArthur, Hainey, and Boyle (2012) examined the "literature on computer games and serious games in regard to the potential positive impacts of gaming on users aged 14 years or above, especially with respect to learning, skill enhancement and engagement" (p. 1). This review paper focused on identifying empirical evidence and on categorizing games and their impacts and outputs. The majority of reviewed papers come under the serious games for education and training classification. We have reviewed data collection in these papers ($n=70$) across Connolly et al.'s categories of affective and motivational outcomes, behavioral change outcomes, knowledge acquisition/content understanding outcomes, motor skill outcomes, perceptual and cognitive skills outcomes, physiological arousal outcomes, and soft skill and social outcomes.

Wattanasoontorn et al. (2013) consider the use of serious games in the health domain area. They provide a survey of serious games for health and define a new classification, based on serious game, health, and player dimensions. For serious game subjects, they classify by game purpose and game functionality, for health, they classify by state of disease and finally for player, they consider two types of player dimensions, player/non-player, and professional/nonprofessional. We have used Wattanasoontorn et al.'s (2013) classification and comparison of health games summary ($n = 91$) which considers the following areas: detection (patient), treatment (patient), rehabilitation (patient), education (patient), health and wellness (non-patient), training for professional (non-patient), training for non-professional (non-patient).

Anderson et al. (2010) describe the use of serious games for cultural heritage; specifically, the use of games to support history teaching and learning and for enhancing museum visits. Their state-of-the-art review includes both a set of case studies and an overview of the methods and techniques used in entertainment games that can potentially be deployed in cultural heritage contexts. Here, we have focused on the former and reviewed data collection as noted in the case studies ($n = 5$).

Girard, Ecalle, and Magnan (2013) review the results of experimental studies designed to examine the effectiveness of video games and serious games on players' learning and engagement. They have attempted to identify all the experimental studies from 2007 to 2011 that have used serious games for training or learning, and assessed their results in terms of both effectiveness and acceptability. Girard et al. (2013) had a two pass process for article inclusion/exclusion where the stricter second pass, only considering randomized controlled trial studies, resulted in only nine articles. Here, we have used the results from their first pass of the literature which resulted in 30 articles ($n = 29$, we excluded one article written in French) published in scientific journals or in proceedings of conferences and symposia across the fields of cognitive science, psychology, human–computer interaction, education, medicine, and engineering where training has been performed using serious games or video games.

The systematic review of Graafland, Schraagen, and Schijven (2012) provides a comprehensive analysis of the use of serious games for medical training and surgical skills training. The authors focus on evaluating the validity testing evident in prior serious games research in the area and identify 25 articles through a systematic search process. Of these, 17 included games developed for specific educational purposes and 13 were commercial games evaluated for their usefulness in developing skills relevant to medical personnel. Of the 25 articles identified by Graafland et al. (2012), six were identified as having completed some validation process and none were found to have completed a full validation process. For the purpose of our study, we considered only articles explicitly identified by Graafland et al. (2012), that appeared in the supplementary information tables ($n = 20$).

Papastergiou (2009) presents a review of published scientific literature on the use of computer and video games in Health Education (HE) and Physical Education (PE). The aim of the review is to identify the contribution of incorporating electronic games as educational tools into HE and PE programs, to provide a synthesis of empirical evidence on the educational effectiveness of electronic games in HE and PE, and to scope out future research opportunities in this area. Papastergiou (2009) notes that the empirical evidence to support the educational effectiveness of electronic games in HE and PE is limited, but that the findings presented in their review show a positive picture overall. We have reviewed data collection methods in the research articles featured in this review ($n = 19$).

Vandercruysse, Vandewaetere, and Clarebout (2012) conducted a systematic literature review where the learning effects of educational games are studied in order to gain more insights into the conditions under which a game may be effective for learning. They noted that although some studies reported positive effects on learning and motivation, this was confounded by different learner variables and different context variables across the literature. Their review initially found 998 unique peer reviewed articles. After removing articles with (quasi) experimental research, only 22 journal articles were finally reviewed. It is these 22 articles that we have included in our data collection review.

Wilson et al. (2009) performed a literature review of 42 identified studies and examined relationships between key design components of games and representative learning outcomes expected from serious games for education. The key design components of fantasy, rules/goals, sensory stimuli, challenge, mystery, and control considered by Wilson et al. (2009) were identified as statistically significant for increasing the "game-like" feel of simulations (Garris & Ahlers, 2001) and key gaming features necessary for learning (Garris, Ahlers, & Driskell, 2002). These were examined in relation to both skill-based and affective learning outcomes. We included all 42 studies in our review.

4 Results

The total number of papers used in this study for data collection was 299. Table 2.1 provides a full list of the references examined, and the literature reviews from which those papers were sourced.

After examination of the 299 papers, and the filtering described in Sect. 2.3, we explored the data collection techniques described in 188 papers spanning 1981–2012. Eighty-four percent of the 188 papers were from the 10 year period of 2003–2012 (see Table 2.2). Also, 51 % of the papers were from the mid-region of this 10 year period, i.e., 2006–2009. However, this does not necessarily indicate a surge in serious game research but is more likely a consequence of publication time frames. Although the literature reviews determined by our search string were published up to 2014, the published research that they reported on was only up to 2012.

In total, 510 data collection techniques were used in the 188 studies. Of these, 33 % of data collection occurred pre-game, 21 % during gameplay, and 46 % in post-game evaluation phases (see Fig. 2.4). On average the total number of data collection methods used per study, across the three phases of pre-game, during gameplay, and post-game, was 2.71 (SD = 1.2).

In terms of specific techniques for the pre-game phase ($n = 169$), 52 % of the studies used questionnaires, 42 % of the studies used some form of test, 4 % of the studies used an interview, 2 % of the studies used an indirect observation, while only a single study employed a focus group in the pre-game phase (see Fig. 2.5).

For the post-game phase ($n = 235$), 46 % of the studies used questionnaires, 37 % of the studies used some form of test, and 13 % of the studies used an interview.

Table 2.1 List and source of all references examined in the data collection process

Anderson, **E**., **F**., **McLoughlin**, **L**., **Liarokapis**, **F**., **Peters**, **C**., **Petridis**, **P**., & **de Freitas**, **S**. (**2010**). Developing serious games for cultural heritage: A state-of-the-art review, *Virtual Reality, 14*(4), 255–275. (*n*=5)
Arnold, D., Day, A, Glauert, J., Haegler, S., Jennings, V., Kevelham, B., Laycock, R., Magnenat-Thalmann, N., Mam, J., Maupu, D., Papagiannakis, G., Thalmann, D., Yersin, B., & Rodriguez-Echavarria, K. (2008); Debevec, P. (2005); Frischer, B. (2008); Gaitatzes, A., Christopoulos, D., & Papaioannou, G. (2004); Jacobson, J., Handron, K., & Holden, L. (2009)
Connolly, **T. M**, **Boyle**, **E. A**., **MacArthur**, **E**., **Hainey**, **T**., **Boyle**, & **J. M**. (**2012**). A systematic literature review of empirical evidence on computer games and serious games, *Computers & Education, 59*(2), September 2012, 661–686. (*n*=70)
Akkerman, S., Admiraal,W., & Huizenga, J. (2008); Anand, V. (2007); Assmann, J. J., & Gallenkamp, J. V. (2009); Aylon, Y., Glaser, C. B., Hall, J. I., Uribe, S., & Fried, M. P. (2005); Backlund, P., Engström, H., Johannesson, M., Lebram, M., & Sjödén, B. (2008); Baldaro, B., Tuozzi, G., Codispoti, M., Montebarocci, O., Barbagli, F., Trombini, E., et al. (2004); Barlett, C. P., Vowels, C. L., Shanteau, J., Crow, J., & Miller, T. (2009); Beale, I. L., Kato, P. M., Marin-Bowling, V. M., Guthrie, N., & Cole, S.W. (2007); Boot, W. R., Kramer, A. F., Simons, D. J., Fabiani, M., & Gratton, G. (2008); Cameron, B., & Dwyer, F. (2005); Carvelho, T., Allison, R. S., Irving, E. L., & Herriot, C. (2008); Castelli, L., Corazzini, L. L., & Geminiani, C. G. (2008); Chiang, Y.-T., Cheng, C.-Y., & Lin, S. S. J. (2008); Chou, C., & Tsai, M.-J. (2007); Connolly, T. M., Boyle, E., & Hainey, T. (2007); Davidovitch, L., Parush, A., & Shtub, A. (2008); De Lucia, A., Francese, R., Passero, I., & Tortora, G. (2009); Eastin, M. S. (2006); Felicia, P., & Pitt, I. (2007); Feng, J., Spence, I., & Pratt, J. (2007); Fu, F.-L., Su, R.-C., & Yu, S.-C. (2009); Gentile, D. A., & Gentile, J. R. (2008); Green, C. S., & Bavelier, D. (2006); Halpern, D. F., & Wai, J. (2007); Hamalainen, R., Oksanen, K., & Hakkinen, P. (2008); Harr, R., Buch, T., & Hanghøj, T. (2008); Higuchi, S., Motohashi, Y., Liu, Y. L., & Maeda, A. (2005); Hogle, N. J., Widmann, W. D., Ude, A. O., Hardy, M. A., & Fowler, D. L. (2008); Houtkamp, J., Schuurink, E., & Toet, A. (2008); Hsu, C.-L., & Lu, H.-P. (2004); Huizenga, J., Admiraal, W., Akkerman, S., & ten Dam, G. (2007); Huizenga, J., Admiraal,W., Akkerman, S., & ten Dam, G. (2008); Ivory, J. D., & Kalyanaraman, S. (2007); Jennett et al. (2008); Jouriles, E. N., McDonald, R., Kullowatz, A., Rosenfield, D., Gomez, G. S., & Cuevas, A. (2008); Karakus, T., Inal, Y., & Cagiltay, K. (2008); Kiili, K., Ketamo, H., & Lainema, T. (2007); Kim, Y., & Ross, S. D. (2006); Lavender, T. (2008); Lee, M., & Faber, R. J. (2007); Lindh, J., Hrastinski, S., Bruhn, C., & Mozgira, L. (2008); Lucas, K., & Sherry, J. L. (2004); Mayer, I. S., Carton, L., de Jong, M., Leijten, M., & Dammers, E. (2004); Miller, M., & Hegelheimer, V. (2006); Nelson, M. R., Yaros, R. A., & Keum, H. (2006); Nomura, T., Miyashita, M., Shrestha, S., Makino, H., Nakamura, Y., Aso, R., et al. (2008); Nte, S., & Stephens, R. (2008); Orvis, K. A., Horn, D. B., & Belanich, J. (2008); Papastergiou (2009); Ravaja, N., Turpeinen, M., Saari, T., Puttonen, S., & Keltikangas-Jarvinen, L. (2008); Riegelsberger, J., Counts, S., Farnham, S. D., & Philips, B. C. (2006); Rossiou, E., & Papadakis, S. (2008); Russell, W. D., & Newton, M. (2008); Salminen, M., & Ravaja, N. (2008); Schneider, L.-P., & Cornwell, T. B. (2005); Schrader, P. G., & McCreery, M. (2007); Schwabe, G., Goth, C., & Frohberg, D. (2005); Stefanidis, D., Korndorffer, J. R., Jr., Sierra, R., Touchard, C., Dunne, J. B., & Scott, D. J. (2005); Stefanidis, D., Scerbo, M. W., Sechrist, C., Mostafavi, A., & Heniford, B. T. (2008); Steinkuehler, C., & Duncan, S. (2008); Sward, K. A., Richardson, S., Kendrick, J., & Maloney, C. (2008); Terlecki, M. S., & Newcombe, N. S. (2005); Vahed, A. (2008); van Reekum, C. M., Johnstone, T., Banse, R., Etter, A., Wehrle, T., & Scherer, K. R. (2004); Wan, C.-S., & Chiou, W.-B. (2007); Weibel, D., Wissmath, B., Habegger, S., Steiner, Y., & Groner, R. (2008); Wijers, M., Jonker, V., & Kerstens, K. (2008); Yalon-Chamovitz, S., & Weiss, P. L. (2008); Yaman, M., Nerdel, C., & Bayrhuber, H. (2008); Yip, F. W. M., & Kwan, A. C. M. (2006)

(continued)

Table 2.1 (continued)

Girard, **C**., **Ecalle**, **J**. & **Magnan**, **A**. (**2013**). Serious games as new educational tools: How effective are they? A meta-analysis of recent studies, *Journal of Computer Assisted Learning, 29*(3), 207–219. (*n*=30)
Annetta, L. A., Minogue, J., Holmes, S. Y., & Cheng, M.-T. (2009); Baker, R. S. J. D., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010); Beale, I. L., Kato, P. M., Marin-Bowling, V. M., Guthrie, N., & Cole, S. W. (2007); Bloomfield, J., Roberts, J., & While, A. (2010); Boot, W. R., Kramer A. F., Simons, D. J., Fabiani, M., & Gratton G. (2008); Brom, C., Preuss, M., & Klement, D. (2011); Ciavarro, C., Dobson, M., & Goodman, D. (2008); Coller, B. D., & Scott, M. J. (2009); Duque, G., Fung, S., Mallet, L., Posel, N., & Fleiszer, D. (2008); Echeverrìa, A., Garcìa-Campo, C., Nussbaum, M., Gil, F., Villalta, M., Améstica, M., & Echeverrìa, S. (2011); Hainey, T., Connolly, T., Stansfield, M., & Boyle, E. A. (2011); Ke, F. (2008); Kebritchi, M., Hirumi, A., & Bai, H. (2010); Kim, B., Park, H., & Baek, Y. (2009); Knight, J. F., Carley, S., Tregunna, B., Jarvis, S., Smithies, R., de Freitas, S., Dunwell, I., & Mackway-Jones, K. (2010); Kobes, M., Helsloot, I., de Vries, B., & Post, J. (2010); Lim, C. P. (2008); Lindström, P., Gulz, A., Haake, M., & Sjödén, B. (2011); Liu, T.-Y., & Chu, Y.-L. (2010); Lorant-Royer, S., Munch, C., Mesclé, H., & Lieury, A. (2010); Lorant-Royer, S., Spiess, V., Goncalves, J., & Lieury, A. (2008); Miller, L. M., Chang, C.-I., Wang, S., Beier, M. E., & Klisch, Y. (2011); Papastergiou (2009); Radillo, A. (2009); Robertson, D., & Miller, D. (2009); Sung, Y.-T., Chang, K.-E., & Huang, J.-S. (2008); Tanes, Z., & Cemalcilar, Z. (2010); Tuzun, H., Yilmaz-Soylu, M., Karakus, T., Inal, Y., & Kizilkaya G. (2009); Watson, W. R., Mong, C. J., & Harris, C. A. (2011); Wrzesien, M., & Raya, M. A. (2010)
Graafland, **M**., **Schraagen**, **J**. **M**., & **Schijven**, **M**. **P**. (**2012**). Systematic review of serious games for medical education and surgical skills training. *British Journal of Surgery, 99*(10), 1322–1330. (*n*=20)
Andreatta, P. B., Maslowski, E., Petty, S., Shim, W., Marsh, M., Hall, T., et al. (2010); Badurdeen, S., Abdul-Samad, O., Story, G., Wilson, C., Down, S., & Harris, A. (2010); Bergeron, B. P. (2008); Bokhari, R., Bollman-McGregor, J., Kahol, K., Smith, M., Feinstein, A., Ferrara, J. (2010); Bowyer, M. W., Streete, K. A., Muniz, G. M., & Liu, A. V. (2008); Caudell, T. P., Summers, K. L., Holten, J., Hakamata, T., Mowafi, M., Jacobs, J., et al. (2003); Cowan, B., Sabri, H., Kapralos, B., Porte, M., Backstein, D., Cristancho, S., et al. (2010); Creutzfeldt, J., Hedman, L., Medin, C., Heinrichs, W. L., & Fellander-Tsai, L. (2010); Dev, P., Heinrichs, W. L., Youngblood, P., Kung, S., Cheng, R., Kusumoto, L., et al. (2007); Heinrichs, W. L., Youngblood, P., Harter, P., Kusumoto, L., & Dev, P. (2010); Knight, J. F., Carley, S., Tregunna, B., Jarvis, S., Smithies, R., de Freitas, S., et al. (2010); Kurenov, S. N., Cance, W. W., Noel, B., Mozingo, D. W. (2009); LeRoy, H. W., Youngblood, P., Harter, P. M., & Dev, P. (2008); Parvati, D. E. V., Heinrichs, W. L., & Patricia, Y. (2011); Rosenberg, B. H., Landsittel, D., & Averch, T. D. (2005); Rosser, J. C., Jr, Lynch, P. J., Cuddihy, L., Gentile, D. A., Klonsky, J., & Merrell, R. (2007); Sadandanan, S., Dryfhout, V. L., & Sosnowski, J. P. (2008); Schlickum, M. K., Hedman, L., Enochsson, L., Kjellin, A., & Felländer-Tsai, L. (2009); Taekman, J. M., & Shelley, K. (2010); Youngblood, P., Harter, P. M., Srivastava, S., Moffett, S., Heinrichs, W. L., & Dev P. (2008)

Papastergiou, **M**. (**2009**). Exploring the potential of computer and video games for health and physical education: A literature review. *Computers & Education, 53*(3), November 2009, 603–622. (*n* = 19)
Bartholomew, L., Gold, R., Parcel, G., Czyzewski, D., Sockrider, M., Fernandez, M., et al. (2000); Beale, I., Kato, P., Marin-Bowling, V., Guthrie, N., & Cole, S. (2007); Chin, M., Paw, A., Jacobs, W., Vaessen, E., Titze, S., & van Mechelen, W. (2008); Ciavarro, C., Meanley, J., Bizzocchi, J., & Goodman, D. (2005); Cullen, K., Watson, K., Baranowski, T., Baranowski, J., & Zakeri, I. (2005); Fery, Y., & Ponserre, S. (2001); Goodman, D., Bradley, N., Paras, B., Williamson, I., & Bizzocchi, J. (2006); Hewitt, M., Denman, S., Hayes, L., Pearson, J., & Wallbanks, C. (2001); Hornung, R., Lennon, P., Garrett, J., DeVellis, R., Weinberg, P., & Strecher, V. (2000); Lieberman, D. (2001); Munguba, M., Valdes, M., & da Silva, C. (2008); Russell, W. (2007); Sell, K., Lillie, T., & Taylor, J. (2007); Silk, K., Sherry, J., Winn, B., Keesecker, N., Horodynski, M., & Sayir, A. (2008); Tan, B., Aziz, A., Chua, K., & Teh, K. (2002); Turnin, M., Tauber, M., Couvaras, O., Jouret, B., Bolzonella, C., Bourgeois, O., et al. (2001); Tuzun, H. (2007); Unnithan, V., Houser, W., & Fernhall, B. (2006); Yawn, B., Algatt-Bergstrom, P., Yawn, R., Wollan, P., Greco, M., Gleason, M., et al. (2000)
Vandercruysse, **S**., **Vandewaetere**, **M**., & **Clarebout**, **G**. (**2012**). Game based learning: A review on the effectiveness of educational games. In M. M. Cruz-Cunha (Ed.), *Handbook of research on serious games as educational*, *business*, *and research tools* (pp. 628–647). Hershey, PA: IGI Global. (*n* = 22)
Annetta, L. A., Mangrum, J., Holmes, S., Collazo, K., & Cheng, M.-T. (2009a); Annetta, L. A., Minogue, J., Holmes, S. Y., & Cheng, M.-T. (2009b); Beale, I. L., Kato, P. M., Marin-Bowling, V. M., Guthrie, N., & Cole, S. W. (2007); Brown, S. J., Lieberman, D. A., Gemeny, B. A., Fan, Y. C., Wilson, D. M., & Pasta, D. J. (1997); Chuang, T.-Y., & Chen, W.-F. (2009); Din, F. S., & Calao, J. (2001); Ebner, M., & Holzinger, A. (2005); Fontana, L., & Beckerman, A. (2004); Goodman, D., Bradley, N. L., Paras, B., Williamson, I. J., & Bizzochi, J. (2006); Huizenga, J., Admiraal, W., Akkerman, S., & ten Dam, G. (2009); Ke, F. (2008); Ke, F., & Grabowski, B. (2007); Miller, D. J., & Robertson, D. P. (2010); Moreno, R., & Mayer, R. (2005); Papastergiou (2009); Prendinger, H., Mori, J., & Ishizuka, M. (2005); Richards, D., Fassbender, E., Bilgin, A., & Thompson, W. F. (2008); Virvou, M., Katsionis, G., & Manos, K. (2005); Warren, S. J, Dondlinger, M. J., & Barab, S. A. (2008); Wrzesien, M., & Raya, M. A. (2010); Yip, F. W. M., & Kwan, A. C. M. (2006); Yu, F.-Y. (2003)
Wattanasoontorn, **V**., **Boada**, **I**., **García**, **R**., **Sbert**, **M**. (**2013**) Serious games for health. *Entertainment Computing, 4*(4), December 2013, 231–247. (*n* = 91)

(continued)

Table 2.1 (continued)

Anchor Bay Entertainment (2008); Anderson, C. (2008); Applied Research Associates Inc (2012); Archimage Inc (2006); Association RMC/BFM (2011); Atkinson, S., & Narasimhan, V. (2010); Bartolome, N., Zorrilla, A., & Zapirain, B. (2010); BBG Entertainment (2009); Blitz Games (2009); Botella, C., Breton-Lpez, J., Quero, S., Baos, R., Garca-Palacios, A., Zaragoza, I., & Alcaniz, A. (2011); BreakAway Ltd (2010); Burke, J. W., McNeill, M. D. J., Charles, D. K., Morrow, P. J., Crosbie, J. H., McDonough, S. M. (2009); Cagatay, M., Ege, P., Tokdemir, G., & Cagiltay, N. (2012); Chan, W., Qin, J., Chui, Y., & Heng, P. (2012); Clawson, J., Patel, N., & Starner, T. (2010); Collision Studios (2009); De Bortoli A, & Gaggi, O (2011); de Urturi, Z., Zorrilla, A., & Zapirain, B. (2011); Deponti, D., Maggiorini, D., & Palazzi, C. (2009); Diehl, L., Lehmann, E., Souza, R., Alves, J., Esteves, R., & Gordan, P. (2011); EAD Nintendo (2007); Edheads (2007); e-Learning Studios (2012); Electronic Arts (2010); EMCO3 (2012); Faculty of Medicine, Imperial College London (2008); fatworld.org (2007); Finkelstein, J., Wood, J., Cha, E., Orlov, A., & Dennison, C. (2010); Fishing Cactus (2010); Fuchslocher, A., Niesenhaus, J., & Krmer, N. (2011); Gago, J., Barreira, T., Carrascosa, R., & Segovia, P. (2010); Gameloft (2006); GENIOUS Interactive (2012); Glasgow Caledonian University (2008); Grau, S., Tost, D., Campeny, R., Moya, S., & Ruiz, M. (2010); Hatfield, D. (2008); HopeLab (2010); Imbeault, F., Bouchard, B., & Bouzouane, A. (2011); Inhalothrapie Kinsithrapie Respiratory Association for Research and Education (2012); Innovation in Learning Inc (2010); Intelligent Systems (2007); Janomedia (2006); Johnston, E., & Duskin, B. (2004); K.T.M. Advance (2010); Kim, J. A., Kang, K. K., Yang, H. R., & Kim, D. (2009); Laikari, A. (2009); Lakeside Center for Autism (2011); LearningGames Lab (2009); Lightning Fish Games (2009); Lin, J. K., Cheng, P. H., Su, Y., Wang, S. Y., Lin, H. W., Hou, H. C., Chiang, W. C., Wu, S. W., Luh, J. J., & Su, M. J. (2011); LudoMedic 2011 (2011); Mc.G.Ill. University (2006); Mckanna, J. A., Jimison, H., & Pavel, M. (2009); Mili,F., Barr, J., Harris, M., & Pittiglio, L. (2008); Miller, J. (2010); Milo Foundation (2012); MIRROR project (2012); Montreal Science Centre (2004); Moya, S., Grau, S., Tost, D., Campeny, R., & Ruiz, M. (2011); Nauta, H., & Spil, T. (2011); Nike Inc (2012); Nintendo (2005); Nordic Innovation Centre (2007); Pervasive Games (2009); QOVEO (2009); Queiros, S., Vilaca, J., Rodrigues, N., Neves, S., Teixeira, P., & Correia-Pinto, J. (2011); RANJ Serious Games (2009); Raylight Srl (2009); Red Hill Studios (2012); Respondesign (2004); Sabri, H., Cowan, B., Kapralos, B., Porte, M., Backstein, D., & Dubrowskie, A. (2010); Scarle, S., Dunwell, I., Bashford-Rogers, T., Selmanovic, E., Debattista, K., Chalmers, A., Powell, J., & Robertson W. (2011); Schnauer, C., Pintaric, T., Kosterink, S. J., & Vollenbroek, M. (2011); Skills2Learn Ltd (2010); Sliney, A., & Murphy, D. (2008); Succubus Interactive (2009); The Diablotines (2011); The Partnership for Food Safety Education (2008); TruSim, Blitz Games Studios (2008); Ubisoft Divertissements (2012); Van Loon, E., Peper, C., van de Rijt, A., & Salverda, A. (2011); Vazquez, M., Santana-Lopez, V., Skodova, M., Ferrero-Alvarez-Rementeria, J., & Torres-Olivera A. (2011); Verduin, M. L., LaRowe, S. D., Myrick, H., Cannon-Bowers, J., & Bowers, C. (2012); Vermont Department of Health (2008); Vermont department of health, Khemia (2008); Vidani, A., Chittaro, L., & Carchietti, E. (2010); Virtual Heroes Inc (2009); Visual Imagination Software (2010); Vtnen, A., & Leikas, J. (2009); Wang, Q., Sourina, O., & Nguyen, M. K. (2010); Warner Bros Entertainment (2008)

Wilson, **K. A**., **Bedwell**, **W. L**., **Lazzara**, **E. H**., **Salas**, **E**., **Burke**, **C. S**., **Estock**, **J. L**., **Orvis**, **K. L**., & **Conke**, **C**. (**2009**). Relationships between game attributes and learning outcomes: Review and research proposals. *Simulation Gaming, 40*(2), 217–266. (*n* = 42)
Adams (1998); Belanich, J., Sibley, D. E., & Orvis, K. L. (2004); Blunt (2007); Bowers & Jentsch (2001); Crown (2001); Day, Arthur, & Gettman (2001); Dennis & Harris (1998); Driskell, J. E., & Dwyer, J. D. (1984); Gander, S. (2002); Garris & Ahlers (2001); Gopher, D., Weil, M., & Bareket, T. (1994); Gremmen, H., & Potters, H. (1995); Habgood, M. P. J. (2005); Habgood, M. P. J., Ainsworth, S. E., & Benford, S. (2005); Lepper, M. R. (1985); Lepper, M. R., & Chabay, R. W. (1985); Malone, T. W. (1981); Marks (2000); Mayer, R. E., Mautone, P., & Prothero, W. (2002); McFarlane, Sparrowhawk, & Heald (2002); Merzenich, M. M., Jenkins, W. M., Johnston, J. P., Schreiner, C., Miller, S. L., & Tallal, P. (1996); Noble, A., Best, D., Sidell, C., & Strang, J. (2000); Orvis, K. A., Orvis, K. I., Belanich, J., & Mullin, L. N. (2008); Pange, J. (2003); Parker & Lepper (1992); Parry, S. B. (1971); Prince, C., & Jentsch, F. (2001); Proctor, M. D., Panko, M., & Donovan, S. J. (2004); Ricci, K. E., Salas, E., & Cannon-Bowers, J. A. (1996); Rieber, L. P. (1996); Rieber, L. P., & Noah, D. (1997); Ronen, M., & Eliahu, M. (2000); Rosas, R., Nussbaum, M., Cumsille, P., Marianov, V., Correa, M., & Flores, P., et al. (2003); Serrano, E. L., & Anderson, J. E. (2004); Squire, Giovanetto, Devane, & Durga (2005); Thomas, R., Cahill, J., & Santilli, L. (1997); Van Eck (2006); Van Eck, R., & Dempsey, J. (2002); Veale, T. K. (1999); Virvou, M., Katsionis, G., & Manos, K. (2005); Westrom, M., & Shaban, A. (1992); Woodman, M. D. (2006)

Table 2.2 Number of serious game papers for each year in the 10 year period from 2003 to 2012

Year	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
No. of papers	2	7	15	10	18	49	19	22	12	3

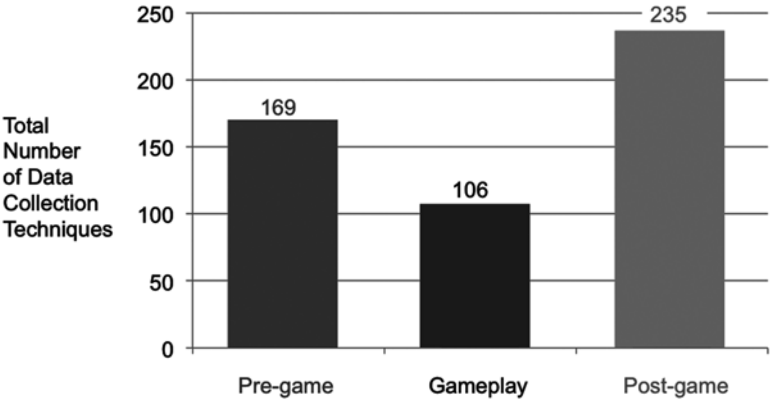


Fig. 2.4 Number of data collection techniques used per phase of study

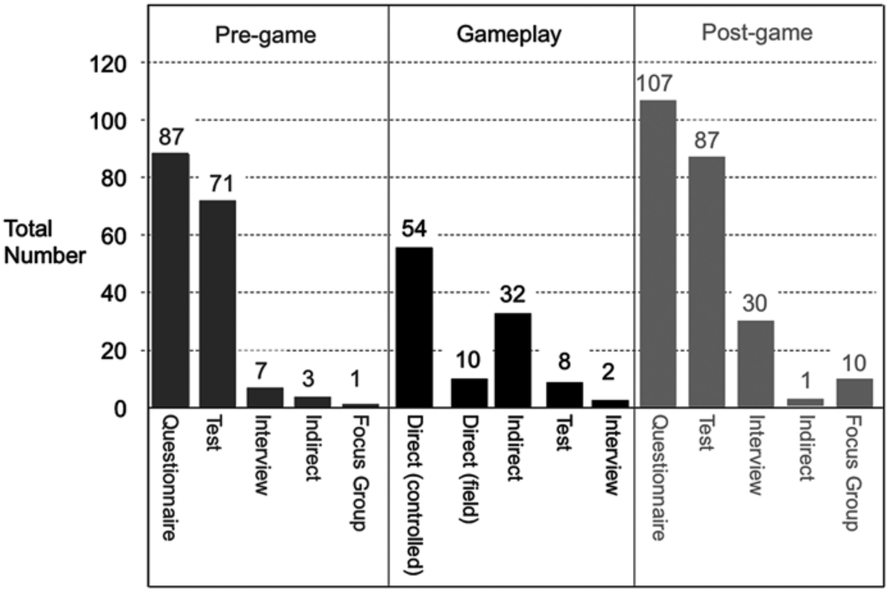


Fig. 2.5 Number of specific data collection techniques used per phase of study

A single study used an indirect observation, while 4 % of the studies employed a focus group in the post-game phase (see Fig. 2.5).

In the context of the specific techniques used during gameplay ($n = 106$), 46 % of the studies used some form of direct observation in a controlled environment, 9 % of the studies used some form of direct observation in the field, 30 % of the studies used an indirect observation method, around 8 % used a test, while two of the studies employed an interview during the gameplay phase of evaluation (see Fig. 2.5).

5 Discussion

5.1 Issues Highlighted Within Our Study Outcomes

The majority of the studies we reviewed used multiple data collection methods (80.3 %, Fig. 2.6). Surveys and questionnaires are good at getting shallow data from a large number of people but are not good at getting, deep, detailed data; participants may try to impress interviewers during interviews (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003); and duration logging systems may not take into account participant thinking time (Lazar, Feng, & Hochheiser, 2010). The studies included in our review were dominated by the use of questionnaires and formal tests (see Fig. 2.5). Also, the majority of data collection occurred post-game, followed by pre-game collection and finally captured during gameplay. Although these results may have been biased by our own sampling techniques, it may also highlight a need for integrating more focus groups and indirect observation techniques into game evaluations.

Even allowing for sampling errors, it was notable that objective techniques such as biometrics or psychometrics, as well as newer techniques such as path tracking or crowd sourcing, were largely absent. This reflects traditional difficulties of capturing gameplay data, that is, if such data collection is not part of the game design, and it also highlights opportunities for new data capture approaches oriented towards data

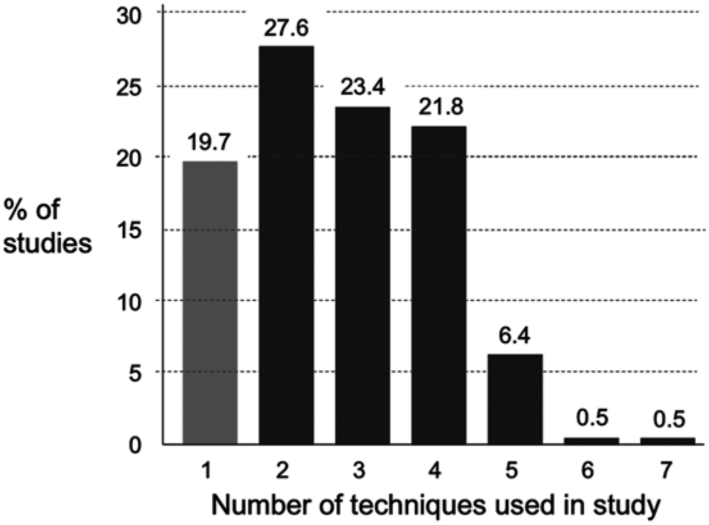


Fig. 2.6 Percentage of studies that used multiple data collection techniques

analytics. Loh (2012) notes that "… few [commercial] developers would actually be interested in 'in-process' data collection unless it somehow contributed to the usability of their games …" and goes on to consider alternative approaches for empirical data gathering via telemetry and psychophysiological measures.

Also, different data collection techniques have inherent biases (Podsakoff et al., 2003). Thus, it is important to consider multiple data collection methods. As Rogers et al. (2011) observe, it is "important not to focus on just one data gathering technique but to use them flexibly and in combination so as to avoid biases" (p. 223). The framework used by Mayer et al. (2014) provides a good examination of the breadth of approaches that can be used in data collection.

Most of the studies were collecting data to demonstrate the use of serious games as an intervention tool, for instance, to demonstrate the impact of a serious game in an educational setting. Thus, a minimum expectation could be for a pretest and posttest, and it would also be desirable to obtain some in game data, e.g., score or duration metric. As seen in Fig. 2.6, only 52.7 % of the studies reviewed used three or more data collection techniques. Exploring this further is outside the scope of this chapter, but is an important area for future research if serious game evaluations and experimental designs are to be considered demonstrably robust.

Another feature highlighted in the data collection that occurred during gameplay was the lack of direct observation in the field (10 %) compared to observations that were made in a controlled environment such as a computer laboratory (54 %). This is understandable, as research by nature tends to occur in university environments, and controlled environments allow contextual variations associated with data collection to be controlled in traditional experimental designs. Again, our own data sampling methods make it difficult to argue the significance of this finding but it still needs to be considered that experimental serious game research might need to be extended to include more situated case studies and perhaps participatory methods.

When reflecting further on the content of the various studies we encountered during our process, as well as some of the problems encountered in the data collection process, a number of generic issues of serious games research were highlighted. These generic issues, which we discuss next, include:

- What data is being collected?
- When data is being collected?
- Where data is being collected?
- Who is involved in data collection?
- Why data is being collected?

5.2 What Data Is Being Collected?

In our study, we found there was a tendency to collect certain types of data during the different phases. During the pre-game, this data tends to include demographic information such as gender, age, nationality, and culture. It was also common to gather data surrounding previous experience and skills with computers, games, and related technology such as simulations and virtual reality. Less common is the collection of data of a participant's attitudes, their intrinsic and extrinsic motivation, learning and personality styles, etc. In many cases, both pre- and posttests were used to generate skill or knowledge performance metrics directly related to intended serious outcomes of the game.

During the gameplay, measures tend to focus on issues of performance. Types of data include game metrics, such as time to complete, number of errors or levels of progress. Less common were measures that examined player approaches to completing the game and measures of experience such as flow, immersion, presence, and the general affective states of the participant (see, for example, Jennett et al., 2008).

Loh (2011) considers gameplay measures with the analogue of Black box and open box approaches in the context of game assessment metrics. Specifically, he defines ex situ data collection where the game environment is a Black box and data is collected without access to internal details. This could be the pre- and post-game collections metrics noted in this chapter (see Fig. 2.5) or psychophysiological measures collected during game sessions. The open box approach supports in situ data collection, for example, log files, game events, or user-generated action data, e.g., *Information Trails* (Loh, 2012). In contrast to psychophysiological measures, such in situ data would have no external noise as the data collection occurred within a closed environment. This could be of significant interest for serious game analytic approaches as a way to triangulate data across collection sources, similar to the use of *immersidata* to collect and index user-player behaviors from gameplay logs and video clips (Marsh, Smith, Yang, & Shahabi, 2006). Also, as noted in the previous section, there have traditionally been difficulties in capturing and using in situ data as it requires access to the internal processes of a serious game (e.g., to collect telemetry data), and it can be problematic to efficiently process the large volumes of data generated. Both topics are prominent in the other chapters of this book and are a focus of ongoing serious game analytics research.

During post-game evaluations, it was more common to obtain subjective feedback concerning game experience and issues surrounding fun and engagement. This phase was also when measures of player satisfaction with the game, such as clarity, realism, aesthetics, and ease of use, as well as perceived suitability were usually made.

Other types of data that might be useful to collect within studies include the quality or experience of any facilitators involved, the general context of field studies such as the interaction with others and their roles in the study, and potential organizational impacts such as management structure and culture (Mayer et al., 2014).

Some of the variation in data collection is related to the intention of studies and whether they relate to measuring the efficacy of serious outcomes, or the usability and quality of the game itself. Many studies address both efficacy and usability issues as they are related. One issue that needs to be considered in relation to what data is collected surrounds player profiling. The importance of this is highlighted in one study that used the specialized "Ravens advanced progressive matrices" to examine the relationship between general cognitive ability and any measured knowledge outcomes from the game (Day, Arthur, & Gettman, 2001). The inference is that underlying individual traits such as cognitive ability might be a good indicator of player performance in learning tasks. This suggests other psychological tests that might assist in measuring player traits such as risk-taking, general personality traits, performance under stress, learning styles, teamwork, and other factors that might be relevant in some applications.

A benefit of adopting these traditional instruments is that they have been validated and are well understood, at least under laboratory conditions. Otherwise, we might also need to question the validity of questionnaires, surveys, and other measuring instruments currently being used in game studies (Boyle, Connolly, & Hainey, 2011; Slater, 2004).

5.3 When Data Is Being Collected?

Our own study identified a variety of data being collected in the post-game, gameplay, and post-game phases of evaluation. Most of the data collected in our study occurred post-game, while the least occurred during the gameplay phase. Arguably there is an opportunity to improve levels of data collection occurring during the gameplay to support a better understanding of how specific game elements relate to the intended serious outcomes.

There are also other aspects of timing that should be considered in data collection. This may partly be related to the whether the study is intent on measuring aspects of the process or purely outcomes (Bowers & Jentsch, 2001). Thus, the relevance of process evaluation versus game efficacy or usability measures may impact on when data is collected.

In terms of learning applications, it may also be important to consider interactions between other forms of instruction that occur before, during, or after the game intervention (Van Eck, 2006). This might also apply to application of serious games for health, where additional treatments may occur in conjunction with game use.

This highlights the issue of deciding when and how often to collect data for evaluating games. Although many studies used mixed methods, data was not necessarily collected over the life of the study. By contrast, in the study by Squire, Giovanetto, Devane, and Durga (2005) games were played over 5 weeks and data was collected over this entire time frame. The time frame of data collection may be influenced by the intent and domain of the study, for example, whether the research is concerned with the direct and immediate influence of playing the game versus the indirect or long-term impact of the game. It is probably important to get short-term feedback involving gaining self-reported, subjective feedback from participants, for example, regarding participant satisfaction, or self-perceived learning as well as immediate changes to attitudes, skills, or knowledge. Medium or longer term data might be required to understand aspects of team or organizational change especially related to social issues.

We also found that the timing of outcome measures varied depending on domain. For example, in some learning applications there may be a greater tendency to measure longer term learning factors such as the time required to transfer or regain knowledge (Day et al., 2001; Dennis & Harris, 1998; Parker & Lepper, 1992). This implies testing skills, not just immediately after completing the game, but also at later intervals such as a few days, weeks, or months to measure the permanence of any immediate outcomes and issues of retention and reacquisition of knowledge.

5.4 Where Data Is Being Collected?

There was a tendency for evaluation to occur in controlled rather than field situations. Data was collected in a variety of contexts including primary schools, secondary schools, universities, and industry settings. One potential issue of study context is what else is happening during the study that might impact on outcomes and yet is not necessarily being reported (Van Eck, 2006).

The location of data collection also indirectly raises issues of cost. For example, onsite studies should be fast and efficient to ensure they do not unnecessarily impact on the time or resources of participating partners (Mayer et al., 2014). It also confirms the need for unobtrusive and perhaps covert data collection techniques (Mayer et al., 2014), not just to improve data validity, but to minimize impact on the standard workflow of participants involved in case studies, for example, the use of in situ methods (Loh, 2011). Stakeholders may also need to be persuaded that more extensive contextual data as well as extended longitudinal data gathering needs to occur beyond the obvious and minimal (Mayer et al., 2014).

5.5 Who Is Involved in Data Collection?

In our study, we identified a range of stakeholders involved in projects including students, teachers, researchers, game developers, and industry partners. All of these various stakeholders are candidates to be involved in evaluation. Such evaluations may need to bear in mind influences related to the motivations of stakeholders surrounding the process and outcomes. For example, the game designer may be enthusiastic to measure the aesthetics, the software engineer the usability, the scientist, the efficacy, and the manager the cost. All stakeholders may also be keen to find positive outcomes whether the motivation is for publication, ongoing employment, or other personal gains. Thus, it may be worthwhile to consider collecting data related to the exact role of various participants in the project and any intrinsic and extrinsic motivations of the parties (Mayer et al., 2014).

Marks (2000) highlights the obvious sampling issues in one project where university students were used to evaluate a game intended to teach military staff. It was not clear that measured effects on such a population would transfer to the intended group. By contrast, in another study three different questionnaires are used for pupils (players), parents, and teachers (McFarlane, Sparrowhawk, & Heald, 2002).

While most studies focus on individuals playing games, there is also interest in evaluating the efficacy of learning team-based, rather than individual, skills. Marks (2000), in considering some of the pros and cons of using computer simulations for team research, highlights the need for measuring the longitudinal impact of skills related to teamwork. Data may need to be collected that considers team performance rather than individual performance where games are designed to teach teamwork (Bowers & Jentsch, 2001). While a number of evaluation models exist that focus on the learning of individuals, there has been less attention given to the data required to assess learning in teams or in larger collectives such as organizations and informal networks (Mayer et al., 2014).

5.6 Why Data Is Being Collected?

During our meta-review, we encountered the use of games across a wide variety of different domains and not surprisingly, we found a variety of expectations about the most appropriate methods of data collection and the types of data collected. For example, in one study participants were partly evaluated on the basis of an essay that reflected on their experience using the game (Adams, 1998). This is in contrast to another learning study that directly measured changes to student knowledge using tests as well as surveying the students and seeking feedback from external stakeholders such as parents and teachers (Crown, 2001).

While it is easy to understand the reasons for such differences, the variations make it harder to compare and contrast data results from different game studies. The usefulness to the serious game community of adopting standardized testing approaches that allow for comparison has been highlight previously (Blunt, 2007; Mayer et al., 2014).

Despite some good work in the area of relating game design features to serious outcomes (Wilson et al., 2009), most studies focus on collecting data to support the message of efficacy rather than data that helps explain why and how they are effective or indeed how to apply design rules that lead to the required efficacy (Garris & Ahlers, 2001; Van Eck, 2006).

6 Conclusions

A complication of data collection for games is that not all games are created equal (Loh, 2009). Van Eck (2006) makes the key point that any taxonomy of games is as complex as learning taxonomies. While not all games are the same, the situation is complicated by the overlap of simulations, virtual reality, and partial gamification of traditional approaches. There is also wide variety in the types of games being used in studies. Some are small in scope and custom built by individuals while others are constructed in multi-discipline projects that involve discipline experts and professional game developers. Other studies simply make use of off-the-shelf games. This range of projects means that the data collection techniques need to be flexible.

In this study, we developed a review process for performing a meta-analysis on data collection techniques used in serious game research. We found that while many studies used a variety of methods, they were not necessarily intended to triangulate findings. The number of data collection techniques also varied considerably, with a number of studies using only a single measure. Our study also highlighted a number of variations and subsequently raised questions around what data is being collected, when data is being collected, how data is being collected, where data is being collected, and why data is being collected.

Our systematic review approach identified a number of significant literature reviews that allowed us to examine data collection processes across broad domain and temporal spaces. However, not all bibliographic sources were included in the search parameters and thus relevant literature has potentially been missed. Despite this, the results provide a representative sample of serious game research that allows us to draw valid conclusions about approaches and issues in data collection.

In summary, the data collected for serious game research is broad in scope, measuring both targeted performance skills, behavioral factors related to both the process and outcomes. For example, the data may be designed to measure changes in knowledge, attitudes, skills, or behavior. The data collected can also be multilevel in scope, designed to measure fine grain individual skills or large-scale organization attitudes. Data is collected using a wide range of objective and subjective methods that may fall across a range of longitudinal scales. The currently used data collection techniques might align more with discipline traditions than necessarily intentions of evaluations. Even though single studies often incorporate a variety of techniques, the data is not necessarily triangulated as might be expected in a true mixed method approach. The review also found that the majority of data collection occurred post-game, then pre-game, and finally during gameplay. This, perhaps, reflects traditional difficulties of capturing gameplay data and highlights opportunities for new data capture (i.e., in situ collection) and analysis approaches oriented towards data analytics. We suggest that more standardized and better-validated data collection techniques, that allow comparing and contrasting outcomes between studies, would be beneficial to the broader serious games community and specifically to those interested in serious game analytics.

References

Adams, P. C. (1998). Teaching and learning with SimCity 2000. *Journal of Geography, 97*(2), 47–55.

Anderson, E., McLoughlin, L., Liarokapis, F., Peters, C., Petridis, P., & Freitas, S. (2010). Developing serious games for cultural heritage: a state-of-the-art review. *Virtual Reality, 14*(4), 255–275. doi:10.1007/s10055-010-0177-3.

Bearman, M., Smith, C. D., Carbone, A., Slade, S., Baik, C., Hughes-Warrington, M., et al. (2012). Systematic review methodology in higher education. *Higher Education Research & Development, 31*(5), 625–640.

Blakely, G., Skirton, H., Cooper, S., Allum, P., & Nelmes, P. (2009). Educational gaming in the health sciences: Systematic review. *Journal of Advanced Nursing, 65*(2), 259–269. doi:10.1111/j.1365-2648.2008.04843.x.

Blunt, R. (2007). Does game-based learning work? Results from three recent studies. In *Proceedings of the Interservice/Industry Training, Simulation, & Education Conference* (pp. 945–955). Orlando, FL: National Defense Industrial Association.

Bowers, C. A., & Jentsch, F. (2001). Use of commercial off-the-shelf, simulations for team research. In C. A. Bowers & E. Salas (Eds.), *Advances in human performance and cognitive engineering research* (pp. 293–317). Mahwah, NJ: Lawrence Erlbaum.

Boyle, E. A., Connolly, T. M., & Hainey, T. (2011). The role of psychology in understanding the impact of computer games. *Entertainment Computing, 2*(2), 69–74.

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education, 59*(2), 661–686. doi:10.1016/j.compedu.2012.03.004.

Crookall, D. (2010). Serious games, debriefing, and simulation/gaming as a discipline. *Simulation & Gaming, 41*(6), 898–920.

Crown, S. W. (2001). Improving visualization skills of engineering graphics students using simple JavaScript web based games. *Journal of Engineering Education, 90*(3), 347–355.

Day, E. A., Arthur, W., Jr., & Gettman, D. (2001). Knowledge structures and the acquisition of a complex skill. *Journal of Applied Psychology, 86*(5), 1022–1033.

Dennis, K. A., & Harris, D. (1998). Computer-based simulation as an adjunct to ab initio flight training. *International Journal of Aviation Psychology, 8*(3), 261–276.

Garris, R., & Ahlers, R. (2001, December). *A game-based training model: Development, application, and evaluation*. Paper presented at the Interservice/Industry Training, Simulation & Education Conference, Orlando, FL.

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation and learning: A research and practice model. *Simulation & Gaming, 33*(4), 441–467.

Girard, C., Ecalle, J., & Magnan, A. (2013). Serious games as new educational tools: How effective are they? A meta-analysis of recent studies. *Journal of Computer Assisted Learning, 29*(3), 207–219. doi:10.1111/j.1365-2729.2012.00489.x.

González, L. S., Rubio, F. G., González, F. R., & Velthuis, M. P. (2010). Measurement in business processes: A systematic review. *Business Process Management Journal, 16*(1), 114–134.

Graafland, M., Schraagen, J. M., & Schijven, M. P. (2012). Systematic review of serious games for medical education and surgical skills training. *British Journal of Surgery, 99*(10), 1322–1330.

Hwang, G.-J., & Wu, P.-H. (2012). Advancements and trends in digital game-based learning research: A review of publications in selected journals from 2001 to 2010. *Journal of Educational Technology, 43*(1), E6–E10. doi:10.1111/j.1467-8535.2011.01242.x. Blackwell.

Jennett, C., Cox, A. L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., et al. (2008). Measuring and defining the experience of immersion in games. *International Journal of Human Computer Studies, 66*(9), 641–661.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering—A systematic literature review. *Information and Software Technology, 51*(1), 7–15.

Lazar, J., Feng, J. H., & Hochheiser, H. (2010). *Research methods in human-computer interaction*. Chichester, England: Wiley.

Loh, C. S. (2009). Research and developing serious games as interactive learning instructions. *International Journal of Gaming and Computer-Mediated Simulations, 1*(4), 1–19. IGI Global.

Loh, C. S. (2011). Using in situ data collection to improve the impact and return of investment of game-based learning. In *Proceedings of ICEM-SIIE 2011, the 61st International Council for Educational Media (ICEM) and the XIII International Symposium on Computers in Education (SIIE) Joint Conference* (pp. 801–811). ICEM-SIIE.

Loh, C. S. (2012). Information trails: In-process assessment for game-based learning. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 123–144). New York: Springer.

Marks, M. A. (2000). A critical analysis of computer simulations for conducting team research. *Small Group Research, 31*(6), 653–675.

Marsh, T., Smith, S. P., Yang, K., & Shahabi, C. (2006) Continuous and unobtrusive capture of user-player behaviour and experience to assess and inform game design and development. In *1st World Conference for Fun'n Games (FNG 2006)* (pp. 79–86). Preston, England: University of Central Lancaster.

Mayer, I., Bekebrede, G., Harteveld, C., Warmelink, H., Zhou, Q., Ruijven, T., et al. (2014). The research and evaluation of serious games: Toward a comprehensive methodology. *British Journal of Educational Technology, 45*(3), 502–527.

McFarlane, A., Sparrowhawk, A., & Heald, Y. (2002). *Report on the educational use of games: An exploration by TEEM of the contribution which games can make to the education process*. Last Access on 11th April, from https://pantherfile.uwm.edu/tjoosten/LTC/Gaming/teem_gamesined_full.pdf

McMahan, R. P., Ragan, E. D., Leal, A., Beaton, R. J., & Bowman, D. A. (2011). Considerations for the use of commercial video games in controlled experiments. *Entertainment Computing, 2*(1), 3–9.

Papastergiou, M. (2009). Exploring the potential of computer and video games for health and physical education: A literature review. *Computers & Education, 53*(3), 603–622. doi:10.1016/j.compedu.2009.04.001.

Parker, L. E., & Lepper, M. R. (1992). Effects of fantasy contexts on children's learning and motivation: Making learning more fun. *Journal of Personality and Social Psychology, 62*(4), 625–633.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879.

Rogers, Y., Sharp, H., & Preece, J. (2011). *Interaction design: Beyond human-computer interaction* (3rd ed.). Hoboken, NJ: Wiley.

Slater, M. (2004). How colorful was your day? Why questionnaires cannot assess presence in virtual environments. *Presence, 13*(4), 484–493.

Šmite, D., Wohlin, C., Gorschek, T., & Feldt, R. (2010). Empirical evidence in global software engineering: A systematic review. *Empirical Software Engineering, 15*(1), 91–118.

Smith, S. P., Stibric, M., & Smithson, D. (2013). Exploring the effectiveness of commercial and custom-built games for cognitive training. *Computers in Human Behavior, 29*(6), 2388–2393.

Squire, K., Giovanetto, L., Devane, B., & Durga, S. (2005). From users to designers: Building a self organizing game-based learning environment. *TechTrends, 49*(5), 34–42.

Van Eck, R. (2006). Digital game based learning: It's not just the digital natives who are restless. *Educause Review, 41*(2), 17–30.

Vandercruysse, S., Vandewaetere, M., & Clarebout, G. (2012). Game based learning: A review on the effectiveness of educational games. In M. M. Cruz-Cunha (Ed.), *Handbook of research on serious games as educational, business, and research tools* (pp. 628–647). Hershey, PA: IGI Global.

Wattanasoontorn, V., Boada, I., García, R., & Sbert, M. (2013). Serious games for health. *Entertainment Computing, 4*(4), 231–247.

Wilson, K. A., Bedwell, W. L., Lazzara, E. H., Salas, E., Burke, C. S., Estock, J. L., et al. (2009). Relationships between game attributes and learning outcomes review and research proposals. *Simulation & Gaming, 40*(2), 217–266.

Part II
Measurement of Data in Serious Games Analytics

Chapter 3
Guidelines for the Design and Implementation of Game Telemetry for Serious Games Analytics

Gregory K.W.K. Chung

Abstract The design of game telemetry requires careful attention to the chain of reasoning that connects low-level behavioral events to inferences about players' learning and performance. Measuring performance in serious games is often difficult because seldom do direct measures of the desired outcome exist in the game. Game telemetry is conceived as the fundamental element from which measures of player performance are developed. General psychometric issues are raised for game-based measurement, and data issues are raised around format, context, and increasing the meaningfulness of the data itself. Practical guidelines for the design of game telemetry are presented, including targeting in-game behaviors that reflect cognitive demands, recoding data at the finest usable grain size, representing the data in a format usable by the largest number of people, and recording descriptions of behavior and not inferences with as much contextual information as practical. A case study is presented on deriving measures in a serious game intended to teach fraction concepts.

Keywords Game telemetry • Behavioral observations • Construct validity • Game-based learning • Embedded assessment • Serious games

1 Introduction

The use of serious games to support teaching and learning is increasing at a rapid rate, as they gain acceptance by teachers, schools, trainers, and policy makers (U.S. Department of Education (DOE), 2012). The integration of games in existing educational media (e.g., textbooks) and across various media platforms—transmedia—is expected to increase (Corporation for Public Broadcasting (CPB) & PBS Kids, 2011). The growing empirical base on the efficacy of serious games for learning

G.K.W.K. Chung (✉)
National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
University of Calififornia, Los Angeles (UCLA), Los Angeles, CA 90095, USA
e-mail: greg@ucla.edu

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based Learning, DOI 10.1007/978-3-319-05834-4_3

suggests that serious games will become an accepted part of formal education and training settings (e.g., Chung, Choi, Baker, & Cai, 2014; Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012; Girard, Ecalle, & Magnan, 2013; Tobias, Fletcher, Dai, & Wind, 2011).

An emerging trend that has potential for a wide-scale and long-term impact on education is the idea that games can provide insights into students' cognitive states based on their in-game behavior—from fine-grained, moment-to-moment play behavior to summary game behavior such as number of levels reached. The capability to automatically capture fine-grained player behavior has generated interest in using games as a means to measure student learning (e.g., Baker, Chung, & Delacruz, 2012; Baker & Yacef, 2009; Chung, 2014; Drachen, Thurau, Togelius, Yannakakis, & Bauckhage, 2013; Loh, 2012; Romero & Ventura, 2007, 2010; Shaffer & Gee, 2012; Shoukry, Göbel, & Steinmetz, 2014; Shute & Ke, 2012; U.S. DOE, 2010). The promise of rich process data, available to a degree never before possible, is to enable us to better understand the complex nature of students' learning processes and to subsequently individualize and personalize the educational experience of students (National Research Council (NRC) (NRC), 2013; U.S. DOE, 2010, 2012, 2013).

To realize the potential of players' moment-to-moment data, the data captured—the game telemetry—must be of high quality. The focus of this chapter is the design of game telemetry, which is the first element in the chain of reasoning connecting player behavior to inferences about their learning. Game telemetry is discussed in the context of measuring player performance as players engage in the game. The goal of this chapter is to describe game telemetry and its uses, identify issues related to the use of game telemetry for measurement purposes, provide guidelines on the design of game telemetry, and present a case study that implemented the guidelines to derive measures from game telemetry.

2 Game Telemetry and Its Uses

Game telemetry is the data associated with specific game events, the state of a game, or other parameters of interest. We use the term *telemetry* to connote the systematic specification, capture, and logging of events that occur in a game (i.e., player-initiated or game-initiated events) or game states to a permanent external store using a predefined record format. We do not mean the logging of unstructured output or the ad hoc capture and storage of events that are based on arbitrary criteria or convenience.

The goal driving the collection of game telemetry is to develop cognitively meaningful measures from a combination of player behaviors and game states. By *meaningful* we mean that the measures should (a) help researchers and game designers interpret why players are performing the way they are, and (b) exhibit a systematic (e.g., statistical) relationship with complementary measures; differentiate between players with different degrees of content knowledge, different degrees of game experience, or different backgrounds (e.g., language skills); and differentiate between players who receive different instructional treatments or different game designs. Our major assumption is that player behavior—what players do at a specific point in the game—is a manifestation of their ongoing cognitive and affective processes (e.g., knowledge, judgment, decision-making, problem-solving, self-regulation, self-efficacy, beliefs, and attitudes).

We conceptualize player-level data at two levels. First is data that are associated with a player's background, performance on some task, degree of prior knowledge, attitudes, demographics—coarse-grained data that is traditionally gathered over a small number of time points and often group administered (e.g., in a classroom). The second type of data is event data that captures the interaction between the player and the game. While game telemetry is confined to the game, we note that the combination of both types of data can provide a richer understanding of players' performance (Chung, 2014).

2.1 Event Data

A fundamental data type useful for telemetry purposes is the event data type (Bakeman & Quera, 2012; Wetzler, 2013). The core components of event data are time stamp, action, and state. The time stamp is when the event occurred, the action is the triggering event, and the state is the information associated with the event. State information is the key contextual information needed to understand the event.

In the context of a game, event data are generated at the time an event of interest occurs. A common event is an action performed by a player (the event), which triggers the data capture. An example of an action is object manipulation (e.g., the player moving an object from one location to another). If we assume the specific object manipulation is useful for understanding player behavior, important context information includes object attributes such as position, object values or state, and the object's unique ID. Object attributes provide information on what the student was interacting with and is informative when the quality of interaction is based on some value of the object. For example, in a math game where a student manipulates objects that embed math rules, the properties of the math object could include the object position, the drag time, and whether the resulting object manipulation is correct or incorrect. This information could be used to estimate the extent to which students appear to be understanding the underlying math concepts. Note that not all actions are important or desirable to capture—the importance of an action is dependent on the game design and intended learning outcomes (Chung & Kerr, 2012).

A second type of data important for understanding player progress is system-related information such as game round, game level, and game-level parameters like resources. System context information helps demark when players progress (e.g., number of rounds completed) and allows segmenting of the data along natural boundaries (e.g., number of errors committed in a level). The inclusion of system information greatly facilitates the post hoc filtering and analysis of the data.

2.2 Uses of Game Telemetry

Three general uses of game telemetry have been reported in the literature. The first use is to increase monetization. In games delivered as part of a subscription service (e.g., XBox or PlayStation Network) or that offer in-game purchases of game-related assets, game telemetry is often used to examine the effectiveness of various monetization strategies. A fuller discussion of these ideas and uses of telemetry is given in Santhosh and Vaden (2013).

The second use of game telemetry is to better understand what players are doing in the game with the focus on modifying the level or game design to improve the play experience. Level designers often strive for the optimal experience—neither too hard (which leads to frustration) nor too easy (which leads to boredom). Hullet, Nagappan, Schuh, and Hopson (2012) examined factors that differentiated new and career players and the usage of various game options, Weber, Mateas, and Jhala (2011) modeled player retention, and Kim et al. (2008) described a system they used to fuse telemetry with players' self-reports. In all cases, the objective was to identify game design elements that could be modified to improve future game experience.

Game telemetry can be represented visually to give designers a visual representation of the data. Heat maps overlaid on the level can be used to show the physical location where player deaths occur and frequency of deaths represented visually. Gagné, Seif El-Nasr, and Shaw (2012) used game telemetry to help answer design-related questions for a real-time strategy game, for example, at what point do players stop playing the game? How often do players lose a level? Are the players doing what the designers expected? Are there specific actions that can be associated with wins and losses? These ideas and uses of telemetry are discussed in detail in Seif El-Nasr, Drachen, and Canossa (2013).

The third use of game telemetry is measurement. Measurement serves an important function in serious games where learning outcomes are an explicit goal. For example, Loh (2011, 2012) describes the use of telemetry to open the "black box" of game-based learning and contrasts in situ with ex situ measures of performance. Telemetry provides in situ measurement that can be used to describe the sequences of events—the process of learning—that players use during the game. Ex situ measurement brackets gameplay and can be used to provide information on players' pre-gameplay skills and knowledge, the impact of gameplay on acquisition of to-be-learned skills and knowledge, whether players attain criterion performance level, or whether players can transfer what they learned in the game to a novel situation. From a measurement perspective, in situ and ex situ measurements are complementary. The added value of in situ measurement is that it can provide information on the processes that presumably underlie the outcome performance.

The concept of in situ measurement is one of the hallmarks of computer-based performance assessments that require students to engage in complex tasks (Baker, Chung, & Delacruz, 2008; Bennett, Persky, Weiss, & Jenkins, 2007; Chung & Baker, 2003; Chung, de Vries, Cheak, Stevens, & Bewley, 2002; Katz & James, 1998;

Koenig, Iseli, Wainess, & Lee, 2013; Mislevy, 2013; Quellmalz et al., 2013; Williamson, Mislevy, & Bejar, 2006). The use of insights and best practices from the measurement community can be directly applied to the measurement in serious games (Baker et al., 2012; Levy, 2013, 2014; Mislevy et al., 2014).

The judicious and systematic instrumentation of the game to record the key player and system events and game states (collectively believed to reflect learning in the game) to illuminate what players are doing and the conditions under which the performance is occurring is the key benefit of game telemetry. The availability of such information, when considered as evidence tied to a theoretical framework, can be a rich source of information about the process of learning, thereby enabling exploratory analyses of the processes occurring in the game (e.g., via data mining procedures) or confirmatory analyses via specific hypothesis testing.

3 Issues in the Use of Game Telemetry for Measurement Purposes

While much progress has been made in the development of algorithms and models using game telemetry (e.g., Baker & Yacef, 2009; Chung & Kerr, 2012; Ifenthaler, Eseryel, & Ge, 2012; Kerr & Chung, 2012b; Loh, 2012; Loh & Sheng, 2014; Mohamad & Tasir, 2013; Romero & Ventura, 2010; Romero, Ventura, Pechenizkiy, & Baker, 2010), less attention has been focused on psychometric issues and data issues.

From a psychometric standpoint, the interactive and open-ended nature of games in general present measurement challenges difficult to model with traditional approaches based on classical test theory, including adjusting for the serial dependence of performance, tasks, and data points; multidimensionality; contingent performance dependent on the problem structure, players' decisions, and system responses; and learning over the course of the task. Games require students to respond to numerous interactive situations that may require different kinds of knowledge at different points in the task depending on choices players make, and often learning and measurement co-occur. The data itself can be generated based on events or by continuous sampling. When data are from educational contexts, the data can have complex hierarchical structures (e.g., interactions nested within games, games nested within an individual, individuals nested within a classroom, classrooms nested within a school, schools nested within a district). These complexities need to be modeled; otherwise, estimates of student learning are likely to be overly optimistic (Cai, 2013; Junker, 2011; Levy, 2013; Mislevy et al., 2014; Mislevy, Behrens, DiCerbo, & Levy, 2012).

In addition to psychometric challenges of game-based measurement, several data challenges have been identified in the literature (e.g., Bousbia & Belamri, 2014; Romero, Romero, & Ventura, 2014; Romero & Ventura, 2007, 2010; Shoukry et al., 2014). First, the lack of standardization makes the output of each learning system unique and thus general tools cannot be used to directly process the data. The lack of standardization also can pose substantial risk to any analysis effort. Problems in the format of the data can require substantial preprocessing, and poor design of the telemetry system can lead to capturing data with low information value (i.e., noise) or not capturing contextual information that can disambiguate the interpretation of various events (i.e., two events that appear to be the same at one level of abstraction resolves to different actions with context data). Decisions in the design of the data format itself, for example, to emphasize human-readability, can inadvertently introduce side effects that result in unstructured data formats. Often 70–80 % of an analyst's time is spent on preparing the data set for analysis (U.S. DOE, 2012).

The Black box nature of data capture means that problems in the data are often discovered during the analysis phase, after the data have been collected. In education and training contexts, shortfalls in the quality of the telemetry are nearly impossible to recover from as recollecting data is generally not feasible (e.g., logistics, costs, and sample contamination). Werner, McDowell, and Denner (2013) provide a painful case study of the challenges confronted when attempting to use telemetry that was not specifically designed to support learning-related questions, and Chung and Kerr (2012) describe an approach that casts data logging as a form of behavioral observation and present a logging framework that has been successfully applied to games.

Finally, an emerging issue is the idea of making the data itself more meaningful. Romero and Ventura (2007, 2010) and others (Bousbia & Belamri, 2014; U.S. DOE, 2012, 2013) recommend that the researchers make better use of the semantic information and educational context information under which the data were captured. These recommendations raise two issues. First, algorithms should be designed to measure the construct as closely as possible instead of relying on what is convenient or easily logged by the system. A coherent design process is needed so that the linkages from the hypothesized construct to features to evidence of those features are explicitly defined and modeled (APA, AERA, & NCME, 2014; Baker, 1997; Cai, 2013; Linn, 2010; Messick, 1995; Mislevy et al., 2014). The second implication is that the broader education context (such as school, district, and community factors) should be incorporated into analyses that examine the effects of various interventions as well as analyses used in validating measures based on data mining. The availability of online public data sets and GIS-based tools makes this recommendation increasingly practical (Tate, 2012).

4 Game Telemetry Design Guidelines

Our perspective on the design of game telemetry flows from the behavioral observation tradition (e.g., Bakeman & Gottman, 1997; Bakeman & Quera, 2012; Ostrov & Hart, 2013). Of paramount importance is that the observations of behavior be systematic—that is, the set of behavioral acts of interest are well defined prior to the observation. Systematic observations require a clear definition, specified a priori, of what to observe and how to code it, a structured sampling method, a reliable method of recording the observation, and high reliability and strong validity evidence of the coding scheme (Bakeman & Gottman, 1997; Ostrov & Hart, 2013). When applying the behavioral observation framework to the design of game telemetry, the two most important properties are (a) the precise definition of the behavior to observe; and (b) the connection between the behavior and a theoretical framework within which the behavior is interpreted. These two components directly address construct validity and they are particularly important in serious games where measures are based on behavior in the game, and the behavior is (somehow) interpreted as evidence of competency on one or more latent constructs. Measuring the processes in a game that presumably lead to learning outcomes is difficult because it is usually the case that it is learning that is of interest rather than the gameplay itself. Thus, typical metrics used in games (e.g., player deaths, number of levels completed) need to be interpreted in light of the level design and how well the game mechanics support the intended learning outcomes. If player deaths result from gameplay unassociated with players interacting with the to-be-learned content, then player deaths may have little connection to learning. Similarly, the use of a particular game mechanic may be of little interest from a gameplay perspective, but it may be of central importance to the measurement of learning. The more directly the game mechanic requires players to interact with the to-be-learned content, the more likely that game mechanic will reflect players' learning processes. The remainder of this section provides guidelines and examples that illustrate components of game telemetry design that flow from the behavioral observation framework.

4.1 Guideline 1: Target Behaviors That Reflect the Use of Cognitive Demands

Because cognitive processes cannot be observed directly, inferences about the use (or nonuse) of a particular cognitive process and the appropriate use (or inappropriate use) of that process can be based only on what learners do in the game—their in-game behaviors and the associated game states (Chung & Kerr, 2012; Drachen, Canossa, & Sørensen, 2013).

The game telemetry specification should target player behaviors that reflect the cognitive demands of interest. By cognitive demands, we mean the set of intellectual skills required of learners to succeed in the game. Examples of broad categories of cognitive demands include adaptive problem-solving, situation awareness, decision-making, self-regulation, teamwork, conceptual and procedural learning of content, and application and transfer of learning. In a game, it is important to conduct a cognitive task analysis that provides insight about the mental operations players invoke during the course of playing the game.

The challenge is in mapping specific in-game behavior to unobservable cognitive processes such that the ambiguity of the datum is minimized. A specific behavioral act can be a manifestation of numerous underlying processes. Judicious structuring of the interaction and the capture of contextual information surrounding the interaction can help eliminate alternative explanations underlying the behavior. Ideally, the design of the game levels and game mechanics will allow only those players who have knowledge of *X* to successfully apply game mechanic *x*. To the extent that is possible, game mechanic *x* becomes a potential measure of *X*.

As an example, if the research question related to a math game asks whether players know that two fractions with unlike denominators cannot be added together (without first converting both fractions to a form with common denominators), then the game should allow and log players' attempts to add unlike denominators, rather than disallow the behavior entirely. Allowing erroneous actions is important because it provides insights into misconceptions (Kerr & Chung, 2012b, 2012c, 2013a, 2013b). Additionally, it is important to know the context in which the attempted addition occurs. An attempted addition of 1/4 to 1/2 when the answer is 3/4 has a different explanation than an attempted addition of 1/4 to 1/2 when the answer is 2/4.

4.2 Guideline 2: Record Data at the Finest Usable Grain Size

By *finest usable grain size*, we mean a data element that has a clear definition associated with it. For example, a data element that refers to "click" is often unusable whereas a datum that qualifies the click (e.g., "clicked on the reset button") is usable. For example, in a fractions math game, logging "attempted addition" is not at the finest usable grain size because some attempted additions have the same denominator and some do not. In this case, the finest usable grain size would be logging an attempted addition with information about the different denominators.

In general, game telemetry should contain sufficient information to describe the context in which the event occurred and in sufficient detail to link the data to a specific school, teacher, period, player, game level, and game state. One way to think about this is to suppose the data were recorded on index cards (e.g., a sorted deck of 150,000 cards composing the game experience of 130 students across 5 teachers, 4 periods, and 4 different versions of the game) and the card deck was dropped: What information would need to be recorded on each index card so that the original card deck could be reconstructed perfectly? Using the same "attempted addition" example, the telemetry would also include the unique ID of the player who made the addition, the game level in which the addition was made, the time at which it occurred, the fraction being added, the fraction it was added to, and any other game state information that would be important in interpreting the specific action.

4.3 Guideline 3: Represent Data to Require Minimal Preprocessing

This guideline may not apply in high volume environments with dedicated programmers and storage constraints. The assumption for this guideline is a typical research environment where programmers may not be available and researchers have little programming experience.

We recommend a *flat file* representation of the data as this format is simple, easy to explain, easily understood, and portable. More efficient and expressive representations are available (e.g., JSON; SQL), but these formats require a reliable connection to a database server and programming skill to extract the data. Regardless of the particular format of the data store, the eventual destination of the data itself is a statistical analysis tool, where often a flat file representation is the easiest format to use for the greatest number of end-users who are generally not programmers.

Our approach has emphasized ease of use by end-users of the data (e.g., the data analyst, researchers) and not computational efficiency. This trade-off is intentional and assumes that multiple data analysts and researchers will touch the data over its life span; thus, making the telemetry format simple and usable is a high priority as shown in Table 3.1.

Table 3.1 Sample flat-file telemetry format

Field	Data type	Description
Serial number	Long integer	Increments from 1 to *n*. Use to uniquely identify each record in the data and to sort the records in the order they were recorded
time stamp	Formatted time of day	The time the data was captured in the following format: mm/dd/yy hh:mm:ss.mmm
game_time	Long integer	The time in seconds since the game was loaded
user_id	Integer	The login ID of the current player
Stage	Integer	The current stage of the game
Level	Integer	The current level of the game
data_code	Integer	The numeric code that uniquely describes this type of data. There should be a 1:1 correspondence between a data code and the type of data logged (e.g., data_description)
data_description	String	A general description of the data being logged by the corresponding data_code
data01	String	data_code specific value
data02	String	data_code specific value
data03	String	data_code specific value
data04	String	data_code specific value
data05	String	Spare
data06	String	Spare
data07	String	Spare
game_state	String	A list of the values that reflect game state (e.g., level configuration, current score, current achievements)

4.4 Guideline 4: Record Descriptions of Behavior and Not Inferences with as Much Contextual Information as Feasible

In general, game telemetry should have the following properties: (a) is a description of behavior and not an inference about why behavior occurs, (b) is unambiguous (i.e., the data point refers to a single event and not a collection of events—the difference between "clicked on button 1" vs. "clicked on a button"), and (c) contains sufficient context information to allow linking of the data element to a specific player at a specific point in the game.

4.4.1 Descriptive

Suppose in a fractions game the game mechanic supports adding two objects where each object represents a fraction. Adding two things incorrectly can be represented descriptively as "incorrect addition" or inferentially as "player does not understand how to add fractions."

The issues with logging inferences are as follows. First, unless validity evidence has been gathered on the specific interpretation, the interpretation may not be accurate. An interpretation layered over the actual event may create restrictions on subsequent data analyses. For example, statements about *what* the player did in the game may not be possible if the data element reflects understanding. Data logged as "does not understand adding fractions" says little about the actual gameplay itself. The inference may subsume multiple events, in which case the subsumed events are unavailable for analyses. This aggregation may lead to uninterpretability of inference data (i.e., an action logged as "student understands adding fractions" immediately followed by "student does not understand adding fractions").

4.4.2 Unambiguous

For maximum flexibility (particularly for statistical analyses), the telemetry should be unambiguous. By unambiguous we mean a 1:1 correspondence between the data element and an event. For example, suppose there are 10 buttons and we are interested in recording button click events. The data should be recorded in such a way to uniquely identify which of the ten buttons was clicked on, as well as support easy aggregation across the ten buttons. The first capability allows us to examine a particular behavioral act, and the latter case allows us to examine a class of behavioral acts. If only the latter capability exists, then there is a loss of information and potentially important behavioral acts may be masked by the aggregation.

4.4.3 Contextualized

The idea of contextualizing data is to encode as much relevant information as possible about the conditions under which the data were generated. The purpose for gathering context information is to rule out alternative explanations for the observed event and in general, to help researchers understand why an event occurred in the game.

Contextual information consists of two classes of information. First, information about the student—background information such as schooling (e.g., school, period, teacher, grade), domain-specific information (e.g., prior knowledge on the topic of the game, game experience), demographic information (e.g., age, sex), and other information that may influence performance and learning in the game (e.g., motivational information). The second class of information is related to the game experience itself. Contextual information during the game can be the values of various game state variables, type of feedback, or any other information that may qualify the data.

5 Case Study: Deriving Measures from Game Telemetry

In general, three types of measures can be derived from gameplay: (a) overall game performance, (b) in-game performance, and (c) in-game strategies. Each type of measure has certain uses and the measure used in an analysis depends on the question being asked. The case study is discussed in the context of a researcher-developed game, *Save Patch*. We first describe the game and its empirical history, and then discuss measures developed from the game telemetry.

5.1 Case Study Game: Save Patch

The game *Save Patch* was designed to teach the concept of a unit in rational numbers (CATS, 2012). The game was designed around two key ideas in rational numbers. The first idea is that all rational numbers (integers and fractions) are defined relative to a single, unit quantity (e.g., a unit of count, measure, area, volume). The second idea is that rational numbers can be summed only if the unit quantities are identical (e.g., 1/4 + 3/4 is permissible but 1/2 + 3/4 is not because the unit or size of the fractions is unequal). These two ideas formed the basis of what we expected to measure from students' gameplay.

The game scenario was to help the character, Patch, move from his initial position to the goal position to free the trapped cat (the cage in the screenshot in Fig. 3.1). Patch could only move by following a path that was specified by ropes, and the distance Patch traveled was determined by the length of the rope segment. Players specify the distance and direction that Patch travels at each sign post by adding rope segments to the sign post.

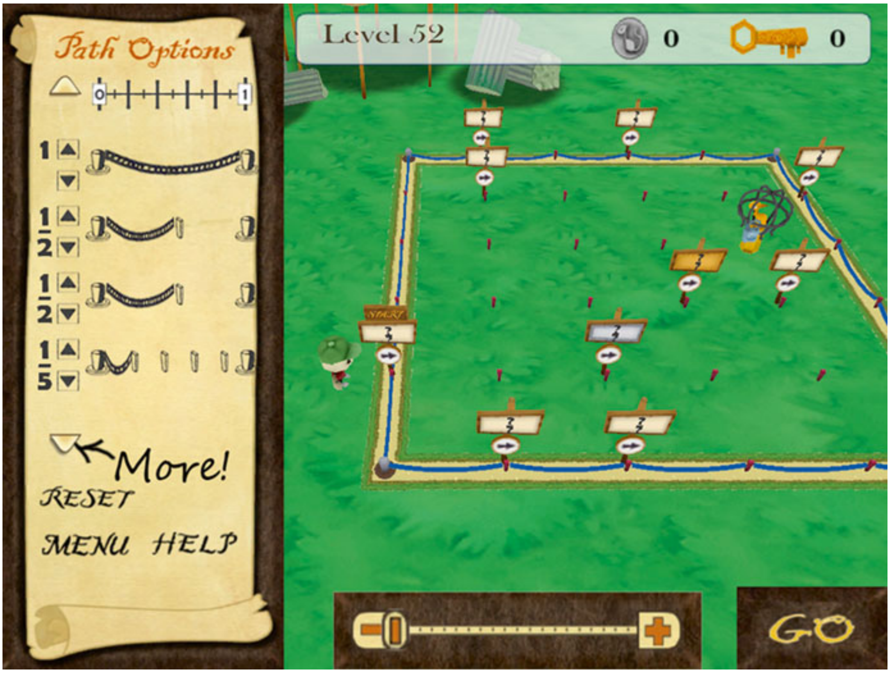


Fig. 3.1 Screen shot of *Save Patch*

Successful gameplay required students to determine the size of the whole unit for a given grid and also the size of any fractional pieces. The second component, additive operations only allowed on like-sized units, was carried out via the game scenario of adding rope segments to the sign post so Patch would travel the appropriate distance. The distance traveled was a function of how many rope segments were added to a sign post. The size of the rope corresponded to a whole unit (1/1) or a fractional unit (e.g., 1/2), and when adding ropes to the sign post, only same-sized rope segments were allowed. This adding operation corresponded to adding fractions with common denominators. A successful solution resulted in Patch traveling from sign post to sign post to the goal position, which mathematically was the sum of all sign post values.

6 Evidence of *Save Patch* as a Learning Game

Save Patch was one game in a suite of four games designed to provide an engaging learning experience for underprepared students in the area of fractions. The effectiveness of the suite of games was demonstrated in a large-scale randomized controlled trial (RCT), where students playing the fractions games outperformed students playing an alternative set of games on a different math topic (effect size of .6) (Chung et al., 2014). *Save Patch* was developed as a testbed and design model for the other games used in the RCT and has been extensively tested in numerous experimental studies testing instructional design options. For example, Kim and Chung (2012) found that conceptual feedback, compared to procedural feedback, resulted in higher scores on a fractions transfer test. Delacruz (2012) found that providing students incentives to use in-game help, compared to no incentives, resulted in higher student performance on a fractions transfer test. Bittick and Chung (2011) found that while a narrative structure around the game character Patch improved students' perceived engagement compared to no narrative structure, there was no difference on math outcome scores. Kerr and Chung (2012a) found a mediation effect of *Save Patch*, suggesting that prior knowledge determined how well students performed in *Save Patch*, and how well students performed in *Save Patch* determined how well they performed on the math posttest. Finally, Kerr and Chung (2013b) found that different types of errors in *Save Patch* were associated with different learning outcomes. Students who had difficulty identifying the unit size were less likely to learn from the game, compared to students who had difficulty identifying the fractional piece size.

6.1 Telemetry Design in Save Patch

The telemetry system in Save Patch was based on the guidelines described earlier. In *Save Patch*, 23 telemetry points are defined for the following categories of information: (a) general information used to describe the conditions under which the game was used (e.g., game build, directory of executable, study condition, student login ID, list of resources, and notes about the game, level, tutorial, and feedback); (b) help system usage; (c) in-game assessment usage (e.g., which assessment item accessed and the player's response); (d) navigation (e.g., which stage and level player advanced to); (e) object manipulation events (e.g., toggled fraction, changed sign direction, scrolled through resources); (f) game states (e.g., player death, level reset, feedback given to student); (g) in-game decisions (e.g., added a rope to a sign post, added ropes incorrectly, closed feedback window).

The most important aspect of the telemetry system is the focus on the behaviors presumed to reflect players' math knowledge (Guideline 1). For example, the telemetry points used to describe overall game performance are player death and level reset. The telemetry points used to describe player strategies are incorrect and correct rope placements. In each case, context information is recorded as well—the value of the rope being added to the sign post, the existing value on the sign post, and the location of the sign post on the gameboard (Guideline 4). This level of abstraction was also determined to be the finest usable grain size (Guideline 3) because it allowed the creation of tokens (or vectors) that could be analyzed in terms of fine-grained behavioral acts (e.g., adding 1/4 to 1/4 at gameboard position 1, 2) as well as in terms of overall occurrence over levels or stages. Chung and Kerr (2012) provide detailed definitions of the various telemetry points and the context information logged. In the following sections, we describe how telemetry was used to measure overall game performance, in-game performance, and in-game strategies.

6.2 Measuring Overall Game Performance

The measure of overall game performance was based on the learning goals for *Save Patch*. The central question we considered was: What behaviors or game states reflect overall achievement in *Save Patch* with respect to learning of the key mathematics content? In *Save Patch*, the overall measure of game performance was the last level reached. This measure reflected a player's progress in the game and reflected the sequencing of content that progressively introduced new content. As seen in Table 3.1, the current level in the game was encoded in each telemetry packet making computing the last level reached trivial.

6.3 Measuring In-Game Performance

Measures of in-game performance were based on an analysis of the cognitive demands required of successful gameplay in *Save Patch*. When designing in-game measures, the two questions we asked were: (a) What in-game behaviors reflect productive and unproductive use of cognitive demand *X*? (b) What behaviors might reflect common errors in the domain?

In *Save Patch*, the in-game performance measures reflected the math knowledge presumably required of the game mechanics. A measure of poor in-game performance was the number of unsuccessful attempts associated with adding fractions operations (e.g., incorrect fraction additions), and overall level performance measures such as the number of level resets and the number of player deaths in a level. Computing these measures was trivial because these telemetry points were uniquely coded [i.e., unique data codes were assigned to each event (see "data_code" in Table 3.1), Guideline 4].

Because the game mechanics were designed to reflect mathematical operations, the use of the game mechanics provided measures of knowledge of the mathematical operations (Guideline 1). Our assumption was that the more directly a game mechanic supported a cognitive operation, the more likely that measure would be sensitive to differences in knowledge. In *Save Patch*, one learning outcome was the idea that only quantities with the same unit can be added together. In fractions, this concept is reflected by addition of fractions with the same denominator. A core game mechanic was adding together objects (e.g., pieces of rope) that represent fractional pieces of a whole unit. The act of adding two pieces was recorded as either a successful addition or an unsuccessful addition. Contextual information such as the value of the numerator and denominator was recorded as well, and if the addition was unsuccessful, where in the solution path the error occurred. The telemetry packet contained information on the nature of the error, when the error occurred, and where on the gameboard the error occurred (Guideline 4). The telemetry packet could be used as part of an aggregated measure (i.e., the number of addition errors in a level) or the telemetry packet could be used to form vectors to be used as part of a data mining procedure (Guideline 2). See Kim and Chung (2012) on the application of survival analysis to examine in-game performance in *Save Patch*, Kerr and Chung (2013a) on the identification of learning trajectories in based on solution attempts in *Save Patch*, and Kerr and Chung (2013b) on the examination of how in-game performance in *Save Patch* mediates the effect of prior knowledge on posttest score.

The key point is that judicious design of the game mechanics to require use of particular knowledge will result in a measure that will be sensitive to the presence or absence of that knowledge. The encoding of the context information in the game telemetry enables the creation of a variety of measures.

6.4 Measuring In-Game Strategies

Compared to in-game performance measures, measures of in-game strategies can be derived from aggregated performance, performance classifications, or other means of describing a player's gameplay over time. The goal of measuring strategies is to be able to summarize how a player's gameplay unfolded over the course of the game level (or other unit of time). Thus, data are gathered over time and subjected to various types of analyses that take order of player events into account (e.g., Markov chain analyses, time series analyses, lag sequential analyses) or sets of co-occurring player events (e.g., cluster analyses, neural network analyses). When we designed measures of in-game strategies, the two questions we asked were: (a) What sets or sequences of in-game behaviors might reflect productive and unproductive use of cognitive demand *X*? (b) What sets or sequences of in-game behaviors might reflect common errors in the domain?

Measures based on the discovery of interesting patterns are more tenuous in that once a pattern is identified the pattern needs to be interpreted in light of the task and the player's presumed knowledge of the domain. As is true of a priori measures, the discovered patterns of player behavior must reflect the targeted knowledge and skills for those patterns to be sensitive to differences in knowledge.

Patterns of player behavior can be identified from game telemetry using data mining techniques such as cluster analysis (Kerr & Chung, 2012b; Merceron & Yacef, 2004; Romero & Ventura, 2007). Cluster analysis groups individual actions into patterns of behavior by determining which actions co-occurred (Berkhin, 2006; James & McCulloch, 1990; Romero, Gonzalez, Ventura, del Jesus, & Herrera, 2009). Two individual actions are considered to belong to the same pattern of behavior (cluster) if they are both made by the same students. Two individual actions are considered to belong to different patterns of behavior (clusters) if the two actions are made by two different groups of students.

In *Save Patch*, our telemetry design enabled the use of cluster analysis for strategy identification because we encoded the major event of interest (e.g., placement of a rope [i.e., adding fractions]) as well as specific contextual information that served to uniquely identify a player action (Guidelines 1, 4). Cluster analysis enabled identification of sets of co-occurring events that reflected the ideal solution (presumably reflecting adequate knowledge of fractions), errors that were consistent with fraction misconceptions, and game strategies that are not mathematical in nature. These groupings were then interpreted, given the level design and targeted math knowledge, as indicators of different strategies students were using to solve game levels. For instance, some players appeared to attempt to solve levels using correct mathematical techniques, others appeared to hold specific mathematical misconceptions, and still others appeared to attempt to solve levels by using "gaming" strategies rather than mathematical techniques. See Kerr and Chung (2012b) and Kerr, Chung, and Iseli (2011) for detailed treatments of the methodology applied to the *Save Patch* data to identify various player strategies. See Levy (2014) on using these strategies as inputs to a dynamic Bayesian network model for diagnostic purposes.

7 Discussion

Our approach to telemetry design is focused solely on supporting the measurement of performance of *play-learners* in a serious game. In serious games, learning of specific content is the desired outcome. The focus on learning outcomes, compared to entertainment or monetization, leads to a different set of design decisions about the game mechanics, what behavior to measure, how to measure that behavior, and how to analyze the resulting telemetry data.

Measuring performance in serious games is often difficult because seldom do direct measures of the desired outcome exist in the game. This situation is far more challenging than analytics for entertainment and monetization, where often the outcome of interest can be directly derived from telemetry with little or no inference (e.g., conversion rate, number of repeat visits, number of in-app purchases). In the case of determining whether learning occurred, evidence of learning must be accumulated from fine-grained game telemetry.

In this chapter, we discussed game telemetry in terms of data types, uses, and its design. The design of game telemetry was developed from the behavioral observation and measurement traditions, which combines two disciplines focused on connecting overt behavior to inferences about learning. A core idea repeated throughout the discussion of game telemetry is the emphasis on having an explicit and coherent connection between the overt, observable behavior, and the latent constructs of interest. The lack of direct measures highlights the importance of having a theoretical framework to situate the game behaviors, game mechanics that exercises the to-be-learned knowledge and skills, and a telemetry design that bakes in validity.

The game telemetry methodology described in this chapter has been adopted by other game efforts at CRESST [e.g., math games for young children (Chung, 2015), physics games for young children (Baker, Chung, Delacruz, & Madni, 2013), and exponent games for remedial college students (O'Neil, Chung, & Williams, 2013)]. Future research will focus on telemetry design with more expressive representations and methods to accommodate continuous behavioral sampling in simulations as well as sensored environments.

Acknowledgments The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C080015 to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education.

References

APA, AERA, & NCME. (2014). *Standards for educational and psychological testing* (2014th ed.). Washington, DC: Author.

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). Cambridge, England: Cambridge University Press.

Bakeman, R., & Quera, V. (2012). Behavioral observation. In H. Cooper (Ed.-in-Chief), P. Camic, D. Long, A. Panter, D. Rindskopf, & K. J. Sher (Assoc. Eds.), *APA handbooks in psychology: Vol. 1. APA handbook of research methods in psychology: Psychological research: Foundations, planning, methods, and psychometrics*. Washington, DC: American Psychological Association.

Baker, E. L. (1997). Model-based performance assessment. *Theory Into Practice, 36*(4), 247–254.

Baker, E. L., Chung, G. K. W. K., & Delacruz, G. C. (2008). Design and validation of technology-based performance assessments. In J. M. Spector, M. D. Merrill, J. J. G. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 595–604). Mahwah, NJ: Erlbaum.

Baker, E. L., Chung, G. K. W. K., & Delacruz, G. C. (2012). The best and future uses of assessment in games. In M. Mayrath, J. Clarke-Midura, & D. H. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 248–299). Charlotte, NC: Information Age.

Baker, E. L., Chung, G. K. W. K., Delacruz, G. C., & Madni, A. (2013, March). *DARPA ENGAGE program review: CRESST—TA2*. Presentation at the ENGAGE PI meeting (Phase II review). Arlington, VA: Defense Advanced Research Projects Agency, Russell Shilling, Program Manager.

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3–17.

Bennett, R. E., Persky, H., Weiss, A. R., & Jenkins, F. (2007). *Problem solving in technology-rich environments: A report from the NAEP technology-based assessment project* (NCES 2007–466). Washington, DC: National Center for Education Statistics.

Berkhin, R. (2006). A survey of clustering data mining techniques. In J. Kogan, C. Nicholas, & M. Teboulle (Eds.), *Grouping multidimensional data* (pp. 25–72). New York: Springer.

Bittick, S. J., & Chung, G. K. W. K. (2011). *The use of narrative: Gender differences and implications for motivation and learning in a math game* (CRESST Report 804). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Bousbia, N., & Belamri, I. (2014). Which contribution does EDM provide to computer-based learning environments? In A. Peña-Ayala (Ed.), *Educational data mining: Applications and trends (Studies in computational intelligence)* (pp. 3–28). Cham, Switzerland: Springer.

Cai, L. (2013). Potential applications of latent variable modeling for the psychometrics of medical simulation. *Military Medicine, 178*(10S), 115–120.

CATS. (2012). *CATS developed games* (CRESST Resource Report No. 15). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K. (2014). Toward the relational management of educational measurement data. *Teachers College Record, 116*(11), 1–16. Retrieved from http://www.tcrecord.org/Content.asp?ContentId=17650

Chung, G. K. W. K. (2015, January). *Updates on final wave of content/outreach: Part II: Learner modeling*. Presentation at the 2015 Ready to Learn Advisors and Partners Meeting. Washington, DC.

Chung, G. K. W. K., & Baker, E. L. (2003). An exploratory study to examine the feasibility of measuring problem-solving processes using a click-through interface. *Journal of Technology, Learning, and Assessment, 2*(2). Retrieved from http://jtla.org

Chung, G. K. W. K., Choi, K.-C., Baker, E. L., & Cai, L. (2014). *The effects of math video games on learning: A randomized evaluation study with innovative impact estimation techniques* (CRESST Report 841). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., de Vries, L. F., Cheak, A. M., Stevens, R. H., & Bewley, W. L. (2002). Cognitive process validation of an online problem solving assessment. *Computers in Human Behavior, 18*, 669–684.

Chung, G. K. W. K., & Kerr, D. (2012). *A primer on data logging to support extraction of meaningful information from educational games: An example from Save Patch* (CRESST Report 814). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education, 59*, 661–686.

Corporation for Public Broadcasting (CPB), & PBS Kids. (2011). *Findings from ready to learn: 2005–2010*. Washington, DC: Author.

Delacruz, G. C. (2012). *Impact of incentives on the use of feedback in educational videogames* (CRESST Report 813). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Drachen, A., Canossa, A., & Sørensen, J. R. M. (2013). Gameplay metrics in game user research: Examples from the trenches. In M. Seif El-Nasr, A. Drachen, & A. Canossa (Eds.), *Game analytics: Maximizing the value of player data* (pp. 285–319). London: Springer.

Drachen, A., Thurau, C., Togelius, J., Yannakakis, G. N., & Bauckhage, C. (2013). Game data mining. In M. Seif El-Nasr, A. Drachen, & A. Canossa (Eds.), *Game analytics: Maximizing the value of player data* (pp. 205–253). London: Springer.

Gagné, A. R., Seif El-Nasr, M., & Shaw, C. D. (2012). Analysis of telemetry data from a real-time strategy game: A case study. *ACM Computers in Entertainment (CIE)—Theoretical and Practical Computer Applications in Entertainment, 10*(3), Article No. 2. doi:10.1145/2381876.2381878.

Girard, C., Ecalle, J., & Magnan, A. (2013). Serious games as new educational tools: How effective are they? A meta-analysis of recent studies. *Journal of Computer Assisted Learning, 29*, 207–219.

Hullet, K., Nagappan, N., Schuh, E., & Hopson, J. (2012). Empirical analysis of user data in game software development. In *Proceedings of the ACM-IEEE International Symposium on Empirical Software Engineering and Measurement* (pp. 89–96). Retrieved from http://dl.acm.org/citation.cfm?doid=2372251.2372265

Ifenthaler, D., Eseryel, D., & Ge, X. (Eds.). (2012). *Assessment in game-based learning: Foundations, innovations, and perspectives*. New York: Springer.

James, F., & McCulloch, C. (1990). Multivariate analysis in ecology and systematic: Panacea or Pandora's box? *Annual Review of Ecology and Systematics, 21*, 129–166.

Junker, B. W. (2011). Modeling hierarchy and dependence among task responses in educational data mining. In C. Romero, S. Ventura, M. Pechenizkiy, & R. S. J. D. Baker (Eds.), *Handbook of educational data mining* (pp. 143–155). Boca Raton, FL: CRC.

Katz, I. R., & James, C. M. (1998). *Toward assessment of design skill in engineering* (GRE Research Report 97–16). Princeton, NJ: Educational Testing Service.

Kerr, D. & Chung, G. K. W. K. (2012a). *The mediation effect of in-game performance between prior knowledge and posttest score* (CRESST Report 819). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kerr, D., & Chung, G. K. W. K. (2012b). Identifying key features of student performance in educational video games and simulations through cluster analysis. *Journal of Educational Data Mining, 4*, 144–182.

Kerr, D., & Chung, G. K. W. K. (2012c). *Using cluster analysis to extend usability testing to instructional content* (CRESST Report 816). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kerr, D., & Chung, G. K. W. K. (2013a). Identifying learning trajectories in an educational video game. In R. Almond & O. Mengshoel (Eds.), *Proceedings of the 2013 UAI Application Workshops: Big Data Meet Complex Models and Models for Spatial, Temporal and Network Data* (pp. 20–28). Retrieved from http://ceur-ws.org/Vol-1024/

Kerr, D., & Chung, G. K. W. K. (2013b). *The effect of in-game errors on learning outcomes* (CRESST Report 835). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kerr, D., Chung, G. K. W. K., & Iseli, M. R. (2011). *The feasibility of using cluster analysis to examine log data from educational video games* (CRESST Report 790). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Kim, J.-K., & Chung, G. K. W. K. (2012). *The use of a survival analysis technique in understanding game performance in instructional games* (CRESST Tech. Rep. No. 812). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Kim, J. H., Gunn, D. V., Schuh, E., Phillips, B. C., Pagulayan, R. J., & Wixon, D. (2008). Tracking real-time user experience (TRUE): A comprehensive instrumentation solution for complex systems. In *Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems* (pp. 443–452). New York: Association for Computing Machinery.

Koenig, A., Iseli, M., Wainess, R., & Lee, J. J. (2013). Assessment methodology for computer-based instructional simulations. *Military Medicine, 178*(10S), 47–54.

Levy, R. (2013). Psychometric and evidentiary advances, opportunities, and challenges for simulation-based assessment. *Educational Assessment, 18*, 182–207.

Levy, R. (2014). *Dynamic Bayesian network modeling of game based diagnostic assessments* (CRESST Report 837). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Linn, R. L. (2010). Validity. In B. McGaw, P. L. Peterson, & E. L. Baker (Eds.), *International encyclopedia of education* (3rd ed., Vol. 4, pp. 181–185). Oxford, England: Elsevier.

Loh, C. S. (2011, September). Using in situ data collection to improve the impact and return of investment of game-based learning. In *Proceedings of ICEM-SIIE 2011, the 61st International Council for Educational Media (ICEM) and the XIII International Symposium on Computers in Education (SIIE) Joint Conference*. Aveiro, Portugal: ICEM-SIIE.

Loh, C. S. (2012). Information trails: In-process assessment of game-based learning. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 123–144). New York: Springer.

Loh, C. S., & Sheng, Y. (2014). Maximum similarity index (MSI): A metric to differentiate the performance of novices vs. multiple-experts in serious games. *Computers in Human Behavior, 39*, 322–330.

Merceron, A., & Yacef, K. (2004). Mining student data captured from a web-based tutoring tool: Initial exploration and results. *Journal of Interactive Learning Research, 15*, 319–346.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*, 741–749.

Mislevy, R. J. (2013). Evidence-centered design for simulation-based assessment. *Military Medicine, 178*(10S), 101–114.

Mislevy, R. J., Behrens, J. T., DiCerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining, 4*(1), 11–48.

Mislevy, R. J., Orange, A., Bauer, M. I., von Davier, A., Hao, J., Corrigan, S., et al. (2014). *Psychometric considerations in game-based assessment*. New York: GlassLab Research, Institute of Play.

Mohamad, S. K., & Tasir, Z. (2013). Educational data mining: A review. *Procedia—Social and Behavioral Sciences, 97*, 320–324.

National Research Council (NRC). (2013). *Frontiers in massive data analysis*. Washington, DC: National Academies Press.

O'Neil, H. F., Chung, G. K. W. K., & Williams, P. (2013). *The effects of game-based instructional feedback on developmental math progress in a Hispanic-serving institution*. Arlington, VA: Office of Naval Research Cognitive Science of Learning Program Review.

Ostrov, J. M., & Hart, E. J. (2013). Observational methods. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology* (Vol. 1, pp. 285–303). Oxford, England: Oxford University Press.

Quellmalz, E. S., Davenport, J. L., Timms, M. J., DeBoer, G. E., Jordan, K. A., Huang, C.-W., et al. (2013). Next-generation environments for assessing and promoting complex science learning. *Journal of Educational Psychology, 105*(4), 1100–1114. doi:10.1037/a0032220.

Romero, C., Gonzalez, P., Ventura, S., del Jesus, M. J., & Herrera, F. (2009). Evolutionary algorithms for subgroup discovery in e-learning: A practical application using Moodle data. *Expert Systems with Applications, 39*, 1632–1644.

Romero, C., Romero, J. R., & Ventura, S. (2014). A survey on pre-processing educational data. In A. Peña-Ayala (Ed.), *Educational data mining: Applications and trends (Studies in computational intelligence)* (pp. 29–64). Cham, Switzerland: Springer.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33*, 125–146.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state-of-the-art. *IEEE Transactions on Systems, Man, and Cybernetics Part C: Applications and Reviews, 40*, 601–618.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. J. D. (Eds.). (2010). *Handbook of educational data mining*. Boca Raton, FL: CRC Press.

Santhosh, S., & Vaden, M. (2013). Telemetry and analytics best practices and lessons learned. In M. Seif El-Nasr, A. Drachen, & A. Canossa (Eds.), *Game analytics: Maximizing the value of player data* (pp. 85–109). London: Springer.

Seif El-Nasr, M., Drachen, A., & Canossa, A. (Eds.). (2013). *Game analytics: Maximizing the value of player data*. London: Springer.

Shaffer, D. W., & Gee, J. (2012). The right kind of GATE: Computer games and the future of assessment. In M. Mayrath, D. Robinson, & J. Clarke-Midura (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 211–228). Charlotte, NC: Information Age.

Shoukry, L., Göbel, S., & Steinmetz, R. (2014). Learning analytics and serious games: Trends and considerations. In *Proceedings of the 2014 ACM International Workshop on Serious Games* (pp. 21–26). Orlando, FL: ACM.

Shute, V. J., & Ke, F. (2012). Games, learning, and assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 43–58). New York: Springer.

Tate, W. F. (2012). Pandemic preparedness: Using geospatial modeling to inform policy in systems of education and health in metropolitan America. In W. F. Tate (Ed.), *Research on schools, neighborhoods, and communities: Toward civic responsibility* (pp. 411–430). Lanham, MD: Rowman and Littlefield.

Tobias, S., Fletcher, J. D., Dai, D. Y., & Wind, A. (2011). Review of research on computer games. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 127–222). Charlotte, NC: Information Age.

U.S. Department of Education (DOE). (2010). *Transforming American education: Learning powered by technology*. Washington, DC: Author.

U.S. Department of Education (DOE). (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief*. Washington, DC: Author.

U.S. Department of Education (DOE). (2013). *Expanded evidence approaches for learning in a digital world*. Washington, DC: Author.

Weber, B. G., Mateas, M., & Jhala, A. (2011). Using data mining to model player experience. In *Proceedings of the FDG Workshop on Evaluating Player Experience in Games*.

Werner, L., McDowell, C., & Denner, J. (2013). A first step in learning analytics: Pre-processing low-level Alice logging data of middle school students. *Journal of Educational Data Mining, 5*(2), 11–37.

Wetzler, M. (2013, June 26). *Analytics for hackers: How to think about event data*. Retrieved from https://keen.io/blog/53958349217/analytics-for-hackers-how-to-think-about-event-data

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.). (2006). *Automated scoring of complex tasks in computer based testing*. Mahwah, NJ: Erlbaum.

Chapter 4
The Dynamical Analysis of Log Data Within Educational Games

Erica L. Snow, Laura K. Allen, and Danielle S. McNamara

Abstract Games and game-based environments frequently provide users multiple trajectories and paths. Thus, users often have to make decisions about how to interact and behave during the learning task. These decisions are often captured through the use of log data, which can provide a wealth of information concerning students' choices, agency, and performance while engaged within a game-based system. However, to analyze these changing data sets, researchers need to use methodologies that focus on quantifying fine-grained patterns as they emerge across time. In this chapter, we will consider how dynamical analysis techniques offer researchers a unique means of visualizing and characterizing nuanced decision and behavior patterns that emerge from students' log data within game-based environments. Specifically, we focus on how three distinct types of dynamical methodologies, Random Walks, Entropy analysis, and Hurst exponents, have been used within the game-based system iSTART-2 as a form of stealth assessment. These dynamical techniques provide researchers a means of unobtrusively assessing how students behave and learn within game-based environments.

Keywords Dynamics • Stealth assessments • Data visualization • Game-based environments

1 Introduction

In this chapter, we discuss how the power of dynamical analyses has the potential to provide researchers with a deeper understanding of students' behaviors within game-based systems and the impact that variations in these behaviors have on learning. The research described in this chapter occurs within the context of iSTART-2 (the Interactive Strategy Training for Active Reading and Thinking-2), an intelligent tutoring system (ITS) designed to support the development of adolescent students' reading comprehension skills (Jackson & McNamara, 2013; Snow,

E.L. Snow (✉) • L.K. Allen • D.S. McNamara
Arizona State University, Tempe, AZ 85281, USA
e-mail: erica.l.snow@asu.edu; laurakallen@asu.edu; danielle.mcnamara@asu.edu

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based Learning, DOI 10.1007/978-3-319-05834-4_4

Jacovina, Allen, Dai, & McNamara, 2014). We first provide a brief overview of the use of log data and dynamical analyses to assess students' behaviors within game-based environments. Subsequently, we describe iSTART-2 and discuss how log data and dynamical analyses have been used as a means of stealth assessment within the context of our game-based environment.

2 The Utility of Log Data Within Game-Based Environments

Computer-based learning environments increasingly incorporate games and game-based features as a means to enhance students' engagement during learning and instruction (Gee, 2003; Johnson et al., 2004; McNamara, Jackson, & Graesser, 2010; Rai & Beck, 2012; Sabourin, Shores, Mott, & Lester, 2012). Although these game-based computer systems vary in their design, structure, and content, a common functionality in many of these environments is the element of user choice. Indeed, many games and game-based environments afford users the opportunity to customize their learning experience by providing them with a variety of choices regarding their potential learning paths. These interactive choices can range from avatar personalization to "choose your own adventure" tasks. Accordingly, users are often required to make decisions about how to interact and behave within the game-based system.

When users are afforded the opportunity to exert agency over their learning path, they will most likely vary in their experiences of the game. Indeed, users' learning trajectories (interaction patterns) vary considerably when they are afforded the opportunity to exert agency within systems (Sabourin et al., 2012; Snow, Jacovina, et al., 2014; Spires, Rowe, Mott, & Lester, 2011). One problem faced by researchers is the *analysis* and *assessmen*t of these interaction patterns, as it can be difficult to quantify the fine-grained changes in users' behaviors. Recently, however, researchers have turned to a novel form of assessment through the use of the log data generated by these systems. Log data has to the potential to capture multiple facets of users' decisions within games, ranging from keystrokes and mouse clicks to telemetry data. Researchers often intentionally program their game-based environments to log all of a user's interactions or choices within the system. When utilized appropriately, this data can provide scientists with a wealth of information concerning students' choices and performance while engaged within game-based systems (Baker et al., 2008; Hadwin, Nesbit, Jamieson-Noel, Code, & Winne, 2007; Sabourin et al., 2012; Snow, Allen, Russell, & McNamara, 2014).

One particular benefit of log-data analyses is that they can act as a form of stealth assessment (Shute, 2011; Shute, Ventura, Bauer, & Zapata-Rivera, 2009). Stealth assessments covertly measure designated constructs (e.g., engagement, cognitive skills) without disrupting the users with explicit tests. In other words, these measures are virtually *invisible* to users. Log data has previously been used as a form of stealth assessment to measure a multitude of constructs, such as students' study habits (Hadwin et al., 2007), self-regulation ability (Sabourin et al., 2012), and gaming behavior (Baker et al., 2008). For instance, Hadwin and colleagues (2007) examined how students varied in their studying patterns within the *g* study system (i.e., a platform designed to aid in students studying behaviors) and how these variations ultimately relate to self-regulative behaviors. This work revealed that log data from students' time within the system was not only predictive of self-regulation, but also captured behaviors that would be missed by traditional self-report measures.

3 Applying Dynamical Analyses to Log Data

Log data generated from game-based systems has proven to be an invaluable assessment tool for researchers. However, researchers have often struggled with ways to quantify patterns that emerge within this type of system data. Indeed, an important goal going forward is for scientists to devise methods for *evaluating* and *quantifying* the variations that manifest within log data. These quantification methods will allow researchers to assess the extent to which behavior patterns can shed light upon students' experiences within game-based environments and how variations in those experiences influence learning outcomes.

Dynamic systems theory and its associated analysis techniques afford researchers a nuanced and fine-grained way to characterize patterns that emerge across time. Dynamic analyses do not treat behaviors or actions as static (i.e., unchanging), as is customary in many statistical approaches, but instead focus on complex and sometimes fluid changes that occur across time. Recently, we have proposed that dynamical systems theory and its associated analysis techniques may be useful for examining behavioral patterns and variations within game-based log data (Snow, Allen, Russell, & McNamara, 2014; Snow, Jacovina, et al., 2014). Current work in this area supports this notion, as dynamical analyses have been successfully applied to log data from adaptive environments to capture the fine-grained behavior patterns enacted by students during various learning tasks (Allen et al., 2014; Hadwin et al., 2007; Snow, Allen, Russell, & McNamara 2014; Snow, Likens, Jackson, & McNamara, 2013; Zhou, 2013). For instance, we have previously used dynamical analyses to classify fluctuations in students' choice patterns within the game-based system iSTART-ME (interactive Strategy Training for Active Reading and Thinking—Motivationally Enhanced; Jackson & McNamara, 2013; Snow, Allen, Russell, & McNamara 2014). These analyses revealed that some students acted in a controlled and decisive manner within the system, whereas others acted more randomly. These behavior classifications would have otherwise been missed without the combination of log data and dynamical analyses.

There are many forms of analysis techniques and methodologies used within dynamical systems theory (Granic & Hollenstein, 2003). The current chapter discusses three of these methodologies (Random Walks, Entropy, and Hurst exponents), which we have used to develop stealth assessments within iSTART-2. First, Random Walks are mathematical tools that generate a spatial representation of a path or pattern that forms within categorical data across time (Benhamou & Bovet, 1989; Lobry, 1996; Snow et al., 2013). This technique has been used in economics (Nelson & Plosser, 1982), ecology (Benhamou & Bovet, 1989), psychology (Allen et al., 2014), and genetics (Lobry, 1996) as a way to visualize changes in patterns over time. Geneticists in particular have used this technique to investigate pairings of genes within gene sequences (Arneodo et al., 1995; Lobry, 1996). Within the context of educational games, this technique can provide a visualization of various learning trajectories or paths within the game. Thus, if students can "choose their own adventure," these tools can provide researchers with a means to track and trace these choices as they manifest across time.

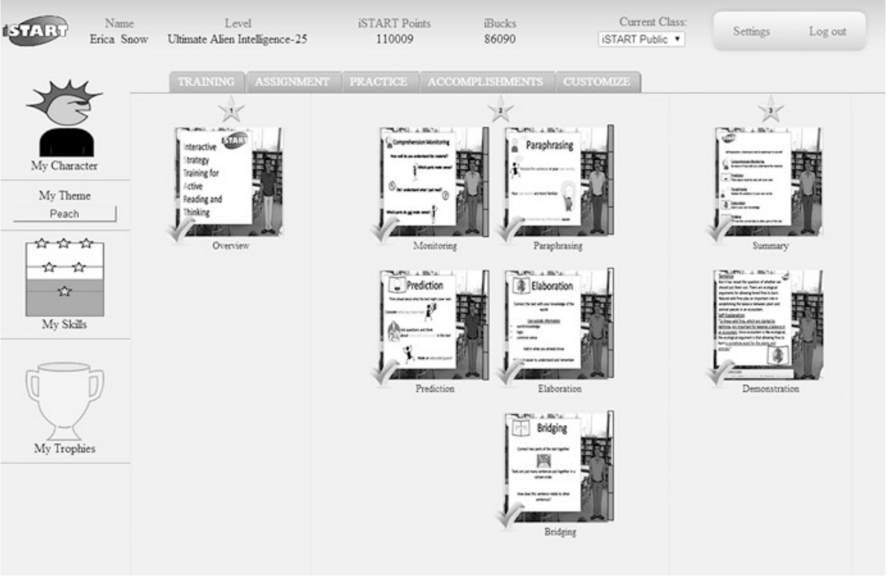
Although Random Walks afford researchers a way to visualize patterns in their data, they do not provide a quantifiable measure of change or fluctuations in those patterns. Thus, other dynamical methodologies, such as Entropy and Hurst analyses, can be used in conjunction with Random Walks to quantify these fluctuations and changes across time. Entropy is a dynamical methodology that originated in the field of thermodynamics (Clausius, 1865) and is used to measure the amount of predictability that exists in a system across time (Grossman, 1953). Specifically, Entropy analyses provide a measure of random (unpredictable) and ordered (predictable) processes by calculating how many pieces of information are contained within a system or time series (Grossman, 1953). Thus, the more information that is present within a time series, the more unpredictable or random the entire series is considered. Similar to Random Walks, Entropy has been used across a variety of domains, from thermodynamics (Clausius, 1865) to linguistics (Berger, Pietra, & Pietra, 1996). Within the context of educational games, this methodology provides a quantifiable measure of the changes in students' behaviors. For instance, if a student makes a variety of different choices within a game, they will produce an Entropy score that contains numerous pieces of information and therefore is indicative of a more unpredictable or random time series. Entropy calculations afford researchers the opportunity to examine the predictability of users' movements and choices within game-based environments.

Similar to Entropy, Hurst exponents (Hurst, 1951) quantify tendencies of a time series. Hurst exponents act as long-term correlations that characterize statistical fluctuations across time as persistent, random, or antipersistent (Mandelbrot, 1982). Persistent patterns are similar to positive correlations, where fluctuations in patterns are positively correlated from one moment to the next. These patterns reflect self-organized and controlled processes (Van Orden, Holden, & Turvey, 2003). In the context of a game, Hurst exponents may be indicative of a student choosing to do the same action or a set of actions repetitively. By contrast, random patterns are said to be independent, where each moment in the pattern does not influence what comes before or after it. These patterns represent a breakdown in control (e.g., Peng et al., 1995). Random patterns within a game could be indicative of a student exploring the interface in an impetuous manner. Thus, the student does not demonstrate a strategy or plan of action. Finally, antipersistent patterns are similar to negative correlations, where the time series demonstrates a corrective process (Collins & De Luca, 1994). These patterns can manifest if a student demonstrates reactive behavior, where their next action within a game is in opposition to what they just experienced. Within the context of educational games, Hurst exponents can provide a fine-grained measure of the relationship between behavior changes. Thus, Hurst affords researchers the opportunity to examine the overall tendency of users' choices within game-based environments. It is important to note the difference between Hurst exponents and Entropy calculations. Hurst exponents capture how each time point (or action) is related to what happens before and after, where correlated actions are considered to be persistent or controlled. Conversely, Entropy provides a quantification of the degree to which the entire time series is predictable versus random.

4 iSTART-2

iSTART (Interactive Strategy Training for Active Reading and Thinking) provides high school students with instruction and practice to use self-explanation and comprehension strategies to understand challenging texts (McNamara, Levinstein, & Boonthum, 2004. It focuses on strategies such as making bridging inferences that link different parts of a text and using prior knowledge to connect the ideas in the text to what the student already knows. When students are provided with instruction to use these strategies, the quality of their explanations improves and their ability to understand challenging texts, such as science texts, is enhanced (McNamara, O'Reilly, Rowe, Boonthum, & Levinstein, 2007; O'Reilly, Sinclair, & McNamara, 2004; Taylor, O'Reilly, Rowe, & McNamara, 2006). iSTART-ME (Jackson & McNamara, 2013) and iSTART-2 (Snow, Allen, Jacovina, & McNamara, 2015; Snow, Jacovina, et al., 2014) are more recent versions of iSTART that provide students with the same comprehension strategy instruction within game-based platforms. These game-based systems were designed to provide adaptive instruction and at the same time enhance students' motivation and engagement through the inclusion of games and game-based features (Jackson & McNamara, 2013).

Within iSTART-2 (see Fig. 4.1), there are two phases: training and practice. Students first engage in training, where they are introduced to a pedagogical agent (Mr. Evans) who defines and explains self-explanation and comprehension strategies and demonstrates how they can be applied to complex science texts. Students are introduced to five comprehension strategies: comprehension monitoring, predicting, paraphrasing, elaborating, and bridging. Each strategy is first introduced and explained in a video narrated by Mr. Evans. At the end of each video, students are transitioned to a *checkpoint*, where they are quizzed on their understanding of the strategy they just learned. After students watch the five lesson videos, they watch a final summary video. In this video, Mr. Evans summarizes the five strategies that the students just learned. Once these videos are completed, students watch as Mr. Evans provides demonstrations on how to combine multiple strategies to better understand complex science texts.



Name Erica Snow

Level Ultimate Alien Intelligence-25

iSTART Points 110009

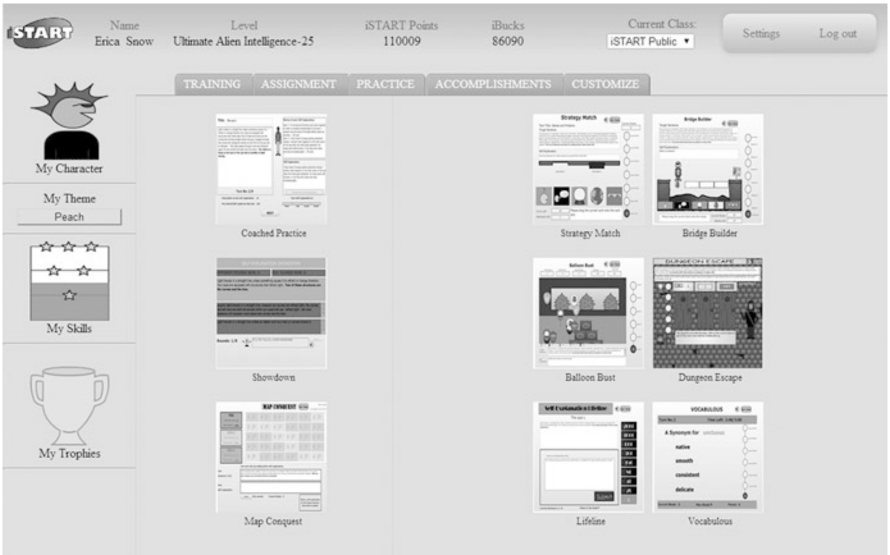
iBucks 86090

Current Class: iSTART Public

My Theme Peach

Fig. 4.1 A screenshot of the iSTART-2 strategy training menu

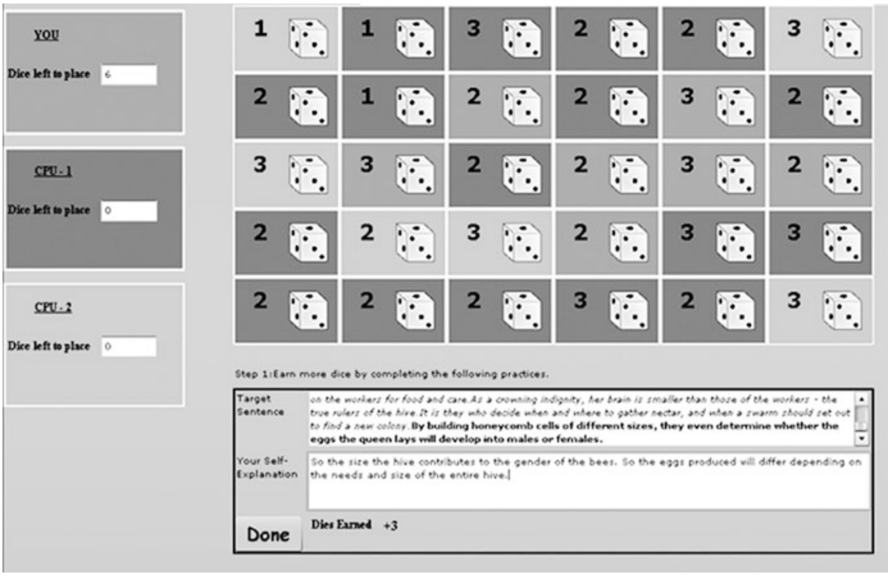
After training, students transition to the practice phase of iSTART-2. During this phase, students engage with an interactive game-based interface, where they can freely choose to self-explain science texts, personalize different aspects of the interface, practice identifying self-explanations within the context of mini-games, or view their personal accomplishments in the system (see Fig. 4.2). Within iSTART-2, there are four different types of game-based features: generative practice, identification mini-games, personalizable features, and achievement screens. *Generative* practice requires students to write their own self-explanations. Within iSTART-2, there are three generative practice environments: Coached Practice, Showdown, and Map Conquest. Coached Practice is a non-game-based method of practice, where students generate self-explanations and then receive feedback from Mr. Evans. Conversely, Showdown and Map Conquest are game-based forms of generative practice. In these games, students generate self-explanations for complex science texts within the context of a game. For example, in Map Conquest, students are asked to generate self-explanations for numerous target sentences. Higher quality self-explanations earn more dice. These dice are then used to conquer neighboring territories (see Fig. 4.3). Students win the game by conquering the most territories; to do this, they must earn a sufficient number of dice by generating high quality self-explanations. Within all three generative practice environments, the quality of students' self-explanations is assessed through an algorithm that relies on both Latent Semantic Analysis (LSA; Landauer, McNamara, Dennis, & Kintsch, 2007) and word-based measures (McNamara, Boonthum, Levinstein, & Millis, 2007).



Name	Level	iSTART Points	iBucks	Current Class:
Erica Snow	Ultimate Alien Intelligence-25	110009	86090	iSTART Public ▾

My Theme: Peach

Fig. 4.2 A screenshot of the iSTART-2 game-based practice menu



YOU
Dice left to place 6

CPU - 1
Dice left to place 0

CPU - 2
Dice left to place 0

1	1	3	2	2	3
2	1	2	2	3	2
3	3	2	2	3	2
2	2	3	2	3	3
2	2	2	3	2	3

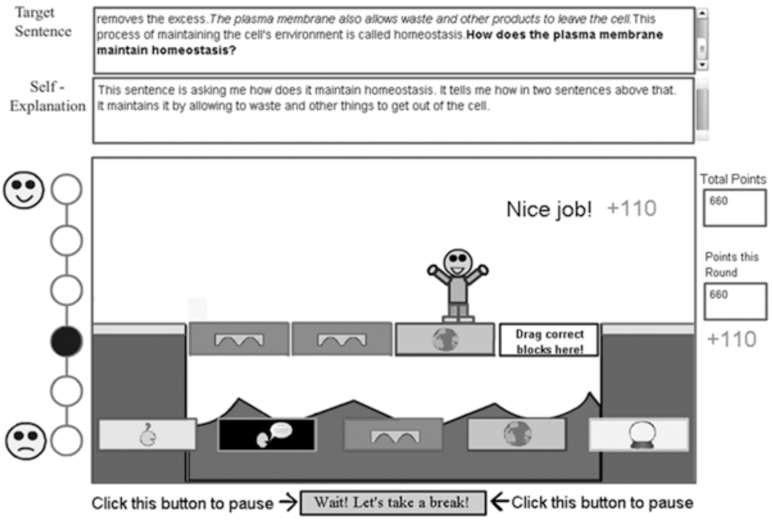
Step 1:Earn more dice by completing the following practices.

Target Sentence: *on the workers for food and care.As a crowning indignity, her brain is smaller than those of the workers - the true rulers of the hive.It is they who decide when and where to gather nectar, and when a swarm should set out to find a new colony.* **By building honeycomb cells of different sizes, they even determine whether the eggs the queen lays will develop into males or females.**

Your Self-Explanation: So the size the hive contributes to the gender of the bees. So the eggs produced will differ depending on the needs and size of the entire hive.

Dies Earned +3

Fig. 4.3 A screenshot of the iSTART-2 generative practice game strategy match



Target Sentence

removes the excess. *The plasma membrane also allows waste and other products to leave the cell.* This process of maintaining the cell's environment is called homeostasis. **How does the plasma membrane maintain homeostasis?**

Self - Explanation

This sentence is asking me how does it maintain homeostasis. It tells me how in two sentences above that. It maintains it by allowing to waste and other things to get out of the cell.

Nice job! +110
Total Points
660
Points this Round
660
+110
Drag correct blocks here!
Click this button to pause →
Wait! Let's take a break!
← Click this button to pause

Fig. 4.4 A screenshot of the iSTART-2 identification game bridge builder

This algorithm scores self-explanations on a scale ranging from 0 to 3, with scores of "0" indicating that the self-explanation is irrelevant and scores of "3" indicating that the self-explanation is relevant, uses prior knowledge, and incorporates information from outside of the text.

Within *identification* mini-games, students are provided the opportunity to practice identifying the five self-explanation strategies. For instance, in Bridge Builder, students are asked to help a man cross a bridge by building the bridge "brick by brick." Each brick represents one of the five self-explanation strategies they have learned. Students are first shown a text and a self-explanation; they must then identify the strategy that was used to generate the self-explanation by placing the corresponding brick on the bridge (see Fig. 4.4). This process repeats until students have helped the man cross the bridge. In total, there are five identification mini-games (see Jackson & McNamara, 2013, for a complete description).

Within iSTART-2, students can earn system points by interacting with texts, either within the context of generative games or identification mini-games. As students collect more points within the system, they subsequently progress through a series of 25 achievement levels (ranging from *Bookworm* to *Ultimate Alien Intelligence*). For students to progress to a new level, they must earn more points than required for the previous level. This mechanic was designed to ensure that students exert more effort as they progress through higher levels in the system. Students also have the opportunity to win trophies in the generative and identification games. These trophies range from bronze to gold and are awarded based on gameplay performance.

iSTART-2 also builds in non-practice game-based features as a way to engage students' interest. These elements include personalizable features and achievement screens. Personalizable features are elements designed to enhance students' feelings of personal investment; they include an editable avatar and changeable background colors. Students can use these elements to customize the system interface. Finally, achievement screens were built into the system to allow students to monitor their progress. Students can use these screens to view their last ten self-explanation scores or any trophies they have won throughout their time in the system.

Overall, the iSTART program has been effective at improving students' use of self-explanations and reading comprehension ability (Jackson & McNamara, 2013). When game-based features are embedded within the iSTART program, students have expressed increased motivation and enjoyment across multiple training sessions (Jackson & McNamara, 2013). Combined, these results suggest that the game-based iSTART system effectively captures users' engagement across multiple training sessions and subsequently improves target skill acquisition.

4.1 iSTART-2 Log Data

Recently, log data from the iSTART programs have been used to develop stealth assessments (Snow, Allen, Russell, & McNamara 2014; Snow, Jacovina, et al., 2014; Snow et al., 2013). This system, like many game-based environments, provides users with agency over their learning paths. Thus, the log data generated from this environment contains a wealth of information regarding variations in students' choices and their influence on learning outcomes.

The log data generated from iSTART-2 contains information about how students interact within the system (choices, time stamps, and language input). For instance, iSTART-2 collects data on every choice a student makes while engaged with the game-based interface. This data provides a detailed list of actions as well as the duration of each action. Table 4.1 provides an example of what this log data looks like. In Table 4.1, there are only five columns (Student ID, Start Time, Stop Time,

Table 4.1 Example log-data from the iSTART-2 system

Student ID	Start time	Stop time	Action	Complete
004	8:45 am	9:00 am	Bridge Builder	Y
004	9:01 am	9:12 am	Map Conquest	Y
004	9:13 am	9:14 am	Avatar Edit	Y
004	9:14 am	9:16 am	Bridge Builder	N
004	9:17 am	9:18 am	Achievement Screen	Y
007	3:00 pm	3:02 pm	Avatar Edits	Y
007	3:03 pm	3:05 pm	Background Edits	Y
007	4:25 pm	4:35 pm	Map Conquest	N
007	4:37 pm	4:45 pm	Showdown	Y
007	4:47 pm	5:01 pm	Balloon Bust	Y

Action, and Complete); however, log data can be much more detailed, as the researcher often dictates the detail of the log data generated from the system. In this simplified example, there are two students (004 and 007) who have each made five choices within the system. The log data presented here reveals the start and stop time of each choice and whether or not it has been completed. This detailed report affords researchers the opportunity to trace each user's learning path within the system. It is important to note that these learning paths constantly vary as iSTART-2 affords users with high levels of agency over their learning path (Snow, Jacovina, et al., 2014).

Although iSTART-2 provides detailed descriptions of each student's interaction path within the system, the log data on its own cannot quantify the variations and fluctuations in behavior patterns that manifest in these data sets. Thus, dynamical analysis techniques are needed to characterize patterns that emerge in this system log data. Because dynamical systems theory treats time as a critical variable, the log data must be first organized chronologically. It is important to note that in order for these methodologies to provide accurate quantifications of users' behaviors, there needs to be some form of time-based classification for each behavior, along with its association with the other behaviors within the system (i.e., chronological or temporal).

4.2 Dynamical Methodologies and Log Data Within iSTART-2

In the following sections, we describe how log data and dynamical analyses can be combined to better understand students' system behaviors. We describe how the three dynamical methodologies discussed earlier (Random Walks, Entropy, and Hurst exponents) have been utilized to covertly assess students' behaviors and the impact of variations in those behaviors on target skill acquisition within iSTART-2. These three techniques provide a novel means of visualizing and categorizing nuances in students' behavior patterns that emerge within log data across time.

4.2.1 Random Walks

Random Walks can provide researchers with a visualization of how students choose to play or interact within game-based environments. These tools are quite flexible, as the researcher can set the parameters and dimensions represented within the walk. For instance, Random Walks have been created that incorporate multiple vectors (Snow et al., 2013; Snow, Allen, Jackson, & McNamara, 2014) and dimensions (Berg, 1993). Indeed, the number of dimensions that can be included when using random walk analyses is, in theory, unlimited. The Random Walks that have been generated for the log data in iSTART-2 have four orthogonal vectors that lie on an X, Y scatter plot (see Fig. 4.5). Each of these vectors corresponds to one of the four

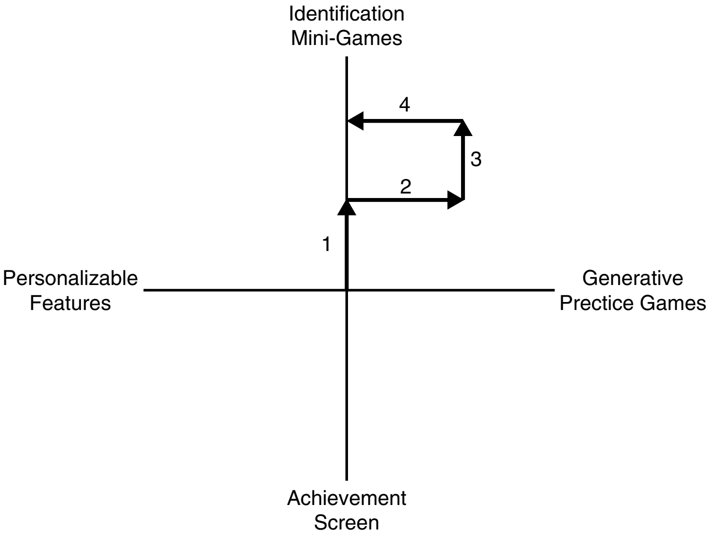


Fig. 4.5 Random walk rule visualization

Table 4.2 Random walk rules within iSTART-2

Game-based Interaction	Movement along *X*, *Y* axis
Generative practice	+1 on *X* axis (move right)
Identification mini-game	+1 on *Y* axis (move up)
Personalizable features	−1 on *X* axis (move left)
Achievement screens	−1 on *Y* axis (move down)

types of game-based features embedded within the system: generative practice, identification mini-games, personalizable features, and achievement screens.

In general, Random Walks follow a set of basic rules that trace movements across categorical data. These rules are predetermined and must stay consistent throughout the entire Walk analysis. Within iSTART-2, these rules dictate how an imaginary particle moves along the *X*, *Y* scatterplot and traces students' movements (i.e., their choice of interactions) between the four orthogonal vectors (i.e., the game-based features). The rules for the Random Walks generated within iSTART-2 are listed in Table 4.2.

Every Walk begins at the origin point (0, 0). An imaginary particle is placed at the origin and only moves after a student has interacted with one of the four game-based features. Every movement of the particle corresponds to the directional assignment established by the researcher. Figure 4.5 demonstrates how the rules described in Table 4.2 would be applied to a student who has made four interaction choices within iSTART-2. This student's sequence of choices is as follows: (1) identification mini-game (move up), (2) generative practice game (move right), (3) second identification mini-game (move up), and (4) personalizable feature (move left).

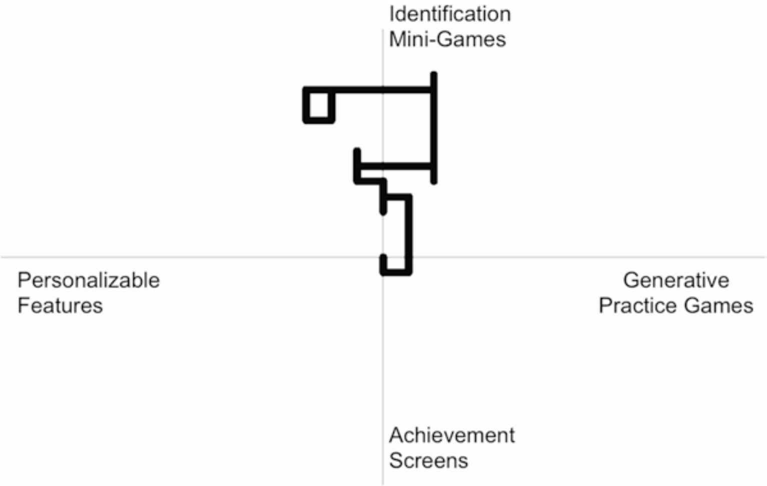


Fig. 4.6 A random walk for one student within the iSTART-2 interface

Random Walks have been applied to over 300 students (across multiple studies) within the iSTART-2 system as a way to visualize various learning paths within the game-based interface. Figure 4.6 reveals what an actual Random Walk looks like for a college student who spent approximately 2 h interacting with the iSTART-2 interface and made 38 total interaction choices. This student's Random Walk provides a visualization of those interactions. From Fig. 4.6, we can see that this student's Walk moved in an upward direction along the *Y*-axis. This indicates that the majority of this student's interactions were with identification mini-games. Indeed, the raw log data reveals that of the 38 total interactions, 22 were with an identification mini-game. Hence, this student's Random Walk provides a means of visualizing fluctuations in these choice patterns as they manifest across time.

Figure 4.6 shows a Random Walk for one student; however, these tools can also be used to visualize differences in interaction patterns (or choices) comparing groups of individuals (Snow, Allen, Jackson, et al., 2014; Snow et al., 2013). For instance, Snow et al. (2013) used aggregated Random Walks to visualize differences in how high reading ability and low reading ability students engaged with game-based features within the iSTART program (see Fig. 4.7). Using this visualization technique, they took the slope of each student's random walk ($n = 40$) and plotted it along the *XY* axis. A median split on pretest reading ability was used to separate students into groups of high reading ability (green slopes) and low reading ability (blue slopes). Results from this visualization revealed that high ability students tended to gravitate more towards identification mini-games whereas low ability students interacted most frequently with the generative practice games (Fig. 4.7). It is important to note that within this random walk, directionality is used only to visualize students' interaction preferences. Thus, Fig. 4.7 reveals that high ability students (green lines) are more likely to select identification mini-games compared

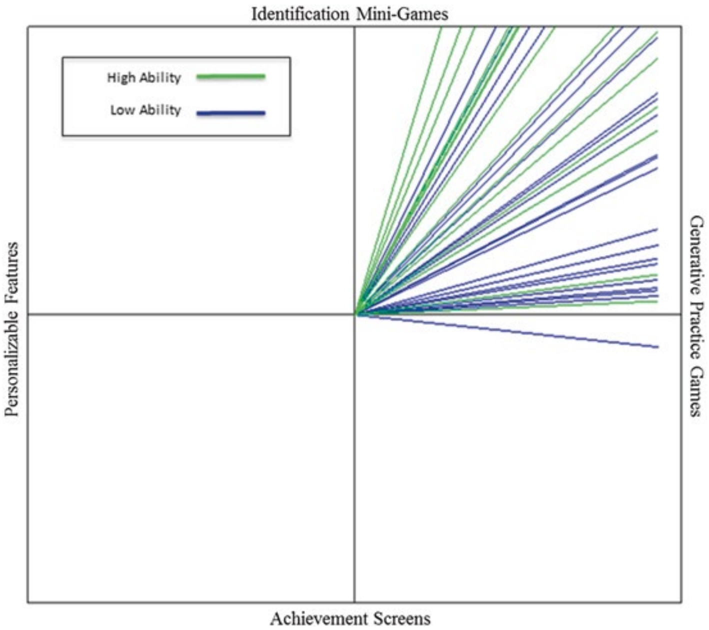


Fig. 4.7 Aggregated random walk for high ($n = 18$) and low reading ability ($n = 20$) students (Figure adapted from Snow et al., 2013)

to the low ability students (blue lines). Overall, Random Walks can be used to trace students' choice patterns within game-based systems. These techniques can be used to track a single student's progress throughout the game or they can be aggregated to provide a visualization of differences in choice patterns comparing two or more groups of individuals.

4.2.2 Entropy

- While Random Walks offer researchers compelling visualizations of students' trajectories within game-based systems, these tools cannot, on their own, quantify variations in choice patterns that emerge across time. Entropy can be used in conjunction with Random Walk analyses to provide an overall quantification of students' interaction patterns. There are many different variations of the Entropy calculation (Bandt & Pompe, 2002; Costa, Goldberger, & Peng, 2002; Shannon, 1951); however, the current chapter focuses on the most widely used Entropy calculation, Shannon Entropy (Shannon, 1951). Equation 4.1 shows the equation for Shannon Entropy. In this equation, $P(x_i)$ represents the probability of a given state (or interaction). In the context of iSTART-2, this formula could be used to analyze log data to calculate how ordered students' choices are across time. Specifically, Entropy for a given student would be calculated by taking the additive inverse of the sum of products calculated by multiplying the probability of each interaction by the natural log of the probability of that interaction. Thus, Entropy scores reflect the degree to which students' interactions within iSTART-2 are ordered (or random) across time. In general, low Entropy scores are indicative of ordered processes, whereas high Entropy scores suggest disorganized or random processes. Thus, if a student's choice pattern is highly organized, they are likely to produce a low Entropy score. Conversely, when a student's choice pattern is disorganized (i.e., interactions within the system are not systematic), the Entropy score will likely be high. Entropy scores are guided by the bits of information presented within a time series. For instance, let's say we flip a fair coin (even probability of heads and tails) twice. If the coin lands on heads both times, Entropy will be zero. Thus, the flip of the coin resulted in uniformed bits of information. However, if we flip the coin and get one heads and one tails, the Entropy of the flips would be 1.0. This is because the maximum possible Entropy increases as the number of possible outcomes (or choices) increases.

$$H(x) = -\sum_{i=0}^{N} P(x_i)\left(\log_e P(x_i)\right) \tag{4.1}$$

Within iSTART-2, we have conducted post hoc analyses using log data in combination with Shannon Entropy to assess how much control students exerted over their learning paths. In one study, it was hypothesized that when students demonstrated higher levels of agency, they would also act in a more controlled and organized manner (Snow, Jacovina, et al., 2014). To test this hypothesis, we conducted a single session study where college students ($n = 75$) freely interacted with the iSTART-2 system for two hours. Every choice that the students made was then categorized into one of the four previously mentioned game-based categories (generative practice identification mini-games, personalizable features, and achievement screens). Entropy analyses were conducted at the end of the study on each student's categorized log data to examine the extent to which the interaction patterns reflected ordered or disordered behavior patterns.

Overall, students varied considerably in their ability to act in a controlled and organized fashion (range = 1.32–2.32, $M = 1.83$, SD = 0.24). Interestingly, results from this study revealed no significant correlation between Entropy scores and the frequency of interactions with any specific feature (i.e., generative practice identification mini-games, personalizable features, and achievement screens). Thus, students' ability to exert controlled interaction patterns was not related to any specific game-based feature. A final analysis examined how variations in students' choice patterns influenced the quality of their self-explanations produced in the generative practice games. A hierarchical regression analysis revealed a significant relation between Entropy and self-explanation quality. Specifically, the students who engaged in more controlled and systematic interaction patterns within iSTART-2 generated higher quality self-explanations than those students who demonstrated disordered behavior patterns.

Entropy analysis has been applied to over 300 high school and college age students' log data (across multiple studies) generated from their time within the iSTART program. This analysis has proven to be a relatively simple way for researchers to examine the overall state of students' choice patterns. Without the use of Entropy, these fine-grained behavior patterns would most likely have been missed. Further, this dynamical methodology can serve as an important stealth assessment when researchers are interested in examining the degree that students' overall behavior patterns are ordered across time.

4.2.3 Hurst

Although Entropy analyses can provide a measure of how ordered students' choices were within a game-based system, this measure does not capture how each choice within the pattern relates to the other choices proceeding and succeeding it. The Hurst exponent has the ability to capture these nuanced fluctuations as they manifest across time (Hurst, 1951). In our recent work, we have calculated Hurst exponents using a *Detrend Fluctuations Analysis* (DFA). A DFA estimates Hurst exponents by first normalizing the time series (or interaction pattern). Once this data is normalized, it is divided into equal time windows of length, n (which may vary for each student). Every window is then fit with a least squares line and the resulting time series is *detrended* by subtracting the local trend of the respective window. This is then repeated as the windows increase exponentially by the power of 2. For each window, a characteristic fluctuation $F(n)$ is calculated; this is the average fluctuation as a function of window size. Finally, $\log_2 F(n)$ is regressed onto $\log_2(n)$, the slope of which produces the Hurst exponent, H. The resulting Hurst exponent ranges from 0 to 1 and can be interpreted as follows: $0.5 < H \leq 1$ indicates deterministic behavior trends, $H = 0.5$ indicates random behavior trends, and $0 \leq H < 0.5$ indicates antipersistent behavior trends.

Within iSTART-2, Hurst exponents have been used in conjunction with log data to examine how fluctuations in students' learning paths influence self-explanations quality (Snow, Allen, Russell, & McNamara 2014). Using this methodology, we were interested in examining how deterministic (and random) patterns of interactions within the game-based environment influenced self-explanations quality (similar to the results from the Entropy analyses). Hurst exponents were calculated for over 80 students (across multiple studies) within the iSTART program. Each of these students spent at least 8 h within the game-based environment and engaged in approximately 275 interactions (i.e., game-based feature choices). Similar to the Entropy analysis, every choice made by students during their time within the system was categorized into one of the four previously mentioned game-based categories and DFA analyses were then calculated using this categorized log data. After the DFA was conducted, each student was assigned a Hurst exponent that quantified the extent to which students' interaction patterns fluctuated in a random or controlled manner.

Results from these analyses revealed that when students engaged in controlled and deterministic patterns of interactions within the game-based system iSTART-2,

Table 4.3 Summary of the benefits and limitations of each methodology

Methodology	Benefits	Limitation
Random walks	Provides visualization of changes in categorical data across time	Cannot quantify variations in choice patterns that emerge across time
Entropy analysis	Provides a statistical measure of the amount of predictability present within a time series or set of interactions	Does not capture how each choice within a pattern relates to the other choices proceeding and succeeding it
Hurst exponents	Provides a long-term correlation of how each choice within a pattern relates to the other choices proceeding and succeeding it	In order to perform a reliable calculation, a large data set with multiple data points is needed

they also demonstrated higher target skill acquisition (Snow, Allen, Russell, & McNamara 2014). The use of Hurst exponents provides researchers with a novel way to look at dynamic movements as they occur over time. Within game-based systems, students are often afforded the opportunity to "choose their own adventure" or personalize their learning path. The Hurst exponents afford researchers a way to examine pattern fluctuations that manifest in students' decisions as they exert agency over their learning path. One limitation of the Hurst exponent analysis is that in order to perform a reliable calculation, a large data set with multiple data points is needed. Although there is no hard-and-fast rule for the exact number of data points needed, in our work, each student completed an average of 275 choices. Understandably, this amount of data may not be readily available for most games. However, one way to combat this issue is to use the Entropy calculation, which requires fewer temporal data points. Although Entropy and Hurst do not measure the same constructs, they are both designed to calculate the relative order of a system or series. The difference, as discussed earlier, is that Hurst focuses on movements between choices (more fine-grained), whereas Entropy measures the overall state of the system. Thus, if a researcher wants to examine patterns of choices or behaviors within a game-based system but they have a smaller data set, Entropy can be calculated to glean an overall measure of a behavior pattern. However, when using Entropy, some fine-grained information will be lost that would otherwise be captured with the Hurst. All three of the methodologies presented here are potentially useful to researchers interested in examining how students interact within game-based environments; however; each has their own benefits and limitations. Table 4.3 provides a summary of the benefits and limitations of each method.

5 Conclusion

Game-based systems often provide students with high levels of agency by allowing them to engage in multiple types of interactions and develop an individualized learning path (Sabourin et al., 2012; Snow et al., 2013). Thus, log data from these systems afford researchers with a unique means of tracing variations in students' choice patterns that may emerge across time. On their own, game-based log data can be complex and provide little insight into students' learning processes and cognitive states. However, the work described in this chapter demonstrates that dynamic techniques can shed light upon variations in students' behaviors within game-based systems and the impact of these variations on learning outcomes.

Dynamic systems analysis treats time as a critical variable which affords researchers the opportunity to not only look at *aggregated* information regarding students' interactions in game-based systems, but to also examine the fine-grained behaviors patterns that emerge across time. While the current chapter focused on how dynamic methodologies have been applied to log data from the iSTART-2 system, these techniques are generalizable to a variety of systems. For instance, Allen et al. (2014) have utilized Random Walks to visualize how high school students demonstrated flexibility in their use of various linguistic properties across 16 prompt-based essays (Allen et al., 2014). Similarly, Random Walks and Entropy analyses have been applied as a way to visualize variations in students' interactions within the game-based writing tutor, Writing Pal (Snow, Allen, Jackson, et al., 2014). Indeed, the tools and methods presented here can be used on any temporal log data.

Future work should focus on the practical use of these techniques within game-based environments to capture the emergence of these complex online behaviors. For instance, dynamical methodologies may inform student models in various adaptive game-based environments. Thus, if a student is engaging in a random interaction loop, dynamical methodologies can potentially "flag" this student and the system can then prompt the student to engage in more controlled patterns. Therefore, these analyses serve to inform and provide game-based systems with information about optimal and non-optimal learning patterns.

In conclusion, this chapter describes preliminary work that serves as a starting point for understanding how dynamical techniques can provide a means to trace and classify students' interactions within game-based environments, as well as other environments that offer multiple choices and pathways. All three of the analysis techniques described here (Random Walks, Entropy, and Hurst exponents) have revealed promising results as to how they can inform researchers about the various ways in which students engage with computer-based systems across time. We conjecture that tracing and modeling choice patterns across time will emerge as a key ingredient in better understanding learning processes.

Acknowledgments This research was supported in part by the Institute for Educational Sciences (IES R305G020018-02; R305G040046, R305A080589) and National Science Foundation (NSF REC0241144; IIS-0735682). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES or NSF. We would like to thank all of the members of the SoletLab for their assistance with data collection.

References

Allen, L. K., Snow, E. L., & McNamara, D. S. (2014). The long and winding road: Investigating the differential writing patterns of high and low skilled writers. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 304–307).

Arneodo, M., Arvidson, A., Badełek, B., Ballintijn, M., Baum, G., Beaufays, J., et al. (1995). Measurement of the proton and the deuteron structure functions. *Physics Letters B, 364*, 107–115.

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research, 19*, 185–224.

Bandt, C., & Pompe, B. (2002). Permutation entropy: A natural complexity measure for time series. *Physical Review Letters, 88*, 174102.

Benhamou, S., & Bovet, P. (1989). How animals use their environment: A new look at kinesis. *Animal Behavior, 38*, 375–383.

Berg, H. C. (Ed.). (1993). *Random walks in biology*. Princeton, NJ: Princeton University Press.

Berger, A. L., Pietra, V. J. D., & Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics, 22*(1), 39–71.

Clausius, R. (1865). *The mechanical theory of heat—With its applications to the steam engine and to physical properties of bodies*. London: John van Voorst.

Collins, J. J., & De Luca, C. J. (1994). Random walking during quiet standing. *Physical Review Letters, 73*, 764–767.

Costa, M., Goldberger, A. L., & Peng, C. K. (2002). Multiscale entropy analysis of complex physiologic time series. *Physical Review Letters, 89*(6), 68–102.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York: Palgrave Macmillan.

Granic, I., & Hollenstein, T. (2003). Dynamic systems methods for models of developmental psychopathology. *Development and Psychopathology, 15*(03), 641–669.

Grossman, E. R. F. W. (1953). Entropy and choice time: The effect of frequency unbalance on choice-response. *Quarterly Journal of Experimental Psychology, 5*(2), 41–51.

Hadwin, A. F., Nesbit, J. C., Jamieson-Noel, D., Code, J., & Winne, P. H. (2007). Examining trace data to explore self-regulated learning. *Metacognition and Learning, 2*, 107–124.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers, 116*, 770–808.

Jackson, G. T., & McNamara, D. S. (2013). Motivation and performance in a game-based intelligent tutoring system. *Journal of Educational Psychology, 105*, 1036–1049.

Johnson, W. L., Marsella, S., Mote, N., Viljhalmsson, H., Narayanan, S., & Choi, S. (2004). Tactical Language Training System: Supporting the rapid acquisition of foreign language and cultural skills. In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC).*

Landauer, T. K., McNamara, D. S., Dennis, S. E., & Kintsch, W. E. (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Lawrence Erlbaum.

Lobry, J. R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. *Molecular Biological Evolution, 13*, 660–665.

Mandelbrot, B. B. (1982). *The fractal geometry of nature*. New York: Freeman.

McNamara, D. S., Boonthum, C., Levinstein, I. B., & Millis, K. (2007). Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 227–241). Mahwah, NJ: Lawrence Erlbaum.

McNamara, D. S., Jackson, G. T., & Graesser, A. C. (2010). Intelligent tutoring and games (ITaG). In Y. K. Baek (Ed.), *Gaming for classroom-based learning: Digital role-playing as a motivator of study* (pp. 44–65). Hershey, PA: IGI Global.

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy trainer for active reading and thinking. *Behavioral Research Methods, Instruments, & Computers, 36*, 222–233.

McNamara, D. S., O'Reilly, T., Rowe, M., Boonthum, C., & Levinstein, I. B. (2007). iSTART: A web-based tutor that teaches self-explanation and metacognitive reading strategies. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 397–420). Mahwah, NJ: Lawrence Erlbaum.

Nelson, C. R., & Plosser, C. R. (1982). Trends and random walks in macroeconomic time series: Some evidence and implications. *Journal of Monetary Economics, 10*, 139–162.

O'Reilly, T. P., Sinclair, G. P., & McNamara, D. S. (2004). Reading strategy training: Automated versus live. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Cognitive Science Society* (pp. 1059–1064). Mahwah, NJ: Lawrence Erlbaum.

Peng, C. K., Havlin, S., Hausdorff, J. M., Mietus, B. S. J., Stanley, H. E., & Goldberger, A. L. (1995). Fractal mechanisms and heart rate dynamics: Long-range correlations and their breakdown with disease. *Journal of Electrophysiology, 28*, 59–65.

Rai, D., & Beck, J. (2012). Math learning environment with game-like elements: An experimental framework. *International Journal of Game Based Learning, 2*, 90–110.

Sabourin, J., Shores, L. R., Mott, B. W., & Lester, J. C. (2012). Predicting student self-regulation strategies in game-based learning environments. In *Intelligent tutoring systems* (pp. 141–150). Berlin, Germany: Springer.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal, 30*, 50–64.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction, 55*, 503–524.

Shute, V. J., Ventura, M., Bauer, M., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning. In U. Ritterfeld, M. J. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). Philadelphia, PA: Routledge/LEA.

Snow, E. L., Allen, L. K., Jackson, G. T., & McNamara, D. S. (2014). Tracking choices: Computational analysis of learning trajectories. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 316–319).

Snow, E. L., Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2015). Does agency matter?: Exploring the impact of controlled behaviors within a game-based environment. *Computers & Education, 26*, 378–392.

Snow, E. L., Allen, L. K., Russell, D. G., & McNamara, D. S. (2014). Who's in control?: Categorizing nuanced patterns of behaviors within a game-based intelligent tutoring system. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 185–192).

Snow, E. L., Jacovina, M. E., Allen, L. K., Dai. J., & McNamara, D. S. (2014). Entropy: A stealth assessment of agency in learning environments. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 241–244).

Snow, E. L., Likens, A., Jackson, G. T., & McNamara, D. S. (2013). Students' walk through tutoring: Using a random walk analysis to profile students. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 276–279). Berlin, Germany: Springer.

Spires, H. A., Rowe, J. P., Mott, B. W., & Lester, J. C. (2011). Problem solving and game-based learning: Effects of middle grade students' hypothesis testing strategies on learning outcomes. *Journal of Educational Computing Research, 44*(4), 453–472.

Taylor, R. S., O'Reilly, T., Rowe, M., & McNamara, D. S. (2006). Improving understanding of science texts: iSTART strategy training vs. web design control task. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 2234–2239). Mahwah, NJ: Erlbaum.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General, 132*(3), 331–350.

Zhou, M. (2013). Using traces to investigate self-regulatory activities: A study of self-regulation and achievement goal profiles in the context of web search for academic tasks. *Journal of Cognitive Education and Psychology, 12*, 287–305.

Chapter 5
Measuring Expert Performance for Serious Games Analytics: From Data to Insights

Christian Sebastian Loh and Yanyan Sheng

Abstract Advances in technology have made it possible to trace players' actions and behaviors (as user-generated data) within online serious gaming environments for performance measurement and improvement purposes. Instead of a Black box approach (such as pretest/posttest), we can approach serious games as a White box, assessing performance of play-learners by manipulating the *performance variables* directly. In this chapter, we describe the processes to obtain user-generated gameplay data in situ using serious games for training—i.e., data tracing, cleaning, mining, and visualization. We also examine ways to differentiate expert-novice performances in serious games, including behavior profiling. We introduce a new Expertise Performance Index, based on string similarities that take into account the "course of actions" chosen by experts and compare that to those of the novices. The Expertise Performance Index can be useful as a metric for serious games analytics because it can rank play-learners according to their competency levels in the serious games.

Keywords In situ data • Expert-novice • Action sequence • Performance metrics • Expertise performance index • Similarity measure

1 Introduction

> Let my playing be my learning, and my learning be my playing.—Johan Huizinga

Although *serious games* can be any meaningful use of computerized game/game industry resources whose chief mission is not entertainment (Djaouti, Alvarez, & Jessel, 2011), the original intent of serious games was to take advantage of the PC

C.S. Loh (✉)
Virtual Environment Lab (V-Lab), Southern Illinois University, 625 Wham Drive,
Mailcode 4610, Carbondale, IL 62901-4610, USA
e-mail: csloh@siu.edu

Y. Sheng
Department of Counseling, Quantitative Methods, and Special Education, Southern Illinois University, 625 Wham Drive, Mailcode 4618, Carbondale, IL 62901-4618, USA
e-mail: ysheng@siu.edu

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based Learning, DOI 10.1007/978-3-319-05834-4_5

gaming technology to create *new tools for improving decision-making skills and performance* (see Chap: 1; Loh, Sheng, & Ifenthaler, 2015; Abt, 1987; Michael & Chen, 2006; Sawyer & Rejeski, 2002). Today's serious games are mostly (~90 %) *message broadcasters* created with the purpose to disseminate information or to push a political leaning. This is why most serious games seem to comprise propaganda- and advertisement-like contents about global warming, military recruitment, conservation of energy, advertisements of online degree programs, etc. Alvarez and colleagues called this category of serious games *message broadcasters.*

1.1 Design-Centric vs. Performance-Centric Game Making

Educational games or edutainment that teaches through a one-way communication channel (e.g., show-and-tell) also fall into this category, albeit in a niche called *educative message broadcasters*. Serious games that broadcast messages usually have no use for tools to improve decision-making skills or performance as game makers consider these games to be well designed enough to teach, instruct, and train. We refer to this as the *design-centric approach* to making serious games.

From the start of Serious Games Initiatives in 2004, there have been several attempts to emphasize the need for assessment component to advance serious games (e.g., Kirkley, Kirkley, & Heneghan, 2007; Kirriemuir & McFarlane, 2003; Michael & Chen, 2006; Sandford & Williamson, 2005; Van Eck, 2006), yet very few serious games come with assessment components.

In an article about why serious games should include *debriefing tools* for feedback and learning (i.e., ad hoc/post hoc assessment tools), the editor of *Simulation & Gaming* journal, David Crookall (2010) said,

> Serious games can easily include tools and modules of various kinds to collect data transparently during play. The data can then be processed to provide material for feedback during play, as in-game debriefing, and also made available as part of the end-of-game debriefing… It is relatively easy, technologically, to build in debriefing data collection into game software. Some wonderful debriefing tools can relatively easily be designed with the same imagination and expertise that go into serious game software and graphics. (p. 908)

As Crookall pointed out, such debriefing tools should be built into serious games, not included as an afterthought. Having the end goal of assessment in mind before the game is even developed is important; we refer to this as the *performance-centric approach* to making serious games.

We understand that the serious games industry is a highly diverse community with all types of learning and training games created to meet the needs of various sectors. We are not trying to change what the industry is doing, except to point out the need for a niche of specialized, performance-centric, immersive games that are created with the intention to *improve decision-making skills and training performance* of play-learners. Play-learners are those who train and learn with serious games because the game environment settings demand a play-learner to "play as they learn, and learn as they play." Examples of serious games created for performance improvement include *Virtual Battlespace* (Bohemia Interactive) for the military and *CliniSpace* (Innovation in Learning) for the healthcare sectors, respectively.

This niche of serious games would most likely benefit from a performance-centric approach leading to the creation of *Serious Games Analytics*. The aim of these performance-centric immersive games is to seek ways to raise play-learners' decision-making skills and (work) performance through play, which can include learning-by-doing (Aldrich, 2005), procedural learning (Torrente et al., 2014), discovery learning, simulations, and other forms of training.

Thanks to the dawn of the "Big Data" era and the proliferation of mobile apps, collecting user-generated data through software applications have become increasingly acceptable by the public. Nowadays, it is quite common for games to collect user-generated data (during gameplay) for profiling and monetization purposes. (Having users agreeing to the terms and conditions before they are given access to the game contents would do the trick.) Gameplay data collection is imperative for debriefing and Serious Games Analytics because, without data, there is no way to measure the performance difference, much less improve it. In this chapter, we will explain some of the processes involved in collecting user-generated data for skills and performance improvement with performance-centric immersive games.

2 Working with Users' Data

In order to assess the performance of play-learners using performance-centric immersive games, we must first deal with the issue of measurement. "How do we measure what play-learners really do in the virtual environment and use that information for performance assessment purposes?" In fact, before we can measure what play-learners do in the virtual environment, we must first find a way to collect user-generated gameplay data within a virtual environment. To do this, we need to understand there are two types of user-generated data that can be collected with serious games: ex situ and in situ data.

2.1 Ex Situ Data and Black Box

Ex situ data are collected "outside the system" from which the object or event under observation lives. User survey data (demographics, feedback) are of this category because they are typically collected in the real world and not from within the game environments. Typically, research data are collected ex situ out of convenience or due to constraints. Constraints can include imminent danger to the researchers (e.g., measuring the temperature of the sun), costs (e.g., sending researchers to Mars), size restrictions (e.g., observing a red blood cell in a human body), or Black box conditions where it is impossible to access the innards of a system.

Black box situations are interesting because they can also be viewed as a case of convenience. Instead of searching for the means to penetrate the constraints of the Black box, researchers can simply choose the easy way out by working with what's

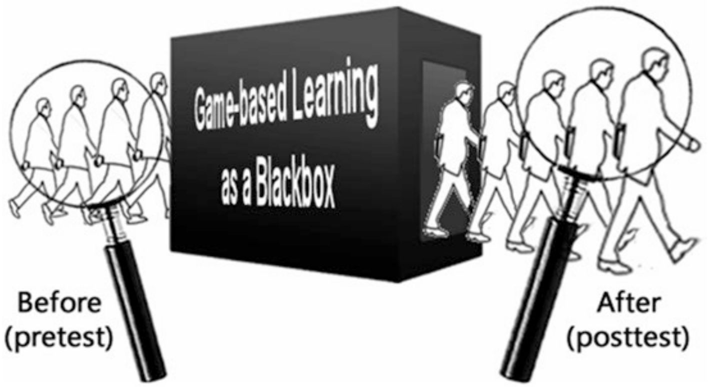


Fig. 5.1 Serious games as a Black box

readily available, the ex situ data. While it may be easier to work with ex situ data, the main disadvantage of this approach is that researchers may never truly know what really occurred within the system. As a result, researchers only obtain an approximation of the event/situation observed.

Researchers unfamiliar with computer programming tend to see digital games as impenetrable Black boxes, and hence, work with ex situ data to find answers for their (research) questions. They are more likely to make use of qualitative protocols, such as: talking aloud, surveys, interviews, focus groups, and analysis of video recordings of the game sessions, to understand why players do what they do in games. These researchers will try to understand the decisions and rationales of players by talking with them, or by having the players explain their actions by reviewing a video recording of their gameplay sessions (similar to After Action Review). This explains why most educator-researchers (who are non-programmers) favor the pretest–posttest methodology for game-based learning research (Bellotti, Kapralos, Lee, Moreno-Ger, & Berta, 2013)—because this is the best methodology in their (somewhat limited) tool box.

In summary, an ex situ approach means that researchers would treat serious games like a Black box and collect user data before *and* after users interact with the serious games (Fig. 5.1). While a change in performance can still be detected through statistical means, the findings from ex situ data cannot effectively explain how game contents affect the performance changes in the play-learners (Loh, 2012a).

2.2 In Situ Data and White Box

Contrary to ex situ, the term in situ means “in place” in Latin. Therefore, in situ data are collected from “within the natural habitat or system” in which an object lives or an event is being investigated. Programming savvy game researchers see serious games as just another software application. Hence, they approach serious games as a White box, open for content manipulation and collection of in situ user-generated data. A good programmer can even create software agents and use them to automate the in situ data collection process.

Obviously, in situ data collection methods are preferable to ex situ ones because they eliminate a lot of "subjective data" obtained from surveys, interviews, and self-reports that simply do not make the cut in high-end research (Quellmalz, Timms, & Schneider, 2009). Once the play-learners' actions have been captured from within the serious games, researchers can then (re)trace what actions players performed within the game, visualize their navigational paths, and make sense of interesting patterns therein (see Loh, 2012b; Scarlatos & Scarlatos, 2010; Thawonmas & Iizuka, 2008).

2.2.1 Behavioral Research Considerations

Because the purpose of researching serious games performance assessment is to understand what play-learners (will) do in serious games, it would obviously require a behavioral analysis approach. As such, we advocate researchers to steer away from self-reports and survey-type evaluations because participants are known to report what they think the investigator wants to hear rather than their own beliefs (also known as *social desirability bias*, Paulhus, 1991). Moreover, people's actions have been found to differ from what they say/claim they would *do*—either due to *over-claiming* (see Roese & Jamieson, 1993), or other reasons (see Fan et al., 2006; Hoskin, 2012). A more objective approach traces what play-learners *actually do* within the game environment directly (i.e., in situ measurement) as empirical evidence. The technique for directly tracing play-learners' actions within a digital game environment is known as *telemetry*.

2.2.2 Telemetry and Information Trails

The term *telemetry* is well known among computer scientists and engineers and has long been associated with remote (ex situ) data collection in the fields of ecology, computer science, biology, and meteorology. For instance, mobile apps make use of *telemetry* to remotely transmit the in situ data collected (from smartphones) to a remote server (over the Internet or Cloud) for storage and analysis.

In serious games, software telemetry is a necessary step for Serious Games Analytics because the technology finally allows game developers and researchers to trace the players' gameplay data without the need for co-location (Joslin, Brown, & Drennan, 2007; Loh, 2012a) and subsequently, to transmit the data to a remote database for storage and analysis. Because there will be no analytics without data, serious games researchers need to understand what *telemetry* is and how to use the technique effectively to collect user-generated data (Moura, Seif El-Nasr, & Shaw, 2011; and Chap. 8 in this volume: Liu, Kang, Lee, Winzeler, & Liu, 2015).

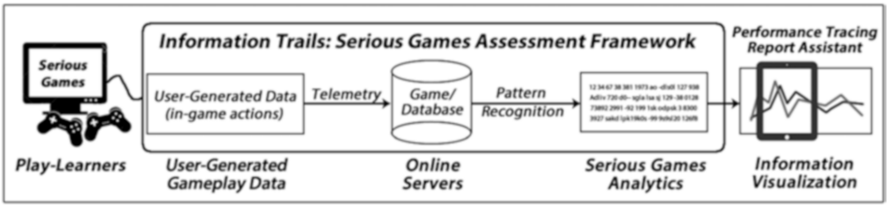

Serious Games
Play-Learners
Information Trails: Serious Games Assessment Framework
User-Generated Data (in-game actions)
Telemetry
Game/ Database
Pattern Recognition
Performance Tracing Report Assistant
User-Generated Gameplay Data
Online Servers
Serious Games Analytics
Information Visualization

Fig. 5.2 Serious games assessment framework: Information Trails

2.3 The Information Trails *Assessment Framework*

More than just telemetry, *Information Trails* (Loh, 2006, 2012b; Loh, Anantachai, Byun, & Lenox, 2007) is a serious games assessment framework (from the field of Instructional Design and Technology) that takes into consideration the need for in situ data collection, telemetry, data mining, and data visualization for performance assessment. There are two parts to *Information Trails*: the online Serious Games data collection framework and a separate visualization component called *Performance Tracing Report Assistant* (PeTRA), as shown in Fig. 5.2.

In order to enable in situ data collection, serious games must be *online*. By this, we mean that the serious games need to be hosted on game servers, allowing players to login (and be tracked) from different locations around the world. Even though it is possible to obtain gameplay data as an exported log file from certain (standalone) games (e.g., Portal 2) because researchers must take the extra steps to: (a) retrieve log files from players, which turns this into an ex situ exercise, and (b) convert/parse the log file into a database before data cleaning and analysis. It would be easier to use online serious games from the onset. But the greatest disadvantage of ex situ assessment is that the process made it impossible for real-time ad hoc assessment reporting. In comparison, Information Trails overcome this limitation: using online in situ data collection to enable ad hoc and post hoc reporting.

One added advantage of using online serious games in the workflow is that most of the analysis processes would already be in place should the opportunity come to migrate to (massively) multiplayer online games (MMOG) for “big data.” To maintain industrial compatibility, researchers should choose either a MySQL server, or an online streaming database server, which is often co-located with the data analysts for easy access to data. This is the telemetric process that is currently being used by MMOG companies to store and track players’ personal data (including credit card information), gameplay data, and in-game transactions.

If such technology to trace user-generated data already exists in the MMOG industry, why are serious games lagging in this area? Crookall (2010) explains that “the problem is that debriefing does not appear to be quite as sexy as the flashy game ware that is usually touted as the game… Funders usually do not understand that learning comes from processing the game experience—that is, in the debriefing. Funders therefore do not see the need to pay for what they see as irrelevant or useless code” (p. 908). Until debriefing and performance assessment tools can rise out of the doldrums of irrelevance, game makers might defer putting in the man-hours needed to create these tools.

Entertainment game makers are lucky because they found a way, namely monetization, to convert user-generated data into post-sale revenue (Moura et al., 2011). Although we are doubtful that monetization would drive serious games, we are fairly certain that something has to happen to break the current stalemate to advance serious games into the next phase. In fact, Crookall's suggestion that serious games clients should insist on having the debriefing or performance assessment tools "be built in as an integral part of both the software and the procedures for running the game" seemed probable.

2.4 Event Listeners

Serious games often take the form of storytelling and role-playing. For example, you may be playing as an Afghan elder who needs to make tough decision to support the American soldiers, or work against them (e.g., *CultureShock: Afghanistan*), a Transportation Security Administration (TSA) agent who need to identify suspicious items from X-Ray images of luggage (e.g., *Checkpoint Screening*), or an American soldier who has to make friends with villagers in a foreign enemy territory (e.g., *Tactical Iraqi*).

These types of serious games all made use of gaming and learning/training events. Gaming events include storylines that provide play-learners with appropriate contexts to draw them into the game. Whereas learning/training events are incidents to problem-solve that are designed to raise critical thinking skills and performance. In the case of *CultureShock*, the training is to empathize with the people of a foreign culture to try and understand their culture and their lives. In the case of *Checkpoint Screening*, the purpose of the training is to increase efficacy of TSA agents to ensure speedy checking without holding up the line, while correctly identifying suspicious items. Lastly, in the case of *Tactical Iraqi*, the aim is to learn to converse in a foreign language (Iraqi) within a short time.

For the game to know what gaming or learning events have occurred, an *event listener* function is necessary. Almost all game engines come with some kind of event listener(s) for the program to keep itself abreast of the myriad of programming events within the system. Task analysis or decision-tree analysis should be performed to properly identify the gaming/learning events for tracing. These events will eventually become nodes in the decision paths taken by the play-learners while interacting with the serious games content. The *event listening* function is an essential part of the serious games analytics equation and should be incorporated into the game engine if possible, or as early in the game development phase as possible.

An example of an event listener can be as simple as an invisible floor trigger. Let's say, in particular serious games, play-learners must reach three checkpoints as part of the training. By placing three invisible floor triggers, A, B, and C, at three different checkpoints, the game system will know the instance a play-learner reach a checkpoint (when he or she stepped on a trigger), and which checkpoint was reached (A, B, or C). Depending on the needs of the researcher, the trigger can be made to keep track of additional pertinent information such as play-learners' names, time of arrival, checkpoint reached, and direction of entry. In the case of a multi-player game, the individual names of play-learners who have triggered gaming or learning events can also be recorded.

2.5 Event Tracers

Working hand-in-hand with the event listener is the *event tracer* function. If the event tracer is to take note of which events got triggered in the game system, then the event tracer is the recorder of those trigger events. An *event tracer* puts a time stamp on the triggered events and injects a permanent entry of the record into the database. In this manner, any triggered event can be recorded and the decision can be jointly determined by the researchers, analysts, and game designers. In other words, taking cues from the *event listeners*, the purpose of the tracer function is to forward the information obtained from the triggered events and place them into the database as permanent records.

A traced event can contain much information, which likely include time-stamped user-actions, game-world coordinates, game variables, health points, item banks, conversation paths, etc. Researchers can also use the tracer function to insert dummy remarks such as "Quest 1 begin" and "Quest 1 end" into the database automatically. During analysis, the time stamp difference between the two dummy remarks would yield the time taken to complete Quest 1, for example.

The in situ data collection process should occur unobtrusively in the background without interfering with gameplay. In other words, serious games assessment should, ideally, be integrated and invisible to the play-learners (Shute & Ventura, 2013)—another reason why pretest/posttest, self-reports, and ex situ data does not work well as serious games assessment. However, being able to record play-learners' actions and behaviors in situ is only half the battle. These actions and behaviors should be convertible into performance metrics that can be shown to measure performance differences and/or good return of investment (Loh, 2012a).

2.6 Data Mining Processes

Once you have the user-generated data you need, the data in the database server should be subjected to a series of data mining processes to produce analytics and actionable insights. Because the gameplay data directly reflect play-learners' in-game decisions and actions, analysis of these data can reveal many insights, including learners' beliefs, behaviors, thought processes, and problem-solving strategies.

By data mining processes, we mean: (1) data storing, (2) data cleaning, (3) data analysis, and (4) data visualization. Stored raw data should be cleaned before use. The data cleaning process commonly involves removing duplicate or extraneous data and/or the filling in of any missing data. It may also be necessary to recode variables before more meaningful information can be gleaned about the players' in-game actions and behaviors. One example is to calculate the duration of an event from raw time stamps, such as taking the time difference between "Quest 1 End" and "Quest 1 Begin" records to calculate how long play-learners took to complete Quest 1.

Once cleaned, the database can then be exported into an XML, or CSV, flat file, and then be imported into any suitable statistical program (e.g., SAS, SPSS, R, or MATLAB), or dedicated data mining package (e.g., JMP by SAS, AMOS by SPSS, Tableau, WEKA) for detailed analysis. Depending on the data mining procedure used, analysts may be able to profile play-learners' characteristics (Thawonmas, Ho, & Matsumoto, 2003), map changes in players' attitudes (Scarlatos & Scarlatos, 2010), categorize patterns of gameplay (Wallner, 2013; Wallner & Kriglstein, 2012, 2013), detect hidden patterns of user behaviors (Drachen, Thurau, Sifa, & Bauckhage, 2013), or compare the (dis)similarity between expert and novice players (Loh & Sheng, 2013, 2014, 2015).

We prefer a quantitative approach to data analysis for SEGA because quantitative methodology is easily automated. Quantitative analysis is also faster than qualitative analysis because the latter requires manual labor (e.g., transcriptions). Some researchers may want to conduct both types of analyses to obtain a spectrum of findings, but given limited time and resources, the choice should be obvious. Compared with qualitative methodologies, quantitative methodologies have greater power of generalization and better cost/benefit ratios, albeit short on the personalization required in many User eXperience (UX) and educational research.

2.7 Information Visualization

The need for a graphical instead of textual presentation of research findings has long been known (see Anderson, 1957; DeSanctis, 1984). It is not surprising that the final, and most important, step in the data mining process would be that of information (or data) *visualization.* In the visualization phase, the analytics (information) are transformed into appropriate graphical forms—never as raw data or log files—for easy communication and discussion by stakeholders who need not understand statistics (see Wallner & Kriglstein, 2013). When done correctly, visualization can reveal information otherwise unobtainable through traditional statistical analysis. In comparison to spatial visualization with GIS programs, gameplay data visualization is very much in its infant stage (for more examples, see Drachen & Canossa, 2011).

For example, *Information Trails* has a visualization component called PeTRA. One of its capabilities includes reporting players' navigational paths traced

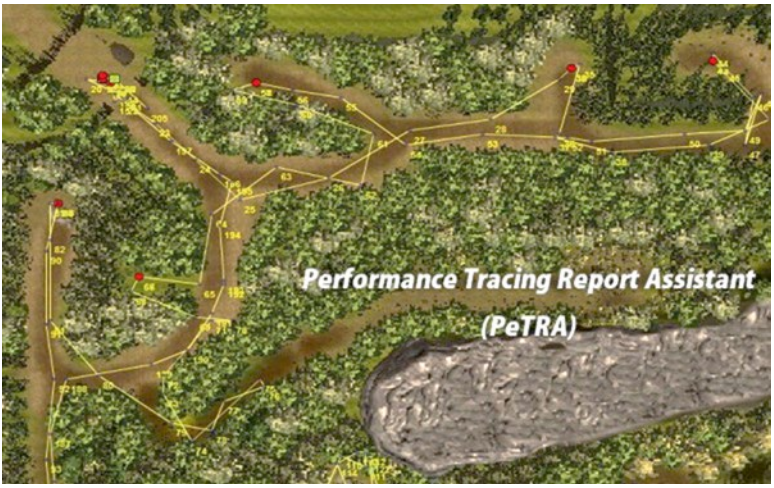


Fig. 5.3 Player's navigational path as revealed in Performance Tracing Report Assistant (PeTRA), the visualization component for Information Trails

during gameplay (Fig. 5.3). PeTRA was designed from the ground up to support formative and summative assessment, thus capable of both ad hoc and post hoc reporting. In ad hoc reporting, the gameplay map trace is generated and displayed in real time. As an external reporting assistant, the purpose of PeTRA is to communicate Serious Games Analytics to stakeholders, who are typically interested in the insights for policy making.

Information visualization is a field of study in its own right and increasingly includes new approaches to visualize spatial and temporal data for reporting and communication purposes (e.g., Kim et al., 2008; Medler & Magerko, 2011; Moura et al., 2011; Thawonmas & Iizuka, 2008). Operationally, the visualization of analytics frequently takes the form of dashboards for easy communication with stakeholders.

3 Collecting User-Generated Data

Great care should be taken to determine what user-generated data would yield meaningful information and what data should be passed over. First, we would like to caution researchers *not* to *over trace*. Researchers should be highly cognizant of the data type and how much information they plan on collecting for the purpose of serious games analytics.

3.1 Big Data vs. Good Data

Researchers have a notorious tendency to over collect data. While any amount of data can be traced, especially in the wake of the age of "Big Data," *over tracing* (e.g., recording every keystroke and mouse-click) can be detrimental to serious games. Not only is over tracing a complete waste of precious computer processing cycles, but an enormous data set can also make analysis much more difficult. Asking the game server to record too many events, too frequently, creates bottlenecks in network traffic and causes lags in game playing (if you are using an online serious game server as we have recommended).

You may fall into the trap of trying to compensate for the network lags by investing in expensive network and server equipment. All these are unnecessary if you would only spend some time considering what kind of data are really needed before collecting them. It is a balancing act that depends very much on what is capable in terms of network technology, computer hardware, and game engine. This may be counterintuitive to readers who think "big data" in serious games, but while *Big Data* can be very tempting, good data is even better.

As mentioned in Sect. 1.1, until MMO serious games become available, it may be too early to discuss about *Big Data*. Researchers and data analysts should learn to put in place the correct procedures until such time. So when *Big Data* do arrive, the data analysis processes could be easily "scaled" to meet the demand.

3.2 Repetition and Behaviors

Having cautioned readers about over tracing, we will now show, by way of examples, how in situ user-generated data can bring about the discovery of interesting information about the play-learners. The key of user-generated data is to find out (gather data about) what play-learners really *do* while inside the game: how did they interact with the game interface, what actions did they choose, how game events affected their decision-making processes, etc. The ultimate goal in this process is to learn what kind of play-learner actions or behaviors lead to an increase in performance.

In one of our game studies (Loh & Byun, 2009), we gave players specific instructions to avoid direct confrontation with a bear (because the animal could easily kill the player when confronted). We discovered that self-professed gamers were the most likely to ignore orders. In contrast, non-gamers (especially female players) were more likely to obey orders and avoid the unnecessary conflict. We believe this is because (male) gamers are so used to being "challenged" in games that they regard the notice to avoid conflict as an invitation-to-try—meaning, they are supposed to try and find a way to kill the bear. This finding showed that direct instruction could easily have an unintended, or reversed effect on players.

In another short text-based game created at the Virtual Environment Lab (V-Lab), a non-player character named Denise was found sleeping in the middle of the road. Players were given a choice to either (a) turn around because Denise blocked the way forward or (b) kick Denise. We purposely refrained from providing any reason for kicking Denise, choosing instead to leave it to the players' imaginations. To our surprise, we found that almost all female players choose to turning around, while male players would have happily kicked Denise.

We presented the same two choices to those who chose to kick Denise—i.e., (a) turn around and (b) kick Denise (again). An even more disturbing trend was observed: male players would happily kick Denise again, and again (and again), displaying fairly aggressive behaviors. When players' repeat their actions over time, this can be detected as a pattern or trend, turning actions into player behaviors (Drachen, Thurau et al., 2013; Thawonmas & Iizuka, 2008; Wallner & Kriglstein, 2013). Our advice for serious games designers is to try to establish player behaviors by presenting multiple opportunities to repeat an action.

3.3 Providing (More Than) Enough Game Actions

Before players' actions can be measured, game designers must first design different ways for the players to solve problems in the games, while the programmers write the functions to enable said actions. For example, in one study, Shute (2011) recorded a particular game event (note: see Sect. 4.1.1 on the limitations of Bayesian Network), requiring players to decide on how to cross a river full of killer fish in *Oblivion*. Shute determined that there were five different methods, namely: (a) swim across river, (b) (cast a spell to) levitate over the river, (c) (cast a spell to) freeze the river and skid across the frozen river, (d) find a bridge over the river, (e) dig a tunnel under the river.

While it appeared that players came up with ingenious ways to solve problems in this given situation, readers should understand that if such game actions were not provided for by the game designers and programmers, there would be no way for the players to execute them. For instance, can a player chop down a tree and use it as a log bridge to cross the river? Is it possible to place large rocks as stepping stone into a shallower part of the river for crossing? Can the killer fish be killed using a poison spell? Can a player fly across the river riding on a dragon? Why isn't there a teleportation spell in *Oblivion* to make crossing the river easier like the portal gun in the game, *Portal*? The possibilities is endless. These game options are most likely not possible in *Oblivion* because they have not been included by the game designer. Compared to some of the fantastic ideas of solving problems in entertainment game, designers of serious games may choose to stick to the real-world solution: Finding a bridge to cross such a river would be considered: (a) logical: what people actually do in the real world, (b) safest: avoid being eaten by killer fish, and (c) most practical: staying dry—albeit meaningless in a digital environment.

3.4 Game Design and Players' Behaviors

Players' behaviors in serious games that are created with the intention to raise performance are notably different from those permissible in entertainment or fantasy games because most, if not all, that a player is allowed to do must resemble what happens in reality. In the real world, all actions carry real-world consequences, some may even result in death. For example, the US National Transportation Safety Board reported that many disastrous airline accidents could be linked directly to flight simulators that only trained pilots to land airplanes in good weather conditions (Levin, 2010). As a result, these pilots were not able to perform the correct actions to land a plane in adverse weathers!

Understanding players' actions and behaviors can help us discover new ways to improve game design. This is what the field of human–computer interaction (HCI) has been trying to tell us for many years (see for example Bellotti, Kapralos, Lee, & Moreno-Ger, 2013; Nacke, Drachen, & Göbel, 2010). To summarize:

1. Simplistic instructions may have unintended and even reversed effects in games.
2. The personality of play-learners can interfere with how instructions are interpreted and which actions are thereby carried out (For example, most female players would choose not to kill in games; *gamer girls* behaved like male players.)
3. The creativity of play-learners is limited by the possibilities of game actions allowed by the designer.
4. Players' actions in entertainment games may not have the same "value" as actions in serious games.

3.5 Game Metrics

Since every industry is different, performance metrics from one industry would be quite different from that of another. For example, a healthcare serious game might trace how many patients successfully recovered for outcome research purposes (Jefford, Stockler, & Tattersall, 2003), while a military serious games may tally how many enemies were successfully defeated (Pruett, 2010). The appropriateness of a performance metric is highly dependent on the aim and the performance outcome required of the (serious) games (Drachen, Canossa, & Sørensen, 2013).

Despite the differences between industries and the kinds of serious games they might commission, there are generally useful metrics across the board for training and learning industries. Given the technological foundation of serious games, it should not surprise anyone that many of these generic metrics were, in fact, borrowed from the fields of HCI, UX and Computer Science.

3.6 Validity of Gameplay Time in Serious Games Research

While reviewing other serious games performance assessment research, we noticed an underreported validity issue related to *gameplay time* in many studies. We found researchers either failed to report the length of gameplay time or employed very short gameplay (less than 10 min) in their studies (Byun & Loh, 2015). For instance, Grimshaw, Lindley, and Nacke (2008), as well as IJsselsteijn, de Kort, Poels, Jurgelionis, and Bellotti (2007) both reported gameplay sessions lasting <10 min as the research condition. In some cases, researchers simply reported the gameplay period using the number of seconds as the reporting unit instead of (the more appropriate) hours and minutes (e.g., O'Rourke, Butler, Liu, Ballweber, & Popovic, 2013). This kind of reporting can be misleading because while 300 s of gameplay may seem reasonable at first glance, it translates to just 5 min—hardly sufficient time for *flow* (Csikszentmihalyi, 1991) or meaningful gameplay engagement (Ermi & Mäyrä, 2007) to occur. Our own studies indicated that 1–2 h of gameplay to be viable for serious games studies involving engagement, without the play-learners becoming fatigue.

The short gameplay period also deviates from real-world practice because players are supposed to spend many hours playing games that they find engaging or motivating. Game companies often published "suggested number of gameplay hours" for their products, with a typical range of 40–60 h. Newer games that take place in a big world (e.g., Skyrim and Far Cry 4) may require even more—e.g., some players have suggested that it may take more than 100 h to complete every mission in Far Cry 4.

Serious games performance assessment studies with 10–15 min of gameplay in a single session are "warm-up" sessions, at best. They hardly qualify as legitimate performance measurement research using serious games. Our experience (Loh & Sheng, 2014, 2015) indicated that 1–2 h of gameplay *per session* to be a much more appropriate time frame for serious games research—without participants becoming bored or fatigued.

3.7 Time of Completion

As suggested in the previous section, gameplay time is an important factor of consideration in serious games research. In fact, one of the most widely used performance metrics in HCI and UX research is that of *time to completion* (Canossa & Drachen, 2009; Smith & Du'Mont, 2009).

The "best time" concept was a useful metric in HCI and UX to measure how long users actually took to complete a given task. Digital games borrowed this metric (equivalent to "speed") and used it as a criteria for the Leaderboard for many first-person shooters, maze, and puzzle games. In such cases, players must compete against themselves, other players, or the game AI (Artificial Intelligence) for a spot on the high-score chart based on how fast they can clear game levels.

While the concept of "best time" (equivalent to "speed") is very intuitive and often makes an effective performance metric for entertainment games, the appropriateness of speed is highly dependent on the learning situations and tasks involved. More specifically, in scenarios where play-learners must think critically before applying their skills or knowledge in problem-solving, speed can actually be detrimental to skill learning.

Research has shown that people who worked (or played) under the pressure of time were often tempted into making hasty decisions and/or taking chances (see Ben Zur & Breznitz, 1981; Pieters & Warlop, 1999; Young, Sutherland, & Cole, 2011). There is little evidence supporting the positive correlation between the speed of completion with the quality of training; indeed, our research indicated time of completion could be negatively correlated with performance (Loh & Sheng, 2014; Loh, Sheng, & Li, 2015). In the workplace, overemphasis of speed can lead to workers rushing to complete a task prematurely.

3.7.1 Caution for Gamification

Gamification and serious games that reward the fastest worker who completes a "job level" may result in risky behaviors and poor decision habits, possibly leading to workplace disasters if left unchecked (Wickens, Stokes, Barnett, & Hyman, 1993). In the end, stakeholders may end up dealing with more costly choices to undo the "damage" caused by bad serious games and to (re)train the play-learners the correct way (Loh, 2012a). It goes to show that not all gamification concepts are appropriate for learning and training.

3.8 Creating New Metrics

Besides time to completion, other prevalent game metrics include the number of kills (i.e., enemy killed), the amount of gold collected, and experience points gained. Occasionally, game designers may break new ground and devise their own metrics either to measure gamers' performance or to better rank them for placement on Leaderboards. One such metric is the *rate of achievement*, which is a hybrid metric created by combining two metrics: (a) number of missions achieved and (b) time period—in this case, *rate* = (a)/(b).

Although such metrics can be rather creative, there is no guarantee that they are suitable for serious games. After all, entertainment and serious games hail from two, very different, domains and are created for different purposes. It is only logical that new metrics need to be crafted for serious games, in order to take advantage of the (more serious) gameplay and to track the skills and performance improvement in play-learners.

Since many training-oriented serious games mimic workplace events (be it health care, military, corporate, or industrial), the serious games contents and instructions should resemble or reflect much of the real world. By tracing, studying, and understanding what play-learners would do (or their course of actions) given certain workplace or training scenarios—e.g., disaster preparation, the data should yield insights for both the game designers and the game users. The same insights from the Serious Games Analytics are, in fact, useful for the production team (game companies) as well as the clients they aim to train. The production team and game designers could learn from the insights and use them as feedback to improve the design of serious games. The game companies or data analysts (hired by the stakeholder) could then convert the user-generated data into analytics or insights to improve the skills and performance of trainees or play-learners.

3.9 Three Different Analytics for Serious Games: Gaming, Testing, and Training

Generally, a two-step progress is needed to convert player-generated data into analytics: (a) tracing the actions of play-learners as they interact with the problem space—be it digital or serious games, as evidence of their cognitive process, skills, and abilities, and (b) analyzing the action sequences obtained by way of statistical or machine learning. Currently, there are roughly three different groups that are interested in game-related analytics:

1. A diverse group of researchers interested in growing *Game Analytics* as a field, where some are interested in advancing the computing technology and game design aspects, and others are interested in the monetization methodologies (see Seif El-Nasr, Drachen, & Canossa, 2013).
2. A diverse group of researchers interested in *Learning Analytics* as a field. They are likely to be associated with the Intelligent Tutoring Systems (i.e., Educational Data Mining) or the Educational Testing industry. The latter group is highly interested in turning digital games into a *testing and measurement* tool where players' *responses* in the test environment can be measured as *performance*. They may use any digital game for their purposes and are not necessarily limited to serious games.
3. This diverse group of researchers perceives Serious Games as a tool for training and raising performance—e.g., the US military, health care, business training. These researchers are interested in all aspects of new metrics and methods to improve "training performance" (not testing and measurement performance), including visualization, engineering, human factor, training, instructional design, etc. They use Serious Games to train and need Serious Games Analytics to improve the design of the game, for the purpose of training performance improvement. (Readers are referred to *Serious Games Analytics—Theoretic Framework*, Chap. 1 in this volume for a longer treatise on the differences among the groups.)

In the following sections, we will examine a number of metrics and cutting edge methods that are being adapted for performance measurement with serious games.

Since various groups use Serious Games for different purposes and agendas, the methods mentioned here (and other found in the rest of this volume) are not specific/restricted to any group.

4 User Performance Measurement for Serious Games

Since Serious Games Analytics is still very much in its infancy, there has not been any clear cut way of categorizing available "analytics" research. We will begin by describing a statistical/machine learning method and follow-up with suggestions as to which of the abovementioned groups of researchers would likely benefit from said method. For example, some of the methods stop short at player profiling. This is because the researchers' original intention was to get a rough idea of their client-base, in regard to how many constituents there were. The method was likely conceived by the Game Analytics group, although the same method can also be used by the Learning Analytics and Serious Games Analytics groups for exploratory purposes.

All in all, any method towards Serious Games Analytics needs to culminate in *(actionable) insights*—i.e., implementable strategies to improve gaming, testing, training performances through the (re)design of serious games, (re)training of play-learners, and remediation of poor performance.

4.1 Decision Analyses by Bayesian Network

A Bayesian Network is a type of probabilistic graphical model, which can simultaneously represent a multitude of relationships between a set of variables in a system. The term was first coined by Judea Pearl in 1985 and has since spawned several varieties: e.g., Bayes(ian) Net(work), Bayes(ian) Model, Belief Network, and Bayesian Belief Network. Researchers represent the conditional relationship (edges) between a set of variables (nodes) using a directed acyclic graph (DAG) and calculate their associated Bayesian probability. If n is the number of the parent nodes, the Bayesian probability of any given node in the DAG is $2n$ and the relationships can be depicted using a *conditional probability table* for the True and False values (see Heckerman, 1995).

Figure 5.4 shows a very simple Bayesian Network with three related nodes [Rain, Sprinkler, Wet Grass] and their corresponding conditional probability tables. For example, the probability (Pr) of finding wet grass, given that the sprinkler was turned off, and that it has rained earlier is:

$$\Pr\left(\text{Wet Grass} = \text{true} \mid \text{Sprinkler} = \text{false}, \text{Rain} = \text{true}\right) = k$$

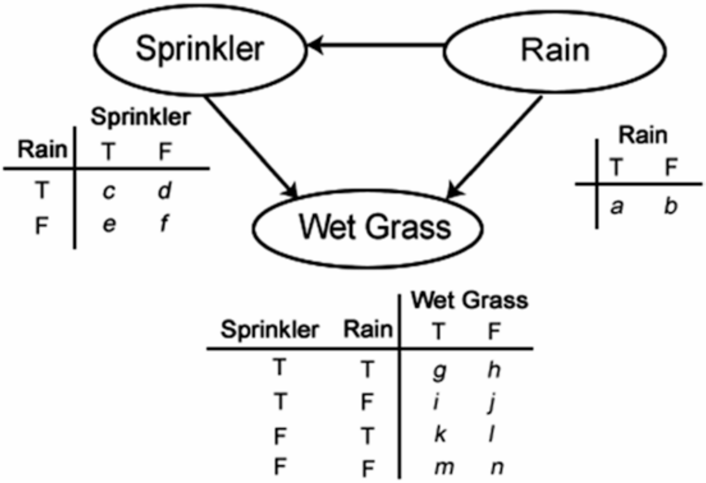


Fig. 5.4 The directed acyclic graph (DAG) of a simple 3-node Bayesian Network

Bayesian Network can be very versatile in modeling the causal and probability relationships of any set of variables. It has been found useful in modeling decision-making systems (Díez, Mira, Iturralde, & Zubillaga, 1997; Oniśko & Druzdzel, 2013) and epistemic games (Rupp, Gushta, Mislevy, & Shaffer, 2010). In these cases, experts are first consulted to create the DAG (what they believed would happen, hence, the name Belief Network) for the decision-making process. The Bayesian probability is then calculated over several iterations of the system to gradually update the *initial model* (prior probability) with new observed occurrence (posterior probability). As the system stabilizes, the researchers will have a model depicting the probabilistic relationships between variables.

4.1.1 Bayesian Networks Are Computationally Prohibitive

Since gameplay is, largely, a series of player decisions, researchers have tried using Bayesian Network to depict the belief systems in game-based learning for "assessment" (Shute et al., 2010). However, because calculation of the Bayesian Network is nondeterministic polynomial-time hard (or NP hard), the approach is considered computationally prohibitive. This could be why many DAGs reported in the Bayesian Network for game-based learning studies depicted *shallow reachability* with very few parent nodes.

In addition, the conditional probability table was seldom reported fully because the number of probability entries required to populate the table increases exponentially (2^n) with the number of parent nodes (n). For example, a node with just four parent nodes would require $\left(2^4 = 16\right)$ entries. This value quickly increases to more than a thousand for a node with 10 parent nodes, and more than a million for a node with 20 parent nodes! This gets even more mindboggling when one considers how many different DAGs (Bayesian Network model) can be produced out of *m*-number of variables because the number of possible Bayesian networks increases super-exponentially:

$$f(m) = (-1)^{i+1} C_i^m 2^{i(m-i)} f(m-i)$$

For example, where $m=3$, possible number of DAGs is: 25; for $m=5$, the possible DAGs increase to 29,281! (A full table is available at: Robinson, 2007, p. 230, and www.bayesnets.com.)

These types of prohibitive computations (even for computers) encouraged researchers working with Bayesian Network to resort to interesting methods to keep the number of variables down. For example, some researchers may claim "domain knowledge" to justify a (simplistic) DAG model created, or modularize the games into standalone rooms with very few choices (see Chap. 12 in this volume: Folkestad et al., 2015). The main idea is to restrict the scope of the gameplay by looking at just one game level, or a single game event (e.g., bridge crossing in *Oblivion*, as reported in Shute, Ventura, Bauer, & Zapata-Rivera, 2009), in order to keep the calculation manageable. Such issues may, unfortunately, preclude Bayesian Network from being used in Big Data research, especially when the serious games industry is increasingly moving towards Serious MMOs. (See Kickstarter initiative by Immersed Games to build *Tyto Online*: www.kickstarter.com/projects/immersed/tyto-online-learning-mmorpg)

As these game modules, rooms, or levels resemble standalone problem spaces akin to multiple-choice test questions (with options to choose from), the assembly greatly endeared Bayesian Network to the Educational Testing industry. For this reason, we contend that the *performance assessment* of game-based learning with Bayesian Network is truly meant for *measurement and testing*, rather than for *training performance* improvement. The confusion in terminology has much to do with the intent of the games, as much game-based learning remains in educative broadcasting and are not created with performance improvement in mind (see Chap. 1 in this volume: Loh, Sheng, & Ifenthaler, 2015).

Thus, even though Bayesian Network has become well established in *testing assessment measurement* (Bauer, 2002), it is not entirely clear as to how the findings can be translated to produce *actionable insights,* which are strategies for remediation or (re)training to raise performance—where testing is not the primary intent.

4.1.2 Limitations of Bayesian Network

Bayesian Network has other limitations when used in conjunction with serious games assessment. Firstly, Bayesian Network is difficult to interpret. For example: "What do the probabilities mean in the real world?" "How does one interpret these probabilities as actionable insights to improve performance?"

Secondly, the type of variables (discrete or continuous) used in the DAG for Bayesian Network can make a big difference. While most of the theory and available models include only discrete variables, models with continuous variables exist in practice. However, a major tradeoff is that once continuous variables are included to create a hybrid Bayesian Network, the model is no longer precise, but an approximation (for more details, see Cobb, Rumí, & Salmerón, 2007).

The third problem is more severe and related to the quality and extent of the prior belief used to depict the expert-created DAG. Niedermayer (1998) explained, a Bayesian Network is "only as useful as this prior knowledge is reliable." This means that delinquent game players (who do not perform task as imagined by the designer) or bugs in the game systems (players doing unexpected actions) could cause the Bayesian Network to fail. More importantly, a wrong model will result in faulty interpretations, which could lead to further problems, even disasters (e.g., pilot training models which led to the airline accidents described in Sec. 3.4).

4.1.3 Inability to Handle Spatial–Temporal Gameplay Data

In today's market, digital game developers are constantly pushing the technology envelope to create bigger game worlds for play (e.g., Skyrim, Dragon Age, World of Warcraft, E.V.E. Online). To trace players within this massive game world, much of the user-generated data are of the spatial–temporal nature. Through these spatial–temporal data, designers are able to pinpoint the exact (spatial) locations of the players and the temporal duration of the gameplay either for troubleshooting during game development or for UX studies pre-game release.

Since the creation of Bayesian Network predated the serious games, it neither understands nor takes into consideration spatial–temporal variables. This may be the biggest downfall for Bayesian Network because it is unable to measure when and where the skills have been acquired or training objectives have been met. Newer research methods that take full advantage of the spatial–temporal gameplay data are needed for serious games analytics. Such methods include movement trajectories analysis (Thawonmas & Iizuka, 2008), game path analysis (Dixit & Youngblood, 2008), GIS (Geographical Information System, Drachen & Canossa, 2011), Expertise Index (Loh & Sheng, 2015), and others.

Given the many innovative approaches to assess training performance in serious games, it is increasingly unclear what *actionable insights* the Bayesian Network models could provide in relation to serious games training. Until new research becomes available to address these concerns, we felt that while Bayesian Network may be suitable for the Educational Testing industry, the serious games industry should look elsewhere for a more fitting model to assess *training* performance for improvement.

5 Performance Measurement and Player Behavioral Profiling

To measure performance assessment with serious games, one needs to observe and trace the play-learners' actions—specifically, what they would do in certain scenarios, and not what they claimed they would do, as evidence of their skills gained from the training/learning. Using data mining and behavior categorization techniques (Moura et al., 2011), these user-generated actions can be aggregated into patterns that are not easily detected using traditional methods. Based on the combinations of behavioral categories obtained, researchers can then develop player (behavioral) profiles using supervised and unsupervised machine/statistical learning techniques, to train them to produce new policies or insights; such as improvement of future game design, formulation of new strategies to (re)train/remediate, monetization, and others.

5.1 Machine/Statistical Learning

Because data analysts would have little to no idea on how to cluster the play-learners, they must first try to reduce the dimensions of the user-generated data into more manageable segments. This is known as the *data exploration* stage, where data analysts make use of unsupervised machine/statistical learning (or segmentation) techniques to divide the play-learners into two or more segments/classes according to the "mix" of fundamental features available, including actions, attitudes, behaviors, needs, etc. Even though analysts can technically divide the play-learner groups into a lot of smaller segments, the approach is not practical from a marketing/advertising standpoint. The rule is to limit the number of segments (two or three) to better focus the advertising efforts.

Once the desired number of segments have been identified (most often with non-hierarchical clustering), data analysts can then make use of this information to predict future users' actions/behaviors based on existing classification and further confirm this prediction using *supervised learning*. This is known as the *data confirmatory* stage.

The (un)supervised learning techniques are available in both *machine learning* (Bishop, 2006) and *statistical learning* (James, Witten, Hastie, & Tibshirani, 2013), depending on one's field of research. The relationship between unsupervised and supervised learning, and their usage for exploratory/confirmatory data analysis, is depicted in Fig. 5.5. Initially, unsupervised learning is used to segment user-generated data into a number of clusters. Using information from these clusters, the play-learners are then differentiated according to their similarities as per certain performance metrics. Supervised machine/statistical learning may then be used to profile player clusters for predictive and prescriptive treatments. Data analysts may propose a

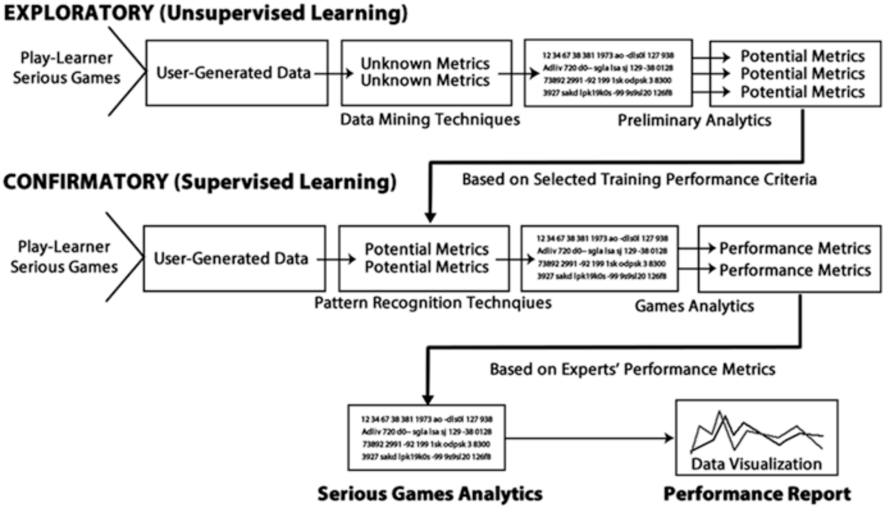


Fig. 5.5 Serious games analytics workflow: from unsupervised to supervised learning

predictive model of play-learners depending on the clustering (for example, experts vs. novices) using supervised learning techniques. Upon verification, the predictive model could be used to assess the performance of new play-learners via a "trained" algorithm or structure.

5.1.1 Clustering Techniques

Clustering technique is a good first-step analysis to be used to examine how various features in a data set (e.g., play-learners, game metrics, attitudes, actions) relate to unique groups. There are many potentially useful methods to analyze play-learners actions in serious games for insights. Some of the common methods that have been considered useful in game analytics are: Cluster Analysis, Archetypal Analysis, Non-negative Matrix Factorization, and principal component analysis (PCA). (An exhaustive account of unsupervised learning methods for game analytics is beyond the scope of this chapter. Readers are referred to Drachen, Thurau, Sifa, & Bauckhage (2013) for more illustration and examples.)

The main purpose of Cluster Analysis is to divide the play-learners into various clusters based on their similarities among one another. Membership of cluster is usually determined by how far (i.e., the distance) a certain unit is from the center of the cluster (or *cluster centroid*). If the centroids represent the "average" profiles (statistically speaking) for that particular cluster, then "archetypic" profiles are unique units that are found on the "edge" of the clusters. An archetype can be seen as a "pure" (or extreme) user before statistical averaging take place. In general, Cluster Analysis is useful for the crafting of general profiles that are representative of certain groups of play-learners; whereas Archetype Analysis is used for the identification of unique “power” play-learners.

PCA—closely related to Factor Analysis is a statistical method that is used in exploratory data analysis to reduce the dimensionality of data space. The results of PCA are usually discussed in terms of component/factor score and can be used to explain the loadings and weights of covariance in a multivariate data set. While PCA is commonly used in pattern recognition, it does not take into consideration the differences in class (or class separability). If class separability is important, a better alternative is linear discriminant analysis (LDA).

Supervised learning techniques are useful for predicting future data classes. If the label for the input data is discrete, the method is known as *classification*, but if the label is continuous, it is known as *regression*. Common supervised learning techniques, such as regression analysis, LDA, and decision trees, can be used to compose predictive models, classify new observations, and predict play-learners' behaviors that are centrally attributed to a category. Data analysts who are interested in just prediction may consider even more advanced techniques, such as Neural Networks (NN) and support vector machines (SVM), which allow for automation.

5.2 Cluster Analysis

At times when classification labels (e.g., experts/novices) are not available, Cluster Analysis can be a very useful unsupervised learning technique. First, performance metrics need to be identified. Once the metrics are identified, play-learners can be divided into two or more groups based on metrics of similarities (i.e., the clustering variables). Clustering variables may need to be normalized, and similarity measures calculated across the entire set of variables to allow for the grouping and comparison of play-learners. Similarity measures are fairly easy to comprehend, with larger values indicate greater dissimilarity, or distance, between persons.

Given that the exact process of assigning players to clusters depends on the selected clustering algorithm, cluster analysis is not an automatic process. Instead, it is an exploratory process that requires choosing and comparing algorithms, defining the number of clusters, etc. In fact, data analysts have over 100 available algorithms (e.g., Estivill-Castro, 2002) to help them decide on how many clusters to form. Given the large number of algorithms with each taking into consideration different sets of assumptions and parameters, there is really no correct way to cluster a data set because the same data set can yield different cluster solutions depending on how the procedures are determined. The best practice is, therefore, to try out different algorithms until a relatively better solution is identified.

Figure 5.6 shows an example of a *k*-means clustering, where two clusters are identified. Cluster 1 comprises experts, while Cluster 2 contained players who completed a game within a certain time frame. A closer investigation of Cluster 2 reveals the players with greater similarities to the expert cluster (overlapped area). Using only the cluster centroid, which is a mean profile of the cluster on each clustering

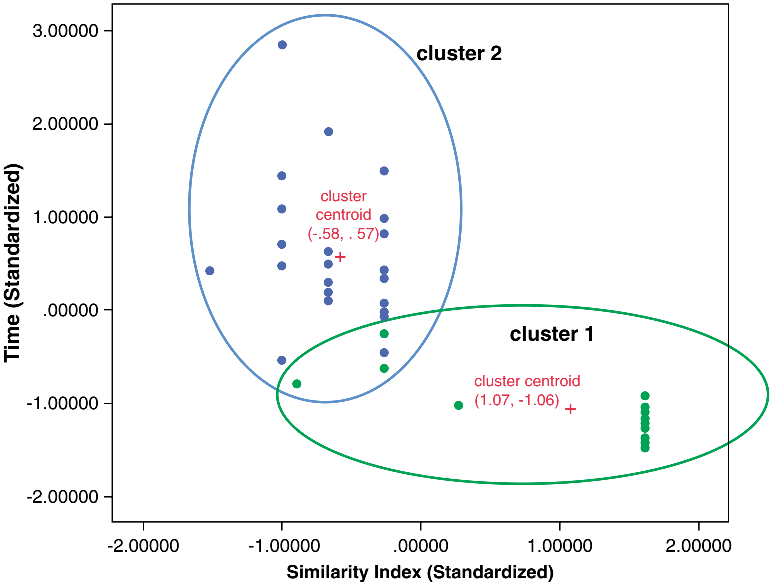


Fig. 5.6 Clustering of players using game metrics (time of completion and similarity index)

variable, researchers can draw conclusion about the profile of the clusters available. For example, Cluster 1 should be profiled as experts, while Cluster 2, novices. Cluster solutions that failed to reveal substantial variations indicates that further explorations are required to identify a better way to cluster the data. The cluster centroid should also be evaluated for correspondence with data analysts' prior expectations, which are often determined based on domain knowledge or practical experience.

5.3 *Linear Discriminant Analysis*

LDA is a better technique than PCA if class separability is an important consideration. It should be used when the data are labeled, or specifically, when play-learner group memberships are already established. The purpose of LDA is to identify the most helpful game metrics that could distinguish between these groups by way of a discriminant model. The usefulness of the model is dependent on its *classification accuracy*: i.e., the ability to predict known group membership correctly.

LDA works by formulating an unobserved variable called the *discriminant function score*, which is a linear function of the best combination of discriminating variables (in this case, game metrics). The discriminant function score can be used

Table 5.1 Classification table for classifying experts and novices using LDA

Actual group	Predicted experts	Predicted novices	Actual total	% Correct (%)
Experts	22	3	25	88
Novices	5	20	25	80
Predicted total	27	23	50	

to predict group memberships of future play-learners. The selection of game metrics in forming the discriminant function can be performed using:

1. Simultaneous procedure—all variables are entered together but only those with relatively higher loadings are interpreted, or
2. Stepwise procedure—the most parsimonious set of maximally discriminating variables are selected.

Once discriminant functions have been determined, data analysts can then assess the contribution of each game metric to a discriminant function by way of discriminant loadings. In addition, prediction accuracy using the developed discriminant function can be assessed with a holdout sample if data sizes are sufficiently large.

Alternatively, a "leave-one-out" cross-validation procedure, such as Jackknife reclassification, can be applied to LDA with smaller data sets. This method is carried out by sequentially holding out one case from the analysis and using the remaining cases to derive the discriminant functions used in classifying that case (Lachenbruch & Mickey, 1968). This process is repeated for all cases in the analysis to yield a prediction accuracy that is "less biased." With either a holdout sample or cross-validation, a classification table (confusion matrix) such as Table 5.1 can be obtained. The example given in Table 5.1 shows group hit ratios of 88 % (for experts) and 80 % (for novices), with an overall hit ratio of $(22+20)/50\times 100\,\% = 84\,\%$.

Hit ratios are usually compared to the proportional chance criterion (0.5 for this example). Based on the rule-of-thumb, hit ratio should exceed the chance criterion by 25 %; since this is true for this example, the accuracy of the predictive model formulated is established.

5.4 Item Response Theory

Item response theory (IRT) is another popular approach used for describing probabilistic relationships between responses on a set of test items and continuous latent traits (e.g., Lord & Novick, 1968; Mislevy, 1985). It is also widely used in educational testing and psychological measurement. In the serious games environment, IRT can be used with game designs involving a series of procedural tasks, where player behaviors represent one of two or more levels. Based on the theory, the probability to carry out a specific in-game action can be modeled using a nonlinear function of the task characteristics and players' latent traits (i.e., competencies).

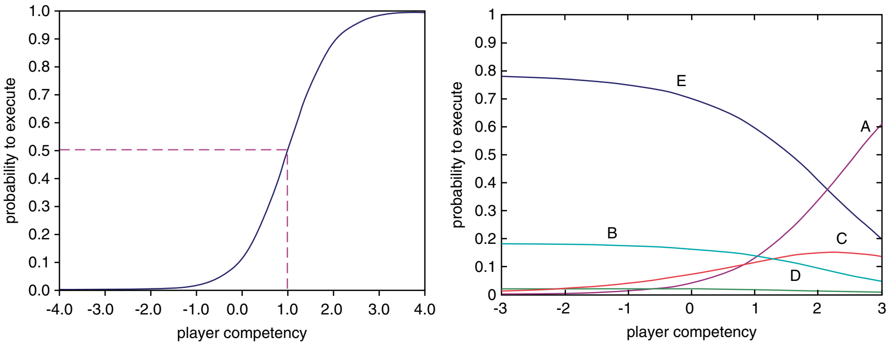


Fig. 5.7 Response function for tasks involving two (*top*) and five (*bottom*) actions

Figure 5.7 shows two different models where a certain task entails two (e.g., complete/not complete; left), or more actions (e.g., found item A, B, and C; right). Play-learners' latent competency levels can then be estimated, and used to differentiate between play-learners' groups (say, expert vs. novice). Alternatively, an IRT model can also be used to help in describing game task characteristics, such as the task's difficulty level, or the ability to distinguish between experts and novices; and to provide game designers with insights to modify or improve a specific game.

6 Conclusions

6.1 From Serious Games Analytics to Insights

Serious Games are more than just (educative) message broadcasters. They have the potential to become tools to raise performance and train decision-making skills in the play-learners. However, to make this happen, debriefing tools (Crookall, 2010) or assessment components will need to be built into the serious games to produce ad hoc/post hoc Serous Games Analytics.

The purpose of Bayesian Network is largely descriptive. It is highly suitable for understanding how play-learners make decisions and for testing and measurement assessment. However, it may not be suitable in serious games that involve a large number of variables, such as MMO serious games. It is highly dependent on the reliability of the expert-created DAG, and unable to handle the spatial–temporal variables found in today's serious games. Although it remains useful for edutainment and epistemic games and may support testing measurements, it may not yield meaningful and actionable insights for prescriptive training performance improvement.

Cluster analysis of players according to their actions and behaviors in (entertainment) games are descriptive in its ability to cluster/categorize players based on their gameplay preferences. The analytics obtained can then be used prescriptively in marketing and advertising to maximize the player data for monetization (Seif El-Nasr et al., 2013).

However, to advance Serious Games as an industry and Serious Games Analytics as a field, we proposed future research to focus on understanding expert performance (such as the Dreyfus model) for descriptive purposes: to identify the development stage of play-learners based on their actions and behaviors in the expert-novice continuum and prescriptively for (re)training and remediation by comparing how similar their actions and behaviors are to a preestablished expert performance baseline (see Loh & Sheng, 2014, 2015).

6.2 Expertise Index as Serious Games Analytics

It is possible to study experts' behavior in detail, given a certain scenario, and to deconstruct them into a series of components/actions using an instructional design strategy called *task analysis* (Jonassen, Hannum, & Tessmer, 1989). These components or action sequences can then be used to facilitate training, and be emulated by novices as they train to competency, to one day become experts—a process that could take up to 10 years. Findings in the area of expertise found experts to possess different reasoning patterns, decision-making procedures, and significantly better problem-solving strategies than novices (see Ericsson, Charness, Feltovich, & Hoffman, 2006). Even the belief systems of experts were different from the novices and have been shown to affect experts' and novices' actions accordingly (Karelaia & Hogarth, 2008).

The difference between experts' and novices' behaviors during problem-solving and decision-making is a very well-studied phenomenon in training and psychology literature (Dreyfus, 2004; Dreyfus & Dreyfus, 1980). In general, novices exhibit a tendency to follow rules *blindly* during problem-solving because they have yet to acquire the context in which those rules operate. As they gradually learn to apply the right rules with the right conditions, they are said to be growing in their *competency*. Competency is demonstrable and observable in a person's chosen *course of action* during problem-solving. Experts, who are so in tune with the tasks at-hand, are able to detect cues that are not obvious to non-experts. As a result, experts can appear (to untrained eyes) to be solving problem based on intuition while breaking or ignoring rules, at will.

The indicators of expert-novice behaviors vary widely and can range from time-to-task-completion rate, to mental representations of knowledge, to specific gaze patterns in scanning for information (Underwood, 2005). Evidences of expert-novice behavioral differences have been reported among airline pilots, teachers, surgeons, nurses, programmers, sportsmen (see for example: Hofer, 2011; Law, Atkins, Kirkpatrick, & Lomax, 2004; Williams & Ford, 2008), as well as digital game players (Boot, Kramer, Simons, Fabiani, & Gratton, 2008).

6.2.1 Competency and Observable Action Sequences

Because a person's competency can be characterized by an *observable* course of action taken during problem-solving, it should be possible to trace the *courses of action* (or *action sequences*) of experts and novices and compare the two sets of traces to determine how closely their actions match. By establishing the performance level of experts as the targeted level of achievement and comparing novice competency to that level, we are able to find the difference in performance between the two. The competency levels of individual novices can then be calculated as an *Expertise Index* and be used in the identification and ranking of play-learners by expertise (Loh & Sheng, 2015). In the case of multiple experts, the *Expertise Indices* may be evaluated using the *Maximum Similarity Index* (or MSI, see Loh & Sheng, 2014).

Once sufficient information on the play-learners' in-game actions and behaviors (i.e., what they actually do in the game) have been captured, data analysts can profile the players and player behaviors, using a categorization method to identify and profile player groups based on their characters and traits, such as playing styles and learning preferences. A detailed explanation regarding the *Expertise Index* for training scenarios with one or more experts is beyond the scope of this chapter and is already available elsewhere (Loh & Sheng, 2014, 2015).

7 Conclusions

In summary, we would like to reiterate that the purpose of Serious Games Analytics is to transform user-generated data traced in situ within the game habitat into *actionable insights.* The question to bear in mind is: what implementable strategies can we derive based on knowledge garnered from Serious Games Analytics in raising human performance and decision-making skills?

It is important that a serious games researcher understands how his/her research interests fit in with the business needs of the Serious Games industry (e.g., cost of production, return of investment, reporting). Some available "assessment methods" are truly testing and measurement assessment methods, and may or may not be suitable for training performance assessment. As such, there are very few serious games assessment frameworks to date.

To fully realized the potential of serious games, researchers will need to innovate and devise new *training performance* metrics and methods to: (1) better measure human performance with serious games (e.g., tracing of in-game actions, inference of cognitive process, categorization of psychological profiles), (2) improve metrics and methods for the measurement of skills, and cognitive abilities, (3) identify likely-expert performance through pattern recognition and focus on distilling trainable aspects, (4) score and distill in-game user-generated data to produce *actionable* insights, and (5) transform analytics into prescriptive, actionable *insights* for the improvement of human performance.

Acknowledgments Research projects from the Virtual Environment Lab (V-Lab) have been made possible in part through funding from the Defense University Research Instrumentation Program (DURIP) from the U.S. Army Research Office. The authors would like to thank Ms. Ariel Yining Loh for her help in editing the chapter.

References

Abt, C. C. (1987). *Serious games* (Reprint). Lanham, MD: University Press of America.

Aldrich, C. (2005). *Learning by doing: a comprehensive guide to simulations, computer games, and pedagogy in e-learning and other educational experiences*. San Francisco: Pfeiffer.

Anderson, E. (1957). A semigraphical method for the analysis of complex problems. *Proceedings of the National Academy of Sciences of the United States of America, 43*(10), 923–927. Retrieved December 12, 2014, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC528552/

Bauer, M. (2002). Using evidence-centered design to align formative and summative assessment. In *Proceedings of the Evidence-Centered Design Approach to Creating Diagnostic Assessments* (ITS2002) (pp. 87–96).

Bellotti, F., Kapralos, B., Lee, K., & Moreno-Ger, P. (2013). User assessment in serious games and technology-enhanced learning. *Advances in Human-Computer Interaction, 2013,* 2. doi:10.1155/2013/120791.

Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., & Berta, R. (2013). Assessment in and of serious games: An overview. *Advances in Human-Computer Interaction, 2013,* 11. doi:10.1155/2013/136864.

Ben Zur, H., & Breznitz, S. J. (1981). The effect of time pressure on risky choice behavior. *Acta Psychologica, 47*(2), 89–104. doi:10.1016/0001-6918(81)90001-9.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer Science+Business Media.

Boot, W. R., Kramer, A. F., Simons, D. J., Fabiani, M., & Gratton, G. (2008). The effects of video game playing on attention, memory, and executive control. *Acta Psychologica, 129*(3), 387–398. doi:10.1016/j.actpsy.2008.09.005.

Byun, J. H., & Loh, C. S. (2015). Audial engagement: Effects of game sound on learner engagement in digital game-based learning environments. *Computer in Human Behavior, 46*, 129–138. doi:10.1016/j.chb.2014.12.052.

Canossa, A., & Drachen, A. (2009). Patterns of play: Play-personas in user-centred game development. In *Proceedings of Breaking New Ground: Innovation in Games, Play, Practice and Theory Conference*. London: DiGRA.

Cobb, B. R., Rumí, R., & Salmerón, A. (2007). Bayesian network models with discrete and continuous variables. In P. Lucas, J. A. Gámez, & A. Salmerón (Eds.), *Advances in probabilistic graphical models* (Studies in fuzziness and soft computing, Vol. 214, pp. 81–102). Berlin, Germany: Springer. doi:10.1007/978-3-540-68996-6_4.

Crookall, D. (2010). Serious games, debriefing, and simulation/gaming as a discipline. *Simulation & Gaming, 41*(6), 898–920. doi:10.1177/1046878110390784.

Csikszentmihalyi, M. (1991). *Flow: The psychology of optimum experience*. New York: Harper Perennial.

DeSanctis, G. (1984). Computer graphics as decision aids: Directions for research. *Decision Sciences, 15*(4), 463–487. doi:10.1111/j.1540-5915.1984.tb01236.x.

Díez, F. J., Mira, J., Iturralde, E., & Zubillaga, S. (1997). DIAVAL, a Bayesian expert system for echocardiography. *Artificial Intelligence in Medicine, 10*(1), 59–73. doi:10.1016/S0933-3657(97)00384-9.

Dixit, P. N., & Youngblood, G. M. (2008). Understanding playtest data through visual data mining in interactive 3D environments. In *Proceedings of the 12th International Conference on Computer Games: AI, Animation, Mobile, Interactive Multimedia and Serious Games, Louisville* (CGAMES 2008). Wolverhampton, England: University of Wolverhampton. Retrieved December 12, 2014, from http://gameintelligencegroup.org/files/2009/09/DixitACMSS2008.pdf

Djaouti, D., Alvarez, J., & Jessel, J.-P. (2011). Classifying serious games: The G/P/S model. In P. Felicia (Ed.), *Handbook of research on improving learning and motivation through educational games* (pp. 118–136). Hershey, PA: IGI Global. doi:10.4018/978-1-60960-495-0.ch006.

Drachen, A., & Canossa, A. (2011). Evaluating motion: Spatial user behaviour in virtual environment. *International Journal of Arts and Technology, 4*(3), 294–314. doi:10.1504/IJART.2011.041483.

Drachen, A., Canossa, A., & Sørensen, J. R. M. (2013). Gameplay metrics in game user research: Examples from the trenches. In M. S. El-Nasr, A. Drachen, & A. Canossa (Eds.), *Game analytics: Maximizing the value of player data* (pp. 285–319). London: Springer. doi:10.1007/978-1-4471-4769-5_14.

Drachen, A., Thurau, C., Sifa, R., & Bauckhage, C. (2013). A comparison of methods for player clustering via behavioral telemetry. In *Proceedings of the 8th International Conference on the Foundations of Digital Games* (FDG 2013) (pp. 245–252). Crete, Greece: Society for the Advancement of the Science of Digital Games.

Dreyfus, S. E. (2004). The five-stage model of adult skill acquisition. *Bulletin of Science, Technology and Society, 24*(3), 177–181. doi:10.1177/0270467604264992.

Dreyfus, S. E., & Dreyfus, H. L. (1980). *A five-stage model of the mental activities involved in directed skill acquisition*. Berkeley, CA: University of California. Retrieved December 12, 2014, from http://www.dtic.mil/get-tr-doc/pdf?AD=ADA084551

Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (Eds.). (2006). *The Cambridge handbook of expertise and expert performance*. Cambridge Handbooks in Psychology. New York, NY: Cambridge University Press.

Ermi, L., & Mäyrä, F. (2007). Fundamental components of the gameplay experience: Analyzing immersion. In S. de Castell & J. Jenson (Eds.), *Worlds in play: International perspectives on digital games research* (pp. 37–53). New York: Peter Lang.

Estivill-Castro, V. (2002). Why so many clustering algorithms: A position paper. *ACM SIGKDD Explorations Newsletter, 4*(1), 65–75. doi:10.1145/568574.568575.

Fan, X., Miller, B., Park, K.-E., Winward, B. W., Christensen, M., Grotevant, H. D., et al. (2006). An exploratory study about inaccuracy and invalidity in adolescent self-report surveys. *Field Methods, 18*(3), 223–244. doi:10.1177/152822X06289161.

Folkestad, J. E., Robinson, D. H., McKernan, B., Martey, R. M., Rhodes, M. G., & Stormer-Galley. (2015). Analytics-driven design: Impact and implications of team member psychological perspectives on a serious games (SGs) design framework. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: Methodologies for performance measurement, assessment, and improvement*. New York: Springer.

Grimshaw, M., Lindley, C. A., & Nacke, L. (2008). Sound and immersion in the first-person shooter: Mixed measurement of the player's sonic experience. In *Proceedings of the Audio Mostly Conference*. Retrieved December 12, 2014, from http://wlv.openrepository.com/wlv/bitstream/2436/35995/2/Grimshaw_CGAMES07.pdf

Heckerman, D. (1995). *A tutorial on learning with Bayesian networks*. Redmond, WA: Microsoft. Retrieved December 12, 2014, from http://research.microsoft.com/pubs/69588/tr-95-06.pdf

Hofer, A. (2011). Exploratory comparison of expert and novice pair programmers. In Z. Huzar, R. Koci, B. Meyer, B. Walter, & J. Zendulka (Eds.), *Software Engineering Techniques* (Vol. 4980, pp. 218–231). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-22386-0_17.

Hoskin, R. (2012). *The dangers of self-report*. Retrieved December 12, 2014, from http://www.sciencebrainwaves.com/uncategorized/the-dangers-of-self-report/

IJsselsteijn, W., de Kort, Y., Poels, K., Jurgelionis, A., & Bellotti, F. (2007). Characterising and measuring user experiences in digital games. In *Proceedings of the International Conference on Advances in Computer Entertainment Technology* (pp. 27–30).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning — with applications in R*. New York, NY: Springer New York. doi:10.1007/978-1-4614-7138-7.

Jefford, M., Stockler, M. R., & Tattersall, M. H. N. (2003). Outcomes research: What is it and why does it matter? *Internal Medicine Journal, 33*(3), 110–118. doi:10.1046/j.1445-5994.2003.00302.x.

Jonassen, D. H., Hannum, W. H., & Tessmer, M. (1989). *Handbook of task analysis procedures*. Westport, CT: Praeger Publishers.

Joslin, S., Brown, R., & Drennan, P. (2007). The gameplay visualization manifesto. *Computers in Entertainment, 5*(3), 6. doi:10.1145/1316511.1316517.

Karelaia, N., & Hogarth, R. M. (2008). *Skill, luck, overconfidence, and risk taking*. Barcelona, Spain: Universitat Pompeu Fabra. doi:10.2139/ssrn.1374235.

Kim, J. H., Gunn, D. V., Schuh, E., Phillips, B., Pagulayan, R. J., & Wixon, D. (2008). Tracking Real-time User Experience (TRUE): A comprehensive instrumentation solution for complex systems. In *Proceeding of the 26th Annual CHI Conference on Human Factors in Computing Systems* (CHI'08) (p. 443). New York: ACM Press. doi:10.1145/1357054.1357126.

Kirkley, J., Kirkley, S., & Heneghan, J. (2007). Building bridges between serious game design and instructional design. In B. E. Shelton & D. A. Wiley (Eds.), *Educational design & use of computer simulation games* (pp. 59–81). Rotterdam, The Netherlands: Sense.

Kirriemuir, J., & McFarlane, A. (2003). Use of computer and video games in the classroom. In *Proceedings of the Level up Digital Games Research Conference*. Utrecht, The Netherlands: Universiteit Utrecht.

Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics, 10*(1), 1–11. doi:10.1080/00401706.1968.10490530.

Law, B., Atkins, M. S., Kirkpatrick, A. E., & Lomax, A. J. (2004). Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment. In *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications (ETRA '04)* (pp. 41–48). New York, NY: ACM Press. doi:10.1145/968363.968370.

Levin, A. (2010, August 30). Simulator training flaws tied to airline crashes. *USA Today*. Retrieved December 12, 2014, from http://usatoday30.usatoday.com/travel/flights/2010-08-31-1Acockpits31_ST_N.htm

Liu, M., Kang, J., Lee, J., Winzeler, E., & Liu, S. (2015). Examining through visualization what tools learners access as they play a serious game for middle school science. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: Methodologies for performance measurement, assessment, and improvement*. New York: Springer.

Loh, C. S. (2006). Designing online games assessment as "Information Trails.". In D. Gibson, C. Aldrich, & M. Prensky (Eds.), *Games and simulation in online learning: Research and development frameworks* (pp. 323–348). Hershey, PA: Idea Group. doi:10.4018/978-1-59904-941-0.ch032.

Loh, C. S. (2012a). Improving the impact and return of investment of game-based learning. *International Journal of Virtual and Personal Learning Environments, 4*(1), 1–15. doi:10.4018/jvple.2013010101.

Loh, C. S. (2012b). Information trails: In-process assessment of game-based learning. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 123–144). New York: Springer. doi:10.1007/978-1-4614-3546-4.

Loh, C. S., Anantachai, A., Byun, J. H., & Lenox, J. (2007). Assessing what players learned in serious games: *In situ* data collection, information trails, and quantitative analysis. In Q. Mehdi (Ed.), *Proceedings of the Computer Games: AI, Animation, Mobile, Educational & Serious Games Conference, Louiseville* (CGAMES 2007) (pp. 10–19). Wolverhampton, England: University of Wolverhampton.

Loh, C. S., & Byun, J. H. (2009). Modding Neverwinter Nights into serious game. In D. Gibson & Y. K. Baek (Eds.), *Digital simulations for improving education: Learning through artificial teaching environments* (pp. 408–426). Hershey, PA: IGI-Global.

Loh, C. S., & Sheng, Y. (2015). Measuring the (dis-)similarity between expert and novice behaviors as serious games analytics. *Education and Information Technologies*, 20(1), 5–19. doi:10.1007/s10639-013-9263-y

Loh, C. S., & Sheng, Y. (2013). Performance metrics for serious games: Will the (real) expert please step forward? In *Proceedings of the Computer Games: AI, Animation, Mobile, Educational & Serious Games Conference* (CGAMES 2013) (pp. 202–206). Louiseville. IEEE. doi:10.1109/CGames.2013.6632633.

Loh, C. S., & Sheng, Y. (2014). Maximum Similarity Index (MSI): A metric to differentiate the performance of novices vs. multiple-experts in serious games. *Computer in Human Behavior, 39*, 322–330. doi:10.1016/j.chb.2014.07.022.

Loh, C. S., Sheng, Y., & Ifenthaler, D. (2015). Serious games analytics: Theoretical framework. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: Methodologies for performance measurement, assessment, and improvement*. New York: Springer.

Loh, C. S., Sheng, Y., & Li, I.-H. (2015). Predicting expert-novice performance as Serious Games Analytics with objective-oriented and navigational action sequences. *Computers in Human Behavior. 49*:147–155. doi:10.1016/j.chb.2015.02.053.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Boston, MA: Addison-Wesley.

Medler, B., & Magerko, B. (2011). Analytics of play: Using information visualization and gameplay practices for visualizing video game data. *Parsons Journal for Information Mapping, 3*(1), 1–12.

Michael, D., & Chen, S. (2006). *Serious games: Games that educate, train, and inform*. Boston: Thomson Course Technology PTR.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80*(392), 993–997. doi:10.1080/01621459.1985.10478215.

Moura, D., Seif El-Nasr, M., & Shaw, C. D. (2011). Visualizing and understanding players' behavior in video games: Discovering patterns and supporting aggregation and comparison. In *Proceedings of the 2011 ACM SIGGRAPH Symposium on Video Games* (pp. 11–15). New York: ACM Press. doi:10.1145/2037692.2037695.

Nacke, L. E., Drachen, A., & Göbel, S. (2010). Methods for evaluating gameplay experience in a serious gaming context. *International Journal of Computer Science in Sport, 9*(2), 1–12.

Niedermayer, D. (1998). *An introduction to Bayesian networks and their contemporary applications*. Retrieved December 12, 2014, from http://www.niedermayer.ca/papers/bayesian/bayes.html

O'Rourke, E., Butler, E., Liu, Y. E., Ballweber, C., & Popovic, Z. (2013). The effects of age on player behavior in educational games. In G. N. Yannakakis & E. Aarseth (Eds.), *Proceedings of the 8th International Conference on the Foundations of Digital Games* (FDG 2013) (pp. 158–165). Chania, Greece: Society for the Advancement of the Science of Digital Games.

Oniśko, A., & Druzdzel, M. J. (2013). Impact of precision of Bayesian network parameters on accuracy of medical diagnostic systems. *Artificial Intelligence in Medicine, 57*(3), 197–206. doi:10.1016/j.artmed.2013.01.004.

Paulhus, D. L. (1991). Measurement and control of response biases. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitude* (pp. 17–59). San Diego, CA: Academic.

Pieters, R., & Warlop, L. (1999). Visual attention during brand choice: The impact of time pressure and task motivation. *International Journal of Research in Marketing, 16*(1), 1–16. doi:10.1016/S0167-8116(98)00022-6.

Pruett, C. (2010). Hot failure: Tuning gameplay with simple player metrics. *Game Developer Magazine*. Retrieved December 12, 2014, from http://gamasutra.com/view/feature/6155/hot_failure_tuning_gameplay_with_.php

Quellmalz, E., Timms, M., & Schneider, S. (2009). Assessment of student learning in science simulations and games. In processings of the *Workshop on learning science: Computer games, simulations, and education*. Washington, DC: National Academy of Sciences.

Robinson, R. W. (2007). Learning the structure of Bayesian networks. In T. D. Nielsen & F. V. Jensen (Eds.), *Bayesian networks and decision graphs* (2nd ed., pp. 229–264). New York: Springer.

Roese, N. J., & Jamieson, D. W. (1993). Twenty years of bogus pipeline research: A critical review and meta-analysis. *Psychological Bulletin, 114*, 363–375.

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *The Journal of Technology, Learning and Assessment, 8*(4), 3–41.

Sandford, R., & Williamson, B. (2005). *Games and learning: A handbook from Futurelab*. Bristol, England: Futurelab.

Sawyer, B., & Rejeski, D. (2002). *Serious games: Improving public policy through game-based learning and simulation*. Washington, DC: Woodrow Wilson International Center for Scholars. Retrieved December 12, 2014, from http://www.seriousgames.org/images/seriousarticle.pdf

Scarlatos, L. L., & Scarlatos, T. (2010). Visualizations for the assessment of learning in computer games. In *Proceedings of the 7th International Conference & Expo on Emerging Technologies for a Smarter World, Incheon, S. Korea* (CEWIT 2010). Retrieved December 12, 2014, from http://ms.cc.sunysb.edu/~lscarlatos/pubs/LLS2010_CEWIT.pdf

Seif El-Nasr, M., Drachen, A., & Canossa, A. (Eds.). (2013). *Game analytics: Maximizing the value of player data*. London: Springer.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–524). Charlotte, NC: Information Age.

Shute, V. J., Masduki, I., Donmez, O., Dennen, V. P., Kim, Y. J., Jeong, A. C., et al. (2010). Modeling, assessing, and supporting key competencies within game environments. In D. Ifenthaler, P. Pirnay-Dummer, & N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge (Part 4)* (pp. 281–309). Boston: Springer. doi:10.1007/978-1-4419-5662-0_15.

Shute, V. J., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. Cambridge, England: MIT Press.

Shute, V. J., Ventura, M., Bauer, M., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning. In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). New York: Routledge.

Smith, S. P., & Du'Mont, S. (2009). Measuring the effect of gaming experience on virtual environment navigation tasks. In K. Kiyokawa, S. Coquillart, & R. Balakrishnan (Eds.), *Proceedings of the IEEE Symposium on 3D User Interfaces* (3DUI 2009) (pp. 3–10). Louisiana: IEEE. doi: 10.1109/3DUI.2009.4811198.

Thawonmas, R., Ho, J.Y., & Matsumoto, Y. (2003). Identification of player types in massively multiplayer online games. In *Proceedings of the 34th annual conference of international simulation and gaming association, Chiba, Japan* (ISAGA 2003) (pp. 893–900).

Thawonmas, R., & Iizuka, K. (2008). Visualization of online-game players based on their action behaviors. *International Journal of Computer Games Technology, 2008*, 1–9. doi:10.1155/2008/906931.

Torrente, J., Borro-Escribano, B., Freire, M., del Blanco, A., Marchiori, E. J., Martinez-Ortiz, I., et al. (2014). Development of game-like simulations for procedural knowledge in healthcare education. *IEEE Transactions on Learning Technologies, 7*(1), 69–82. doi:10.1109/TLT.2013.35.

Underwood, J. (2005). Novice and expert performance with a dynamic control task: Scanpaths during a computer game. In G. Underwood (Ed.), *Cognitive processes in eye guidance* (pp. 303–323). Oxford, England: Oxford University Press. doi:10.1093/acprof:oso/9780198566816. 003.0013.

Van Eck, R. (2006). Digital game-based learning: It's not just the digital natives who are restless. *EDUCAUSE Review, 41*(2), 16–30.

Wallner, G. (2013). Play-graph: A methodology and visualization approach for the analysis of gameplay data. In G. N. Yannakakis & E. Aarseth (Eds.), *Proceedings of the 8th International Conference on the Foundations of Digital Games* (FDG 2013) (pp. 253–260). Crete, Greece: Society for the Advancement of the Science of Digital Games.

Wallner, G., & Kriglstein, S. (2012). A spatiotemporal visualization approach for the analysis of gameplay data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI'12) (pp. 1115–1124). New York: ACM Press. doi:10.1145/2207676.2208558

Wallner, G., & Kriglstein, S. (2013). Visualization-based analysis of gameplay data—A review of literature. *Entertainment Computing, 4*(3), 143–155. doi:10.1016/j.entcom.2013.02.002.

Wickens, C. D., Stokes, A., Barnett, B., & Hyman, F. (1993). The effects of stress on pilot judgment in a MIDIS simulator. In O. Svenson & A. J. Maule (Eds.), *Time pressure and stress in human judgment and decision making* (pp. 271–292). Boston: Springer. doi:10.1007/978-1-4757-6846-6_18.

Williams, A. M., & Ford, P. R. (2008). Expertise and expert performance in sport. *International Review of Sport and Exercise Psychology, 1*(1), 4–18. doi:10.1080/17509840701836867.

Young, M. E., Sutherland, S. C., & Cole, J. J. (2011). Individual differences in causal judgment under time pressure: Sex and prior video game experience as predictors. *International Journal of Comparative Psychology, 24*(1), 76–98.

Chapter 6
Cluster Evaluation, Description, and Interpretation for Serious Games

Player Profiling in Minecraft

David J. Cornforth and Marc T.P. Adam

Abstract This chapter describes cluster evaluation, description, and interpretation for evaluating player profiles based on log files available from a game server. Calculated variables were extracted from these logs in order to characterize players. Using circular statistics, we show how measures can be extracted that enable players to be characterized by the mean and standard deviation of the time that they interacted with the server. Feature selection was accomplished using a correlation study of variables extracted from the log data. This process favored a small number of the features, as judged by the results of clustering. The techniques are demonstrated based on a log file data set of the popular online game Minecraft. Automated clustering was able to suggest groups that Minecraft players fall into. Cluster evaluation, description, and interpretation techniques were applied to provide further insight into distinct behavioral characteristics, leading to a determination of the quality of clusters, using the Silhouette Width measure. We conclude by discussing how the techniques presented in this chapter can be applied in different areas of serious games analytics.

Keywords Cluster evaluation • Cluster description • Cluster interpretation • Player profiles • Cognitive performance

D.J. Cornforth (✉)
University of Newcastle, Australia, University Drive, ICT3.12, Callaghan,
NSW 2308, Australia
e-mail: david.cornforth@newcastle.edu.au

M.T.P. Adam
University of Newcastle, Australia, University Drive, ICT3.62, Callaghan,
NSW 2308, Australia
e-mail: marc.adam@newcastle.edu.au

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based
Learning, DOI 10.1007/978-3-319-05834-4_6

1 Introduction

Evaluating player profiles in a serious game is an important consideration when evaluating the success of the game (Asteriadis, Karpouzis, Shaker, & Yannakakis, 2012; Loh & Sheng, 2013). Not all players play a game with the same abilities, expectations, or approach (e.g., Astor, Adam, Jerčić, Schaaff, & Weinhardt, 2014; Feldmann, Adam, & Bauer, 2014; Loh & Sheng, 2014). Understanding player preferences may help inform the provision of game infrastructure. For example, the time at which players prefer to play could be used to assist in planning of bandwidth and server requirements in an online game. Player profiling is an analytical approach that uses empirical data collected from a game and can be used to help identify groups of players. The data in many cases already exist on the server; indeed the amount of information available in server logs may represent a huge, untapped resource. These logs should be mined using machine learning techniques, to yield valuable information on player types and on their behavior. Such machine learning tools are widely used and mature in their development, and there is a huge amount of free code online that can be used to implement systems to sift through the logs and build player profiles. These profiles can then inform the results gained from serious games that are being used to study cognitive performance, for example. Player groups may be based on fundamental differences between approach and performance. Some players may adopt an aggressive posture in game play, while others may employ more subtle tactics to achieve their goals. As well as assessing the effectiveness of a game, this approach can be used to help tailor the challenges within the game so they better meet the serious intention of the game.

In this chapter, we describe several useful approaches that can be used to characterize player behavior in games using cluster analysis. We discuss the various parts of the clustering process, including how to collect data, how to clean the data, how to select data features for the analysis, some basic clustering approaches, and how to assess the quality of the outcomes. Moreover, we outline the particular challenges inherent in applying this type of analysis to data derived from serious games. This will include a discussion of the quantitative measures that can be used to assess the quality of the clusters and how this also informs the selection of parameters such as the data features to be used in the process. In particular, we describe the Silhouette Width, a quantitative measure that may be used to obtain a measure of cluster quality. Moreover, we discuss methods for visualizing player categories derived from the cluster analysis and other aspects of the data. To better illustrate the concepts of cluster analysis as applied to player profiling, we collect information from real-time sessions of the online game *Minecraft* (minecraft.net). Minecraft is a very popular online game that is frequently used as a serious game for educational purposes (e.g., Danish Geodata, 2014; Duncan, 2011; Ekaputra, Lim, & Eng, 2013; Short, 2012).

The remainder of this chapter is organized as follows. In Sect. 2 we outline the theoretical background for cluster analysis. Section 3 provides background information on the online game Minecraft. In Sect. 4 we conduct cluster analysis based on a Minecraft data set.

2 Theoretical Background

Cluster analysis is a tool derived from computational intelligence and is a well-known technique in the machine learning community. Cluster analysis uses tools that can automatically search a set of data for naturally occurring groups or clusters. Clustering has been used with success in market research, for example, to identify groups of customers who can be provided with tailored advertising of products and service (Lee, Kim, & Kim, 2008). Indeed variations of the technique have been used across a broad section of application domains, including astronomy (Jang & Hendry, 2007), agriculture (Ruß & Kruse, 2011), and geology (Honarkhah & Caers, 2010). Although the computation steps of clustering are automatic, it must also be seen within the context of a process that begins with raw data and ends with new domain knowledge. These steps typically include (1) data cleaning and transformation, (2) feature selection, (3) clustering, (4) cluster evaluation, (5) cluster description, and (6) cluster interpretation (Nesbitt & Cornforth, 2013).

Data cleaning acknowledges the fact that data may contain errors and inconsistencies (Han, Kamber, & Pei, 2011). With the data set used in this work, this is unlikely to be an issue since the data were collected directly from server logs and were not compiled from multiple sources.

Data transformation arises from the fact that the features derived from the raw data sets may not be suitable for the clustering process. As many clustering algorithms depend on numeric data only, and conceptualize a data record as a point in d-dimensional space, the raw data must be transformed into ratio-scaled numeric variables. Every step of the data cleaning and transformation requires some domain knowledge in order to ensure that decisions are taken that are consistent with the way the data set was collected and with the type of events being described. Thus, a cluster analysis must pay particular attention to the application area being studied and the nature of the data available. Some features may be numeric but may not be suitable to use for the calculation of summary statistics. An example that is particularly relevant for games is the time of day feature used in this study. This feature must be treated as a circular variable instead of a linear variable in order to obtain meaningful and consistent summary statistics.

Feature selection refers to the process of choosing a subset of features from those available in order to find the best partition of the data into clusters (Han et al., 2011). After the data have been clustered using the machine learning algorithms, the quality of the clusters may be assessed in the cluster evaluation stage. If the clusters are not well separated, they are unlikely to yield any definitive knowledge about the natural groups of players. This may entail going back to the feature selection stage, and selecting a different subset of features in order to improve the quality of clusters. Once the data set has been partitioned by clustering, a cluster description stage attempts to describe each cluster, and cluster interpretation extracts knowledge from these results in order to describe the groups of players found in the data.

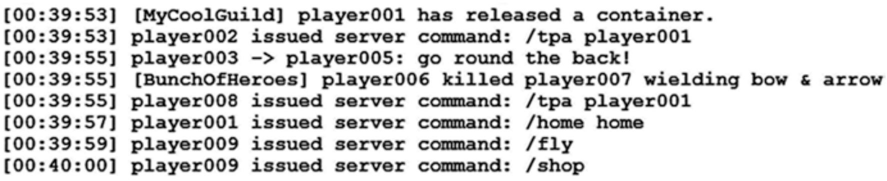
Clustering can inform the practitioner of the structure and type of phenomena being studied, as it automatically searches a database for groups or clusters that occur naturally. However, for any given data set there are a huge number of possible partitions of the data set, and clustering algorithms are not guaranteed to find the "best" partition, however that might be measured. Quantitative measures can be used to assess the quality of the clusters formed by this process, and so can guide the selection of parameters, including data features or variables, to be used in the analysis. If the clustering process does not yield good quality clusters according to the quantitative measure used, then the cluster process many be repeated using a different set of features.

In this chapter, we apply clustering to the area of serious games and combine it with analysis techniques. We describe a quantitative measure of cluster quality and discuss how this might be used to choose clusters that provide insight into the data set.

3 Minecraft Data

Minecraft is a type of *sandbox* game, characterized by a wide variety of gameplay but a lack of definite goals (Duncan, 2011). In this game, players build structures, both underground and above ground, mine minerals and convert them to manufactured products. Players may concentrate on survival, competition, constructing environments, building a personal economy through the collection of valuable items, technology growth though simulated manufacturing processes, or even on exploring electrical circuits and computational logic by using built-in components. This provides a great potential for educational uses of the game (Ekaputra et al., 2013). The appearance of the game is characterized by blocks of material with shading and texture, so that the resolution of objects is relatively low, compared to a game where maximum realism is sought. Minecraft has featured in a number of reports where a range of serious applications have been suggested, for example educational (Waxman, 2012) and resource building. For example, the entire country of Denmark was modeled in Minecraft, including buildings (Danish Geodata, 2014).

The data used in this study consists of log entries from actual play sessions on a server supporting online gameplay. Each log entry contains the time, information about the type of event that led to the entry, the players involved, and free text fields containing messages or commands. An example is shown in Fig. 6.1. Users can be



```
[00:39:53] [MyCoolGuild] player001 has released a container.
[00:39:53] player002 issued server command: /tpa player001
[00:39:55] player003 -> player005: go round the back!
[00:39:55] [BunchOfHeroes] player006 killed player007 wielding bow & arrow
[00:39:55] player008 issued server command: /tpa player001
[00:39:57] player001 issued server command: /home home
[00:39:59] player009 issued server command: /fly
[00:40:00] player009 issued server command: /shop
```

Fig. 6.1 Example of server logs from Minecraft

identified by their unique name. For example, Fig. 6.1 shows the user known as *Player003* sending a message to *Player005* instructing the latter in a tactical encounter. Another player called *Player006* kills *Player007*, while *Player009* is seen to issue several server commands. The data set consists of 127,765 such log entries covering a variety of encounters between players.

4 Analysis

Clustering is well known in the machine learning community, but must be seen within the context of a process which begins with raw data and ends with new knowledge of the particular application area being studied (Han et al., 2011). In this work, the task is to look for natural groups or cluster of players in the Minecraft server log data. In order to find clusters, it is necessary to extract some features which can be used to describe the characteristics of a player. Features can be based on analysis of the data available and can take a variety of forms. Some simple features, that are easily understandable and have an obvious explanation, will be extracted from the data logs of an online game server. There are a variety of clustering algorithms, but the simplest of these require a distance measure to categorize players. The Euclidean distance measure is the most commonly used, and measures distance between two players, where each player is imagined as a point in a d-dimensional space. The ordinates of such a point are the features used (Lloyd, 1982).

4.1 Data Transformation

The data set examined in this study comprised 24 h of data and approximately 128,000 entries in a server log of an online version of Minecraft. Custom software was prepared to scan the Minecraft server logs and extract the required information. Each log entry was assigned to one of the following types:

- Craft Scheduler Thread
- Server thread/ERROR
- Server thread/INFO
- Server command
- Message (sent from one player to another)
- Player killed due to events in the game (not killed by another player)
- One player killed another
- Server thread/WARN
- User authenticator

The time of the log entry requires special treatment. Time of day is a circular quantity and so aggregations such as mean and standard deviation cannot be obtained in the usual way. As an illustration, consider finding the average value of three times: 22:30, 23:00, and 23:30. By adding together these values (taking account of the minutes and seconds) and then dividing the result by 3, one would obtain the correct mean of 23:00. However, if these times were all moved forward by 1 h, the values would be 23:30, 24:00, and 00:30. Applying the same process to these times would result in a mean of 16:00 which is the wrong answer, while the answer one would expect is 24:00. Because time of day is a circular quantity, it must be analyzed using circular statistics (Mardia, 1975). Here, the time of day is viewed as a point on a circle, and extracted into its conceptual orthogonal components using trigonometry functions:

$$cosTime = \cos\left(2\pi \cdot \frac{TimeVal}{24}\right) \tag{6.1}$$

$$sinTime = \sin\left(2\pi \cdot \frac{TimeVal}{24}\right) \tag{6.2}$$

where *TimeVal* is the 24-h format time converted into a single number where, for example, 3:15 p.m. would be coded as 15.25.

Using these new variables, meaningful summary statistics may be obtained, based on the *von Mises* distribution (Mardia, 1975). Using the mean values of these variables, a reliable value of the mean day of week can be obtained:

$$\mu = 24 \cdot atan2\left(\overline{cosTime}, \overline{sinTime}\right) \tag{6.3}$$

where $\bar{x}$ indicates the mean of x. Using the example mentioned above, the times of 23:30, 24:00, and 00:30 would be coded as 23.5, 24, and 0.5, respectively. Using Eqs. (6.1) and (6.2) above would result in values of 0.99, 1, and 0.99 for *cosTime*, and in −0.13, 0, and 0.13 for *sinTime*. The average *cosTime* is 0.99 and the average *sinTime* is 0. Using the inverse tangent function of Eq. (6.3): *atan2*(0.99, 0) provides the correct answer of 0, corresponding to midnight.

The custom software prepared by the authors extracted information for 941 players. The data set was converted to the ARFF format (Attribute Relation File Format) (ARFF, 2014) for use with the Weka machine learning package (Witten & Frank, 2005). By assigning log entries into the cases listed above, a number of features were obtained. All features used are listed in Table 6.1. The column labeled “Score” refers to a correlation study explained in the next section.

The number of times each player was featured in any capacity in the log entry was calculated, as *count* in Table 6.1. The fractional values *fracCmd*, *fracInfo*, *fracSent*, and *fracRecv* were calculated as the number of times each player gave a command, featured in an information text, sent or received a message, divided by *count*. Player kills were recorded separately depending on whether the player was killed by another player (*victim*) or whether the player was killed by another feature of the game (*killed*). This can happen when the player walks off a cliff, is drowned, is

Table 6.1 List of features derived from player logs, with an explanation of how the feature was derived from log data and their relative score expressing correlation with other features

Short name	Explanation	Score
count	Number of times player mentioned in log file	1.09
killed	Number of times player killed by game event	0.01
kills	Number of times player killed another player	0.13
victim	Number of times player killed by another player	0.35
fracCmd	Fraction of log entries involving a command	0.17
fracInfo	Fraction of log entries involving an information	0.50
fracSent	Fraction of log entries involving a message sent by player	0.18
fracRecv	Fraction of log entries involving a message received by player	0.47
meanCmndLen	Average length of commands entered by this player	0.44
meanInfoLen	Average length of information text strings involving this player	0.53
meanSentLen	Average length of messages sent by this player	1.28
meanRecvLen	Average length of messages received by this player	1.41
meanTheta	Mean time of logs for this player (using circular statistics)	0.01
meanBigR	Standard deviation of time of logs for this player	0.56

encased by sand, is consumed by fire, dies from hunger or poisoning, or when the player is killed by a Non-Player Character (NPC), known as a “mob,” which includes zombies, wolves, spiders, creepers, skeletons, endermen, silverfish, and a dragon. These lists provide an idea of the richness of the modeling environment available in Minecraft. In addition to counting various log entries, for all types that include a free text field, the average length of the text field is calculated for each player.

4.2 Feature Selection

The choice of features to include in the clustering step can have a very significant effect upon the outcome. Some features may be discarded as they are closely correlated with others, and so contribute little information, or the practitioner may create new features using mathematical transformations based on existing ones. It is essential that features are selected according to rigorous and repeatable criteria. This is recognized as a serious issue to the extent that there is a body of literature devoted to the topic; for example, Mitra, Murthy, and Pal (2002) provide a useful summary of some approaches.

Feature selection can be automated and a variety of methods exist for this. These are usually divided into ranking methods and wrapper methods (Sun, Todorovic, & Goodison, 2010). Ranking methods assign some score to each feature, so that a subset of higher scoring features can be selected. Wrapper methods evaluate the performance of the particular data mining methods with different combinations of features chosen. Feature selection then becomes a search through the possible combinations of features. Both approaches can be treated as an optimization problem, and methods ranging from gradient descent to genetic algorithms have been applied (for example, Huang & Wang, 2006; Inza, Larranaga, Etxeberria, & Sierra, 2000; Perkins, Lacker, & Theiler, 2003; Wang, Yang, Teng, Xia, & Jensen, 2007). However, it should be noted that just because a particular subset of features provides a high score, it does not follow that the resulting partition of the data set will provide new knowledge for the practitioner: the clusters may be well separated but it does not mean they represent useful knowledge.

In this work a correlation study, using Pearson correlation, provided the *r* value for each pair of features in the data. This allowed the 14 features to be ordered by calculating a composite correlation score for each feature. A list of features is provided in Table 6.1. The first column gives a short name for the feature. The second column explains what this feature represents. The third column gives the correlation score. Smaller values of this score indicate features that are relatively uncorrelated with the others. Features were chosen in ascending order of this correlation score. The intention was to ensure that features that were uncorrelated with others were included in the clustering process.

Although the procedure mentioned above provides a list of ranked features, there is no simple answer about how many features to include in the clustering, but one approach is discussed in the section on cluster evaluation below.

4.3 Clustering

When applied to data of players in Minecraft, clustering has the potential to identify types of players, or attributes that characterize certain types of players. A variety of algorithms exist to form a partition of a data set, and many of these are accessible via freely available software. Perhaps the most well known is *k-means* clustering (Lloyd, 1982). This is an example of a centroid-based method, and relies on some measure of distance, usually Euclidean. If the number of clusters *n* is specified in advance, it begins by randomly choosing *n* records as the centroids of *n* clusters. Every record in the data set is assigned to its nearest centroid, and therefore to the cluster its centroid represents. In the next round, each centroid is moved to the mean of all the records belonging to that cluster. Again, each record is assigned to its nearest centroid, and therefore to the corresponding cluster. The process repeats until some error measure, usually based on the squared distance of all points to centroids, has fallen below some threshold. Each record has now been assigned to a cluster.

A variation on *k-means* estimates not only a mean for each cluster, but models each cluster as a Gaussian kernel, estimating variance as well. Now each data record has membership of each cluster, but a weighted membership defined by its a posteriori probability according to the kernel function. This is known as the Expectation Maximization (EM) algorithm. In the Expectation step, the membership probabilities of each record in each cluster are recalculated given the existing cluster centers. In the Maximization step, cluster centers and covariance matrices are recalculated, so that the position of the cluster centers is changed. This is an iterative procedure but has a known guaranteed convergence (Witten & Frank, 2005).

This work used the data mining toolkit known as Weka, as it is free, open source, and widely accepted in the machine learning community (Witten & Frank, 2005). It is thus easily accessible for serious games analytics. The clustering was performed using EM for every combination of features from 2 to the full 14, in ascending order of the *Score* column shown in Table 6.1. For each combination of features, the number of clusters ranged from 2 to 10 clusters. In all, 130 clustering operations were performed using the Weka software.

4.4 Cluster Evaluation

There are many quantitative measures available for assessing the quality of clusters (Ackerman & Ben-David, 2008; Halkidi, Batistakis, & Vazirgiannis, 2001; Hubert, 1985; Jain, 2010; Strehl, Ghosh, & Mooney, 2000). Such measures allow the practitioner to assess the validity of a data partition produced by automated clustering. However, this step is often neglected in clustering analysis (Bolshakova & Azuaje, 2003). The quality of the results of the clustering process would be expected to reflect the choices made of selected features, clustering algorithm, and other parameters. This can be facilitated by the use of quantitative measures of cluster quality, and there is literature devoted to this effort (Handl, Knowles, & Kell, 2005). Such measures can, for example, favor a relatively smaller distance between data points within clusters, and favor a relatively larger distance between data points in different clusters. The concept of distance requires that any two records can be compared, and the distance between them can be a measurable quantity.

The Silhouette Width was introduced by Rousseeuw (1987) and has subsequently been used to validate clusters found, for example for genome expression data (Bolshakova & Azuaje, 2003). The Silhouette Width has been used for clustering of players in serious games (Asteriadis et al., 2012) but in that study was used for selection of number of clusters only and not for selection of the number of features used. The Silhouette Width (SW) has been shown to have desirable properties (Breaban & Luchian, 2011), and can be calculated for each record *i*:

$$\mathrm{SW}(i) = \frac{separation(i) - cohesion(i)}{\max\{cohesion(i), separation(i)\}} \tag{6.4}$$

In Eq. (6.4), *cohesion*(*i*) is a measure of the average distance between record *i* and the other records in the same cluster C_i, where d_{ij} is the distance between record *i* and another record *j*, and n_i is the number of records in cluster *i*:

$$cohesion(i) = \frac{1}{n_i} \sum_{j \in C_i} d_{ij} \tag{6.5}$$

In Eq. (6.4), *separation*(i) is a measure of the average distance between record i and all records in the nearest cluster C_l. Here, the *min*() denotes that the average distance is computed between record i and each cluster that i does not belong to. The minimum of these is taken, indicating the distance between i and the closest other cluster:

$$separation(i) = \min\left(\frac{1}{n_l}\sum_{j\in C_l} d_{ij}\right) \tag{6.6}$$

Distance d_{ij} is calculated as the Euclidean distance between instance i and instance j. If record x_i has m features, the difference must be calculated for each feature, then squared and summed:

$$d_{ij} = \sqrt{\frac{1}{N}\sum_{k=1}^{m}\left(x_{ik} - x_{jk}\right)^2} \tag{6.7}$$

A measure of the overall cluster separation is then obtained from the average Silhouette Width ($\overline{\mathrm{SW}}$) over all N records in the data set:

$$\overline{\mathrm{SW}} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{SW}(i) \tag{6.8}$$

The value of $\overline{\mathrm{SW}}$ ranges from −1 (completely meshed clusters) to +1 (well-separated clusters).

Extra code was prepared by the authors and integrated with the Weka package (Witten & Frank, 2005) in order to calculate the Silhouette Width measures, and to perform post-processing and collation of results. The number of features selected was varied between 2 and 14, in the order of the Score given in Table 6.1. So, for example, the first attempt at clustering used the features *killed* and *meanTheta*, as these have the lowest correlation with other features. As the EM algorithm allows the number of clusters to be specified beforehand, this number was varied between 2 and 10 clusters. This provides 130 combinations of parameters resulting in 130 possible partitions of the data set.

Figure 6.2 shows the results of these 130 runs of the clustering software, with the $\overline{\mathrm{SW}}$ plotted against the number of features used, from 2 to 14. Every point on the graph is a partition of the dataset made by running the EM algorithm for a different set of features (from 2 to 14), and the specifying a desired number of clusters (from 2 to 10). At the left side of the graph, it can be observed that the $\overline{\mathrm{SW}}$ is highest for partitions using only two features, achieving a maximum of 0.96, for two features and eight clusters. This is an extremely high value for the $\overline{\mathrm{SW}}$ and approaches the maximum value of 1, indicating well-separated clusters. As the number of features is increased, the $\overline{\mathrm{SW}}$ reduces, indicating that clusters are not so well separated in higher dimensional spaces. The exception seems to be partitions using 11 features

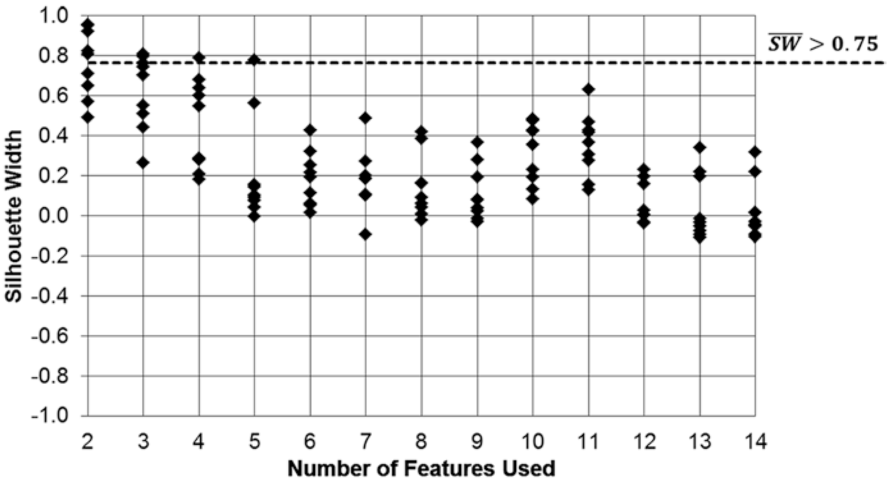


Fig. 6.2 Average Silhouette Width against number of features used for all partitions

that have slightly higher scores. This figure shows an ideal number of features between 2 and 5, suggesting that the features *killed*, *meanTheta*, *kills*, *fracCmd*, and *fracSent* are useful in separating players into different groups.

One might ask whether there is any advantage to selecting a subset of features instead of using all the features available in the dataset, and what the measures of cluster quality can tell us about this as a reasonable choice. According to this graph, choosing all the features would be a bad choice, as the $\overline{\text{SW}}$ reduces as the number of features chosen increases. This is important, as it clearly shows that use of all features would result in a poor division of the data that would be less likely to shed any light on the nature of these events. The importance of feature selection is well illustrated by this graph.

Figure 6.3 plots the same 130 partitions, but here the $\overline{\text{SW}}$ is plotted against the number of clusters used, from 2 to 10. From this figure, the higher values can be observed for between 7 and 10 clusters, although all selections of the number of clusters results in high scores (0.75 or more).

In order to find partitions that may lead to insights into the data set, the choices must be narrowed down, using the value of $\overline{\text{SW}}$ for guidance. It is desirable to select a partition with a high value for the $\overline{\text{SW}}$. A list of all partitions with $\overline{\text{SW}} > 0.75$ is given in Table 6.2, in order of decreasing $\overline{\text{SW}}$.

As this work is concerned with choosing the number of features to use, only one partition needs to be selected for each set of features. Some of the differences in values for $\overline{\text{SW}}$ are very small, and therefore it makes sense to select only a representative sample. Also, a large number of clusters are not desirable as it is more difficult to interpret and obtain useful conclusions from such a partition. Bearing in mind these considerations, two partitions were selected for further study. These are the partition for two features and eight clusters (Partition I), with the $\overline{\text{SW}}$ of 0.96,

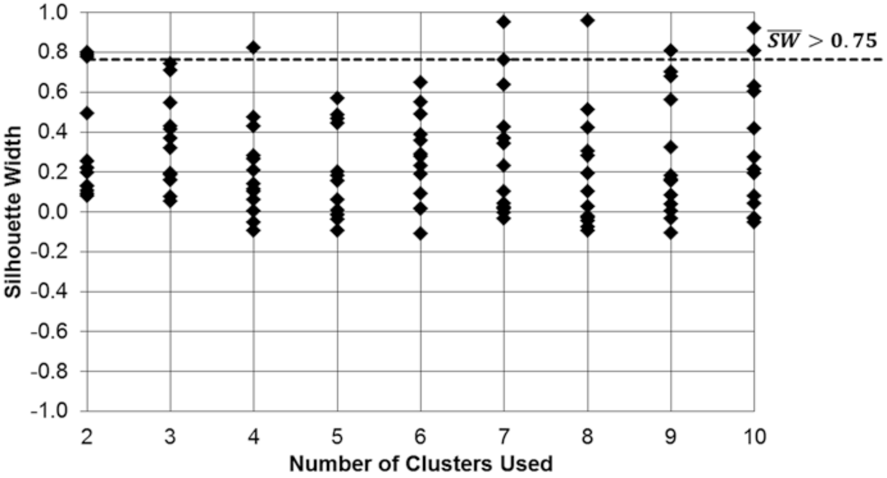


Fig. 6.3 Average Silhouette Width against number of clusters used for all partitions

Table 6.2 The partitions that achieved the highest score according to the $\overline{\text{SW}}$

Features	Clusters	$\overline{\text{SW}}$
2	8	0.96
2	7	0.95
2	10	0.92
2	4	0.82
2	9	0.81
3	10	0.81
3	2	0.80
4	2	0.79
5	2	0.78
3	7	0.76

and the partition for five features and two clusters (Partition II), with the $\overline{\text{SW}}$ of 0.78. These will be examined in more detail below.

In addition to these partitions, a range of partitions using only two features was also examined using 2-dimensional scatterplots. These will be used to illustrate the effectiveness of data transformation and simple visualization.

4.5 Cluster Description

The partitions selected using the cluster quality measures in the previous step were described using the statistical properties of each cluster. For each feature selected, the mean (cluster centroid) and 95 % confidence intervals were calculated using the student's *t* distribution. Any feature having a mean significantly different from the means of the same feature in all other clusters was deemed to be useful in describing that cluster. In this way, each cluster was described in terms of its significant attributes (where feature values are significantly different from all other clusters). For example, a cluster could be described as having an unusually high number of disturbances in one particular region, when compared with other clusters.

4.6 Cluster Interpretation

Clusters formed by the machine learning algorithms unfortunately are not accompanied by descriptions of why events were grouped in that way, and so it is up to the domain expert to draw conclusions or to identify phenomena that are of interest. Some groups are obvious and provide no further information, as they describe features of the domain that are well known to such an expert. On the other hand, some clusters provide unexpected groupings and it is these that the practitioner will focus upon.

The first partition uses two features and eight clusters (Partition I). As this uses only two features, it may be easily visualized in a 2-dimensional scatterplot. Figure 6.4 shows this scatterplot. In the figure, the feature *meanTheta*, shown on the horizontal axis, has been transformed from a mean calculated using circular statistics, into a time of day. This represents the average time of day that the player interacted with the game server. The first cluster, (cluster 0) was omitted for clarity as it contained only two players, who both died a relatively large number of times.

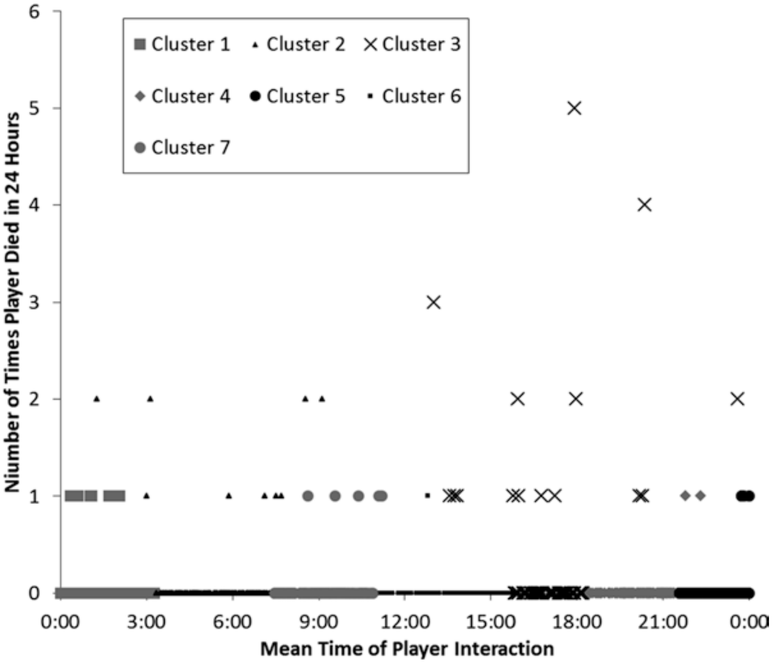


Fig. 6.4 Number of times player died in 24 h for clusters in Partition I

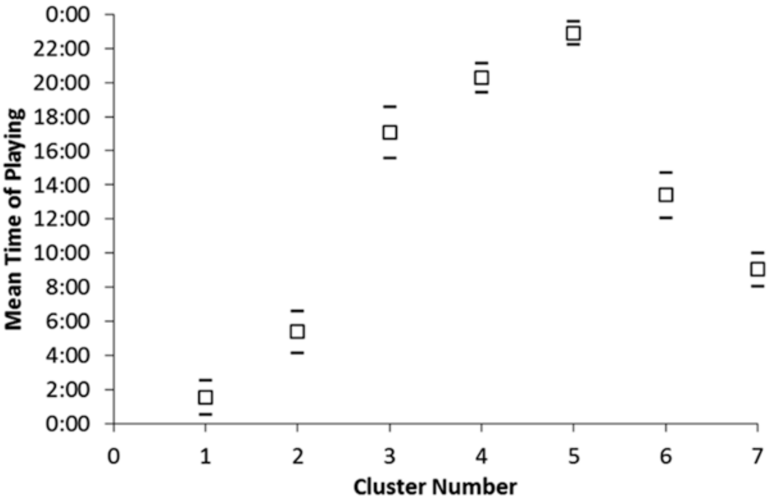


Fig. 6.5 Mean time of playing for clusters in Partition I

The figure clearly shows seven clusters of players who play at different times of day. Cluster 3 is of interest as it shows players who play in the afternoon (mean time 4 p.m. to 6:30 p.m.) that also suffer more deaths during the game. This may indicate a group of relatively unskilled players, perhaps school students beginning to learn the game, who play at this time of day.

Figure 6.5 shows the range of the mean time features for each of the seven clusters in the partition of two features and eight clusters. The mean ± one standard deviation is shown. As above, cluster 0 was omitted because of the low numbers of players. All clusters are distinct, showing that this method has identified distinct groups of players who play at different times of the day. Figure 6.6 shows the same partition but for the other feature, the number of times the player dies in play (not killed by another player). Here the difference between cluster 3 and the other clusters is clearly seen.

The second partition to be identified by a high value of the $\overline{\mathrm{SW}}$ is the partition for five features and two clusters (Partition II). A 5-dimensional feature space is not easily visualized with one graph, so Fig. 6.7 shows each feature individually. The feature which distinguishes best between the two clusters is clearly the number of times a player has killed another. A relatively high number of kills may identify either a player who has an aggressive style of gameplay, or a relatively skilled player.

4.7 Additional Partitions

Scatterplot visualization identified interesting partitions from the player kills vs commands issued, and from players kills vs. player killed. The first of these is shown in Fig. 6.8. Here the number of times a player killed another player is plotted

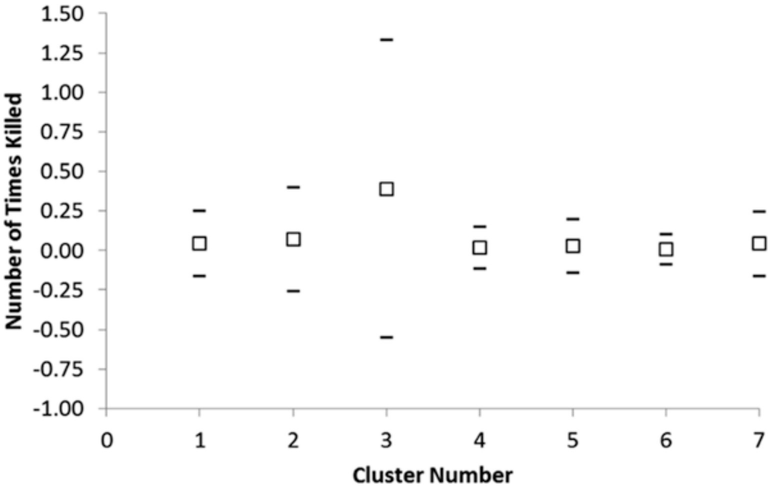


Fig. 6.6 Number of times killed for clusters in Partition I

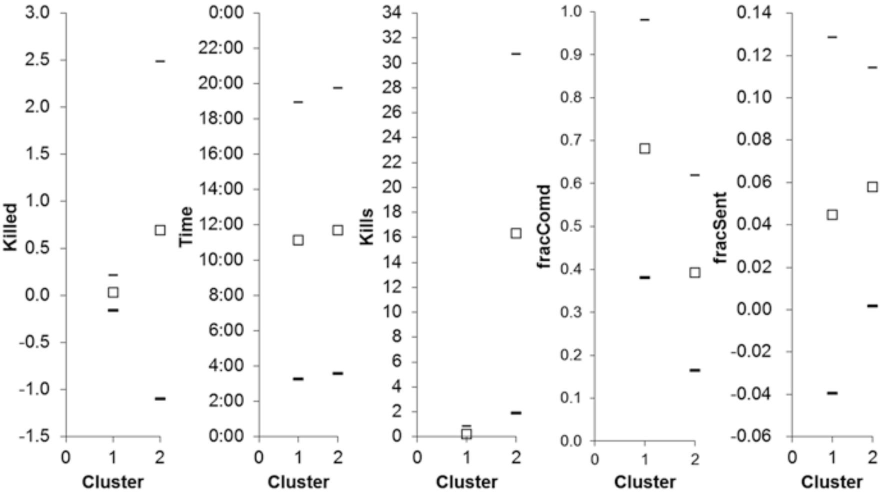


Fig. 6.7 Scatterplots for clusters in Partition II

against the number of commands issued by that player. There is no obvious pattern of clusters, and cluster analysis supported this observation. However, both features were transformed by dividing by the number of log entries found for each player. Figure 6.9 shows the transformed features. The vertical axis shows the number of times a player killed another, divided by the number of log entries for that player. The horizontal axis shows the number of times a player issued a command, divided by the number of log entries for that player. Now a distinct cluster is easily visible

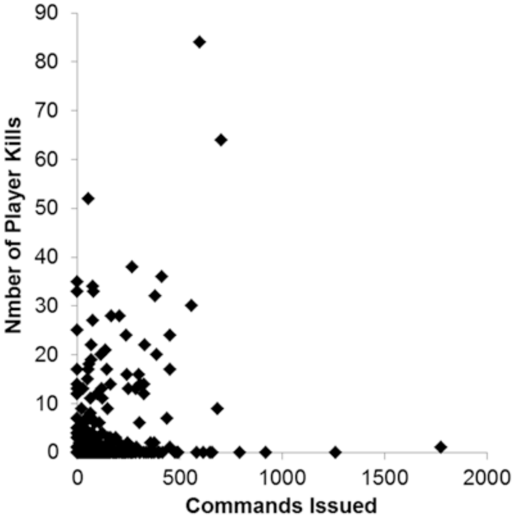


Fig. 6.8 Number of player kills against commands issued

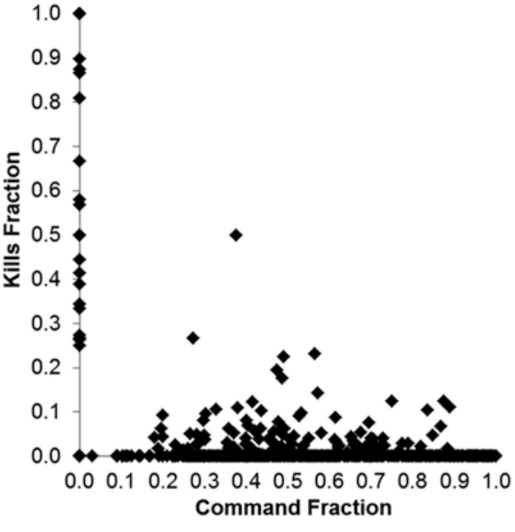
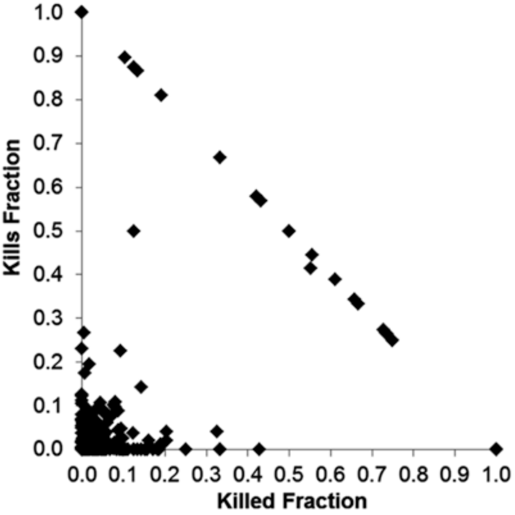


Fig. 6.9 Kills fraction against command fraction

along the vertical axis, representing a group of players who have a high number of kills but who do not issue any server commands. These commands require the player to type into a text box, so it is possible that players may find that issuing commands distracts from a competitive game. If one can assume that the number of times a player killed another is a measure of success in a competitive game, then it may that some players have developed the strategy of minimizing typed commands in order to win. The use of transformed features in this way illustrates the role of visualization and domain knowledge in identifying clusters of players.

Fig. 6.10 Number of player kills against commands issued



Scatterplot visualization identified an interesting relationship between players who killed and their targets. Figure 6.10 shows this relationship, where the vertical axis shows the number of times a player killed another, divided by the number of log entries for that player, as in Fig. 6.9. But here the horizontal axis shows the number of times a player was killed (either by another of by events in the game) divided by the number of log entries for that player. Most players have a relatively low number for both of these and are indicated by a large cluster near the origin of the graph. However, there is a very distinct cluster that forms a straight line through the middle of the graph. This cluster of players shows a strong negative correlation, where the more times a player kills another, the less likely that player is to die. This provides an insight into gameplay, depending on the type of game being played. In some games, players who kill others will rise in level and thereby will become harder to defeat. In other games, players who attack first are less likely to be themselves attacked. This result is a good illustration of the type of knowledge that can be gained from the analysis of player types. In serious games, for example, such clusters may identify players who have been able to exploit an unexpected loophole in the game to gain an advantage which undermines the serious goals of the game. This loophole can then be closed to force these players to engage better with the serious goals of the game.

5 Applications

The cluster analysis techniques discussed in this chapter are essential methods for serious games analytics. In particular, we demonstrated how to use log file data to cluster Minecraft players into different behavioral groups with distinct characteristics. While Minecraft is first and foremost a very popular online game, it is often also used as a serious game for educational and research purposes, because it realistically simulates ecology, chemistry, and physics aspects of the real world (Ekaputra et al., 2013). In Denmark, for instance, Minecraft has been used to model the entire country (Danish Geodata, 2014). Moreover, Minecraft was used to facilitate learning scientific and mathematical concepts (Short, 2012). In such educational settings, distinguishing different player types based on log file and other behavioral data can be essential for improving learning performance, as it allows (1) to systematically investigate how learning depends on individual characteristics, and (2) to adjust the learning approach accordingly. For instance, while specific learning elements might work for one specific category of players they might have no or even detrimental effects for other players. By applying cluster evaluation, description, and interpretation techniques, researchers and practitioners are able to determine such moderating influences and provide players with a learning environment that is tailored to their individual characteristics. Importantly, the techniques discussed in this chapter are not only applicable for Minecraft, but for any serious game that stores behavioral data on an individual player level. Byun and Loh (2015), for instance, found that game sound can have a positive influence on learner engagement. By combining this important result with the techniques discussed in this chapter, researchers and practitioners can disentangle how the learning engagement of different groups of players is affected by specific sound elements, which in turn allows for a personalized and adaptive learning process (Lehmann, Hähnlein, & Ifenthaler, 2014).

In this context, it is important to highlight that behavioral and learning performance differences can be both conscious and unconscious in nature. While some players might be aware of their own behavioral characteristics, e.g., above-average aggressiveness as measured by kills in Minecraft, other players might not be able to describe their own behavior and how it relates to their learning performance. For instance, Jerčić et al. (2012) and Astor et al. (2014) developed a serious game that supports financial decision-makers with learning emotion regulation capabilities. The approach is based on the rationale that emotion regulation capabilities are essential for making advantageous financial decisions and uses heart rate measurements for providing the players with a live feedback on their current level of emotional arousal. The authors found that the player's ability to control their own level of emotional arousal is moderated by their individual emotion regulation approach and that "biofeedback is to some extent processed unconsciously" (Astor et al., 2014, p. 268). Using cluster analysis for unobtrusively determining systematic behavioral differences and learning performance is therefore a promising approach for further improving such learning environments. Also beyond the scope of the individual player level, player characteristics play an important role. For instance, Feldmann et al. (2014) used a serious game for letting teams agree to allocate funding to a given set of service innovation project proposals. The authors grouped players according to their personality trait "Openness to Experience" and found that teams with a higher score on this personality trait have a stronger tendency to select radical projects than other teams. Again, applying cluster evaluation, description, and interpretation techniques can be an important additional element in disentangling how the design of the serious game as well as behavioral and personality characteristics are related to each other.

6 Conclusions

This chapter has described several techniques for evaluating player profiles based only on log files available from a game server in serious games analytics. Features were extracted from these logs in order to characterize players. Using circular statistics, players were able to be characterized by the mean and standard deviation of the time that they interacted with the server. Feature selection was accomplished using a correlation study of variables extracted from the log data. This process favored a small number of the features, as judged by the results of clustering. Automated clustering was able to suggest groups that Minecraft players fall into. Visualization of the data assisted in identifying these clusters.

The techniques described in this chapter can be applied in any serious game that stores behavioral data on an individual player level. They are therefore an important element of the serious games analytics toolset and can be combined with other techniques described in this book to identify behavioral patterns of different groups of players. The domain knowledge gained from this approach can be used to inform the design and adaptation of a particular game in order to provide the player with a personalized and adaptive environment.

Results show that it is possible to characterize players by the time of day that they play the game. However, there are also other groups, including players who are relatively successful in removing other players from the game. Some players appear to minimize their use of text boxes to issue commands during the game, and also have a relatively high success in removing other players, indicating a possible winning strategy. Other players engage in removing others and find that the more successful they are in removing other players, the less likely they are to be removed themselves. This suggests a method for identification of competitive strategies that might arise in serious games.

Acknowledgments We would like to thank Masahiro Takatsuka who supplied the data used in this study, and Samuel Cornforth who assisted with decoding server logs, and provided details of gameplay in Minecraft.

References

Ackerman, M., & Ben-David, S. (2008). Measures of clustering quality: A working set of axioms for clustering. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 21 (NIPS)*.

ARFF (2014). Retrieved July, 2014, from http://weka.wikispaces.com/ARFF

Asteriadis, S., Karpouzis, K., Shaker, N., & Yannakakis, G. N. (2012). Towards detecting clusters of players using visual and gameplay behavioral cues. *Procedia Computer Science, 15*, 140–147.

Astor, P. J., Adam, M. T. P., Jerčić, P., Schaaff, K., & Weinhardt, C. (2014). Integrating biosignals into information systems: A NeuroIS tool for improving emotion regulation. *Journal of Management Information Systems, 30*(3), 247–278.

Bolshakova, N., & Azuaje, N. (2003). Cluster validation techniques for genome expression data. *Signal Processing, 83*, 825–833.

Breaban, M., & Luchian, H. (2011). A unifying criterion for unsupervised clustering and feature selection. *Pattern Recognition, 44*, 854–865.

Byun, J., & Loh, C. S. (2015). Audial engagement: Effects of game sound on learner engagement in digital game-based learning environments. *Computers in Human Behavior, 46*, 129–138.

Danish Geodata Agency (2014). *Denmark in Minecraft.* Retrieved July, 2014, from http://eng.gst.dk/maps-topography/denmark-in-minecraft/#.U9dwS-OSySo

Duncan, S. C. (2011). Minecraft, beyond construction and survival. *Well Played: A Journal on Video Games, Value and Meaning, 1*(1), 1–22.

Ekaputra, G., Lim, C., & Eng, K. I. (2013). Minecraft: A game as an education and scientific learning tool. In Information Systems International Conference (ISICO) (pp. 237–242).

Feldmann, N., Adam, M. T. P., & Bauer, M. (2014). Using serious games for idea assessment in service innovation. In *ECIS 2014 Proceedings*, Tel Aviv, Israel (pp. 1–17).

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems, 17*, 107–145.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques. Morgan Kaufmann series in data management systems*. Burlington, MA: Morgan Kaufmann. ISBN: 0123814790, 9780123814791.

Handl, J., Knowles, J., & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics, 21*(15), 3201–3312.

Honarkhah, M., & Caers, J. (2010). Stochastic simulation of patterns using distance-based pattern modeling. *Mathematical Geosciences, 42*, 487–517.

Huang, C. L., & Wang, C. J. (2006). A GA-based feature selection and parameters optimization for support vector machines. *Expert Systems with Applications, 31*, 231–240.

Hubert, A. (1985). Comparing partitions. *Journal of Classification, 2*, 193–198.

Inza, I., Larranaga, P., Etxeberria, R., & Sierra, B. (2000). Feature subset selection by Bayesian networks based optimization. *Artificial Intelligence, 123*(1–2), 157–184.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*, 651–666.

Jang, W., & Hendry, M. (2007). Cluster analysis of massive datasets in astronomy. *Statistics and Computing, 17*(3), 253–262.

Jerčić, P., Astor, P. J., Adam, M. T. P., Hilborn, O., Schaaff, K., Lindley, C. A., Sennersten, C., & Eriksson, J. (2012). A serious game using physiological interfaces for emotion regulation training in the context of financial decision-making. In *ECIS 2012 Proceedings*, Barcelona, Spain (pp. 1–13).

Lee, M.-Y., Kim, Y.-K., & Kim, H.-Y. (2008). Segmenting online auction consumers. *Journal of Customer Behaviour, 7*(2), 135–148.

Lehmann, T., Hähnlein, I., & Ifenthaler, D. (2014). Cognitive, metacognitive and motivational perspectives on preflection in self-regulated online learning. *Computers in Human Behavior, 32*, 313–323.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129–137.

Loh, C. S., & Sheng, Y. (2013). Measuring the (dis-)similarity between expert and novice behaviors as serious games analytics. *Education and Information Technologies, 20*, 5–19.

Loh, C. S., & Sheng, Y. (2014). Maximum Similarity Index (MSI): A metric to differentiate the performance of novices vs. multiple-experts in serious games. *Computers in Human Behavior, 39*, 322–330.

Mardia, K. V. (1975). Statistics of directional data. *Journal of the Royal Statistical Society, Series B, 37*(3), 349–393.

Mitra, P., Murthy, C. A., & Pal, S. K. (2002). Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(3), 301–312. doi:10.1109/34.990133.

Nesbitt, K., & Cornforth, D. (2013). Quality assessment of clusters of electrical disturbances: A case study. In *Proceedings of the 8th IEEE Conference on Industrial Electronics and Applications (ICIEA 2013)* (pp. 247–254).

Perkins, S., Lacker, K., & Theiler, J. (2003). Grafting: Fast, incremental feature selection by gradient descent in function space. *Journal of Machine Learning Research, 3*, 1333–1356.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*(1), 53–65.

Ruß, G., & Kruse, R. (2011). Exploratory hierarchical clustering for management zone delineation in precision agriculture. In P. Perner (Ed.), *Proceedings of the 11th International Conference on Advances in Data Mining: Applications and Theoretical Aspects (ICDM'11)* (pp. 161–173). Berlin: Springer.

Short, D. (2012). Teaching scientific concepts using a virtual world: Minecraft. *Teaching Science, 58*(3), 55–58.

Strehl, A., Ghosh, J., & Mooney, R. (2000). Impact of similarity measures on web-page clustering. In Proceedings of the Workshop of Artificial Intelligence for Web Search, AAAI 2000 (pp. 58–64).

Sun, Y., Todorovic, S., & Goodison, S. (2010). Local learning based feature selection for high dimensional data analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(9), 1610–1626.

Wang, X., Yang, J., Teng, X., Xia, W., & Jensen, R. (2007). Feature selection based on rough sets and particle swarm optimization. *Pattern Recognition Letters, 28*(4), 459–471.

Waxman, O. (2012, September 21). MinecraftEdu teaches students through virtual world-building. *Time.*

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques with Java implementations*. San Francisco, CA: Morgan Kaufmann.

Part III
Visualizations of Data for Serious Games Analytics

Chapter 7
Comparative Visualization of Player Behavior for Serious Game Analytics

Günter Wallner and Simone Kriglstein

Abstract Telemetry opens new possibilities for the assessment of serious games through the continuous, unobtrusive, monitoring of in-game behavior. Data obtained through telemetry thus not only contains information about the outcomes but also about the intermediate processes. In this sense, telemetry data can be of value for various stakeholders of serious games, including developers, educators, and learners themselves to increase the effectiveness of the intervention. In doing so, particular significance should be attached to differences among individuals and demographic groups in order to understand and better accommodate for these variations. However, the large amounts of data gathered via telemetry can make it challenging to derive meaningful information from it. Visualizations can support this process by providing a means to explore, to compare, and to draw insights from the data sets. In this chapter, we discuss three common visual design strategies that facilitate comparative data analysis. Several examples, drawn from the game-based learning literature and related areas as well as two detailed case studies are used to illustrate how these strategies can be leveraged in the context of serious game analytics.

Keywords Game telemetry • Player behavior • Visualization • Visual comparison

1 Introduction

Serious games have the potential to promote knowledge transfer by being engaging and entertaining. This, however, poses the challenge to find a proper balance between entertainment and fun while at the same time conveying knowledge or promoting behavioral change. Moreover, developers of serious games have an ethical responsibility to ensure that the game does not cause negative unintended

G. Wallner (✉)
University of Applied Arts Vienna, Oskar Kokoschka Platz 2, Vienna 1010, Austria
e-mail: guenter.wallner@uni-ak.ac.at

S. Kriglstein
Vienna University of Technology, Argentinierstrasse 8, Vienna 1040, Austria
e-mail: simone.kriglstein@igw.tuwien.ac.at

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based
Learning, DOI 10.1007/978-3-319-05834-4_7

consequences, like building incorrect mental models of the taught concepts (Warren, Jones, & Lin, 2011) as re-teaching can be costly (Loh, 2012), difficult, and time consuming (Warren et al., 2011). Careful evaluations with the intended target audience are therefore necessary to validate the effectiveness of game-based learning applications. In contrast to entertainment games, however, educational games have to cater to a more diverse player audience, as pointed out by Magerko, Heeter, and Medler (2010), making it important to consider differences among learners and demographic characteristics, such as gender or age, when designing and evaluating educational games. As de Freitas and Jarvis (2006) emphasized, preferences and differences of learners should be taken into account as early as possible to better integrate learning outcomes into the gameplay itself. Gender, for example, has been shown to influence, among others, the need for achievement, challenge, and competition in games as well as the preference regarding genre and game speed (cf. Heeter, Lee, Magerko, & Medler, 2011; Steiner, Kickmeier-Rust, & Albert, 2009). Age-related factors like deficits in short-term memory (Wouters, van der Spek, & Van Oostendorp, 2009) or the role of brain maturation on problem-solving abilities (Gelderblom & Kotzé, 2009) can also have an effect on player behavior and the educational outcome.

However, commonly used evaluation methods, like the widely used pre- and posttest design, to assess the effectiveness of serious games (cf. Becker & Parker, 2011; Bellotti, Kapralos, Lee, Moreno-Ger, & Berta, 2013) have been criticized to neglect the intermediate processes and rather view serious games as a black box (Kriz & Hense, 2006; Loh, 2012). Yet, understanding why a serious game works (or not) is considered to be equally important (cf. de Freitas & Oliver, 2006; Kriz & Hense, 2006) and can inform the design of future games.

Data collection via telemetry has therefore received increasing attention as a means to capture the in-game behavior of individual players. In educational games research, telemetry is valued as being able to continuously monitor the learner (Anolli & Confalonieri, 2011) while at the same time being objective and noninvasive as playing is not disrupted (Linek, Öttl, & Albert, 2010). In this respect, telemetry also helps to overcome the traditional dichotomy between learning and assessment, as emphasized by Anolli and Confalonieri (2011), where assessment usually takes place after the intervention. Instead, telemetry data provides continuous feedback on the progress of learners, hence supporting formative assessment (cf. Bellotti et al., 2013). Subsequently, this knowledge can be used to provide ongoing feedback to users or to dynamically make adjustments to the learning environment (cf. Bellotti et al., 2013; Shute, Ventura, Bauer, & Zapata-Rivera, 2009). Put differently, telemetry data offers opportunities to assess whether users actually used the intended processes to achieve a learning goal or not. This, however, naturally raises the question how the tracked in-game behavior relates to learning performance. To address this complex issue, assessment frameworks which require to specify how learners must behave to provide evidence about the skills and competencies to be conveyed, such as Evidence Centered Design (e.g., Mislevy & Riconscente, 2005), have been adopted for the analysis of behavioral log data. Others have relied upon comparing the logged behavior with reference solutions by experts (Fardinpour, Reiners, & Dreher, 2013; Loh & Sheng, 2013). However, it is beyond the scope of this chapter to review these different approaches in detail. An introduction to Evidence Centered Design and related frameworks can, for example, be found in Plass et al. (2013) and Shute et al. (2009).

In this chapter, we are instead concerned with another major challenge associated with telemetry data, namely how to extract meaningful information from the wealth of collected data and how to clearly communicate this information. Graphical representations of the data take on an important role in this context, as they enable us to explore and draw insights from the data in an efficient and effective way (cf. van Wijk, 2005). This includes, among others, the examination of differences among individuals or groups. Comparative visualizations can be of benefit for at least three major stakeholders of game-based learning applications:

Developers/Researchers: First, telemetry data and visualizations thereof can aid developers or researchers in assessing their game in regard to (a) game design aspects and (b) pedagogical effectiveness. However, especially the former is considered to be an often overlooked aspect in serious game development (cf. Moreno-Ger, Torrente, Hsieh, & Lester, 2012; Olsen, Procci, & Bowers, 2011; Warren et al., 2011). Yet, issues pertaining to usability, playability, balancing, or difficulty can all negatively affect player experience and, ultimately, knowledge transfer. Visualizations can be helpful in this regard to uncover such issues and to tailor a game to the different needs of different demographic groups within the game's target audience. Concerning the second point, visualizations can assist in determining if differences in the educational outcome can be attributed to differences in in-game behavior.

Teachers/Instructors: Visualizations aid teachers in monitoring and comparing the progress of their students (Govaerts, Verbert, Klerkx, & Duval, 2010; Minović & Milovanović, 2013), helping them to ensure that all students are on track (Loh, 2012) or to spot students which face problems or need more attention (Govaerts et al., 2010; Serrano-Laguna, Torrente, Moreno-Ger, & Fernández-Manjón, 2012) and, if necessary, to make responsive changes to adjust the learning process to the abilities of individual learners. However, visualization capabilities of, for example, current learning management systems are reported to be nonexistent or very basic (Dawson, 2010) and to rather focus on displaying post-training outcomes than the learning process itself (cf. Ritsos & Roberts, 2014).

Players/Learners: Lastly, visualizations can be used by the learners themselves to monitor their progress (Govaerts et al., 2010) and to understand how they perform compared to their peers (Duval, 2011; Govaerts, Verbert, Duval, & Pardo, 2012). This, in turn, can foster self-reflection, increase motivation, and encourage competition and collaboration among students (see, e.g., Govaerts et al., 2010).

In the following, we will discuss three common visual design approaches which facilitate comparative analysis and show, by reviewing some examples from the game-based learning literature and related fields, how these approaches can be utilized for serious game development and evaluation. We conclude the chapter with two illustrative case studies.

2 Comparative Visualization

There exist different ways how visual structures can be composed in order to compare data sets to discover similarities and differences. For example, Graham and Kennedy (2010) present different representation techniques for the analysis and comparison of tree structures. A general taxonomy, which is independent of data types, applications, and domains, was introduced by Gleicher et al. (2011) who present three strategies for the visual design of data sets. One strategy, called *juxtaposition*, displays the visualizations of the data sets separately and places them side by side. *Superposition*—another strategy—places the visualization of the data sets in the same coordinate system by overlaying or alternating them. The third strategy, *explicit encoding*, directly represents the relationships between the data sets. Although these three categories by Gleicher et al. (2011) are quite general, different adaptations were made in the last years, especially for explicit encoding. For example, Beck, Burch, Diehl, and Weiskopf (2014) define the strategy *integration* which can be seen as a form of explicit encoding since the different visualizations of data sets are integrated into one. Beside juxtaposition and superposition, Javed and Elmqvist (2012) draw a distinction between *integration*—juxtaposed views which use explicit visual links to relate objects, *overloading*—different visualizations utilize the same space but without a one-to-one spatial linking, and *nesting*—one or more visualizations are nested inside another visualization, which can as well be considered as different strategies for explicit encoding.

In this chapter, however, we will follow the taxonomy by Gleicher et al. (2011). As an introductory example, Fig. 7.1 illustrates these three categories, by means of two heat maps. These heat maps were generated from replay data from a *Starcraft 2* match between two players.

One heat map shows the death locations of units from the first player and the other shows the death locations from the second player. In Fig. 7.1 (top), these two heat maps are placed side-by-side. The lower left image shows the superposition of both heat maps using alpha blending. The lower right image encodes the differences explicitly by subtracting the values from the second heat map from the values of the first heat map and mapping the resulting difference values to a color gradient (cf. Houghton, 2011). In this case, positive values are mapped to the color gradient of the first player whereas negative values are mapped to the color gradient of the second player. The color therefore reflects which player had more casualties in particular areas. The brightness reflects the magnitude of difference. Compared to the superposition in this particular case, the explicit encoding highlights the differences better since the superposition of the two heat maps causes occlusions that make the gradient of the underlying heat map hard to gauge.

In the following sections, each of the three categories will be discussed further to show their potential for comparative analysis.

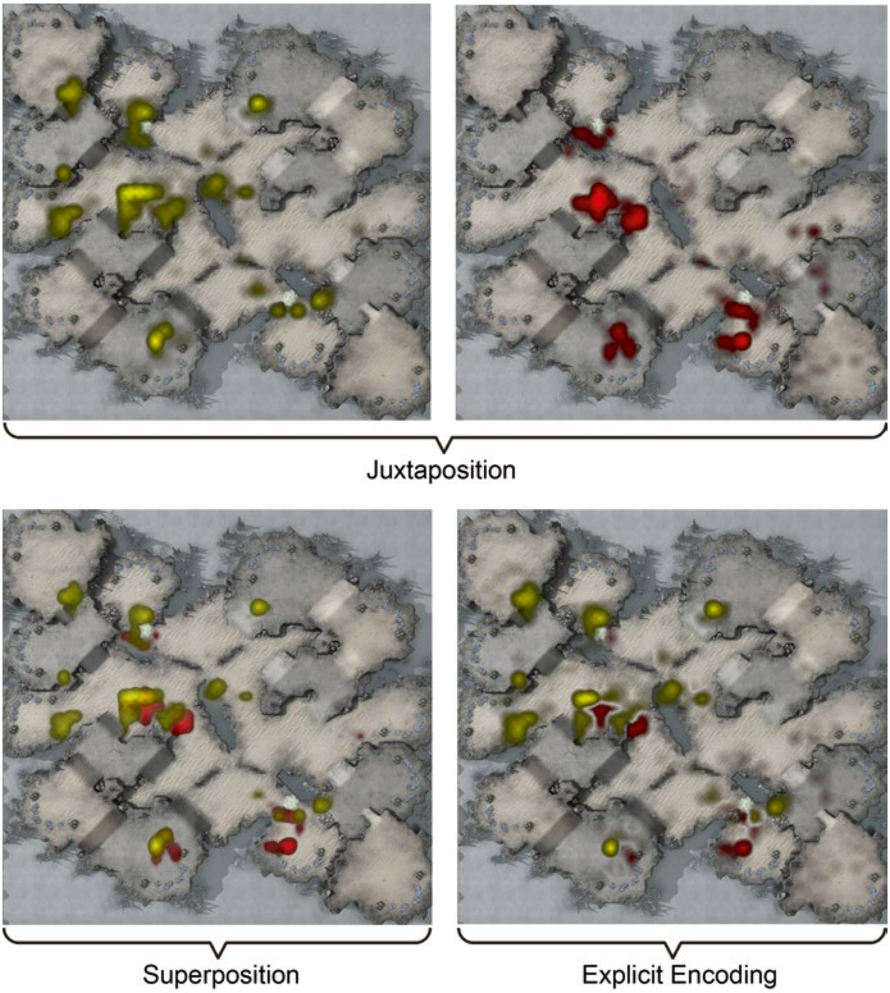


Fig. 7.1 Three basic approaches for comparative visualization illustrated by means of heat maps

2.1 *Juxtaposition*

The juxtaposition approach displays each visualization separately in their own view. According to Roberts (2005), side-by-side representations are useful for exploring and comparing differences and similarities. Javed and Elmqvist (2012) point out that this approach is the most prominent because of its flexibility in how to arrange and visualize different data sets and because of its ease of implementation. The usage of side-by-side comparison for different data sets can be found in many different applications, for instance, to compare different text versions (cf. Ferster & Shneiderman, 2012), glyphs (e.g., Ward, 2002), or trees and graphs (see, e.g., Federico, Aigner, Miksch, Windhager, & Zenk, 2011; Munzner, Guimbretière, Tasiran, Zhang, & Zhou, 2003).

Gleicher et al. (2011) distinguish between time and space for the separation. Juxtaposition in space shows the visualizations of the data sets in parallel to facilitate efficient and effective comparisons across different visualizations (Kirk, 2012; Yau, 2011). Space juxtaposition is sometimes also referred to as *small multiples*, a term introduced by Tufte (1990). Small multiples present a series or grid of small, thumbnail-sized visualizations of data sets, in order to get answers "*directly by visually enforcing comparisons of changes*, *of the differences among objects*, *of the scope of alternatives*" (Tufte, 1990, p. 67). However, since the viewer has to compare separate visualizations it may be difficult to see the relationships between them, as pointed out by Gleicher et al. (2011). Therefore, particular attention should be attached to the design of small multiples. For example, visual cues, like highlighting matching objects between the different views, can assist viewers to spot relationships.

Animation, i.e., a sequence of visualizations, can be seen as a juxtaposition in time, according to Gleicher et al. (2011), if "*it predominantly requires the use of the viewer's memory and attention shifts to make connections between objects*" (Gleicher et al., 2011, p. 294). Over the last years, several studies compared animation and static visualizations, like small multiples, regarding the extraction of different kinds of information (cf. Boyandin, Bertini, & Lalanne, 2012) or patterns (see, e.g., Griffin, MacEachren, Hardisty, Steiner, & Li, 2006). Archambault, Purchase, and Pinaud (2011), in turn, compared the performance of animation and small multiplies in regard to the visualization of dynamic graphs and assessed the effects of mental map preservation on both representations. The different studies show that animation and small multiples have different advantages and disadvantages depending on the type of task to be performed. For example, Robertson, Fernandez, Fisher, Lee, and Stasko (2008) present a user study that shows that the usage of animation seems to be successful for presentation tasks but static depictions, like small multiples, may be more effective for analysis tasks. Although studies show that motion can be helpful to follow changes in data (cf. Kriglstein, Pohl, & Smuc, 2014; Kirk, 2012) points out that the usage of animation seems not to be the best method for comparison tasks, and it may be more effective to use small multiples.

2.2 Superposition

In contrast to the juxtaposition approach, the *superposition* approach visualizes the different data sets in such a way that they share the same visual space by overlaying or alternating the visualizations on top of each other in the same coordinate system. The usage of such overlaid visualizations to analyze the differences and commonalities of data sets ranges from the comparison of graphs or trees (see, e.g., Brandes, Dwyer, & Schreiber, 2004) to heat maps (e.g., Drachen & Canossa, 2011) in various domains.

To distinguish between the representations of different data sets in the same space, one possible solution is the use of different colors (see, e.g., Erten, Kobourov, Le, & Navabi, 2003), including techniques—as outlined by Gleicher et al. (2011)—like color weaving (e.g., Urness, Interrante, Marusic, Longmire, & Ganapathisubramani, 2003), attribute blocks (cf. Miller, 2007), or semitransparency (e.g., Federico et al., 2011). If more than two dimensions are from interest, a 2.5D technique can be useful in order to stack the different visualization on top of each other (see, e.g., Brandes & Corman, 2003; Brandes et al., 2004; Tominski, Schulze-Wollgast, & Schumann, 2005).

Overlaying visualizations in the same view has the advantages that it is easier to understand them in the context of each other and that the viewer can compare the data sets without having to split the attention between more than one view (cf. Gleicher et al., 2011; Roberts, 2005). As pointed out by Javed and Elmqvist (2012), a further advantage is that the full available space can be used. However, they also note that stacked or overlaid visualizations can lead to visual clutter or may occlude interesting information. Such occlusion problems, as mentioned by Roberts (2005), can especially occur in 2D representations and can lead to misunderstandings about how many objects are in fact visualized.

2.3 Explicit Encoding

According to Gleicher et al. (2011), the *explicit encoding* approach explicitly visualizes the relationships between the different data sets in a dedicated visualization in order to support viewers to detect, for example, differences, correlations or similarities between them. One of the most widespread techniques to visually encode the differences between data sets is the use of different colors. Examples include, among others, Andrews, Wohlfahrt, and Wurzinger (2009), Guerra-Gómez, Buck-Coleman, Pack, Plaisant, and Shneiderman (2013), and Kriglstein, Wallner, and Rinderle-Ma (2013) for tree and graph visualizations or Beck, Burch, and Weiskopf (2013) for matrix visualizations. Another way is to draw lines between the objects in order to trace the relationships between the visualized data sets (see, e.g., Dwyer, Hong, Koschützki, Schreiber, & Xu, 2006; Holten & van Wijk, 2008; Stewart et al., 2001).

In contrast to superposition and juxtaposition, explicit encodings can minimize the viewer's effort by providing a visual encoding that allows the viewer to directly see the relationships between the data sets in a single visualization. Piringer, Pajer, Berger, and Teichmann (2012) found out that the explicit visualization was for a precise comparison of differences very valuable in contrast to the juxtaposition approach. However, Gleicher et al. (2011) point out that a prerequisite is to have prior knowledge of the data sets and their possible relationships to be able to specify the relationships which should be depicted graphically. This can lead to restrictions concerning exploration and detection of unknown but interesting correlations between the different data sets. Furthermore, since the explicit encoding approach can influence the visual structure of the different data sets, a conflict can occur between viewers' mental model of the original visualizations of the data sets and the modified visualization representing the explicit encoding of the relationships. A possible solution is a hybrid approach which depicts not only the explicit encoding but also the individual visualizations of the data sets in separate views (see, e.g., Andrews et al., 2009; Guerra-Gómez et al., 2013; Kriglstein et al., 2013).

3 Comparative Visualization in Serious Game Analytics

While serious games have the potential to stimulate learning, individual differences among learners can also mitigate the effectiveness of the treatment for different types of individuals. Personal characteristics such as gender and gender-related differences regarding challenge, competition, or sensation seeking (e.g., Heeter et al., 2011; Steiner et al., 2009), age (e.g., O'Rourke, Butler, Liu, Ballweber, & Popović, 2013), or differences in self-efficacy (e.g., Ketelhut, 2007; Rowe, Shores, Mott, & Lester, 2010) or visual attention (e.g., Arthur et al., 1995) have shown to be capable of influencing the success of educational games, as have genre preferences and experience in playing games (e.g., Heeter et al., 2011; Magerko et al., 2010; Steiner et al., 2009). Consequently, special emphasis should be placed on recognizing and accommodating these differences. As elaborated earlier, visualizations can assist in this task by providing a means to explore and draw insights from behavioral player data. Unfortunately, examples which utilize visualization techniques other than traditional charts to draw comparisons are still quite sparse in game-based learning research and related fields such as eLearning (see also Ritsos & Roberts, 2014). From among these, we discuss some recent works in the following. In general, these can be roughly divided into two categories: (a) papers that propose new visualization approaches or tools and (b) studies that use existing graphical representations (e.g., heat maps) as part of their analysis. While in the latter case it is not always obvious which comparison strategy has specifically been employed, these examples show how visualizations can be of value in determining differences.

With regard to the first category, Andersen, Liu, Apter, Boucher-Genesse, and Popović (2010) proposed a graph-based visualization tool to understand player strategies and to uncover common points of confusion. For that purpose a game is considered to be composed of a set of states (represented by the vertices of the graph) between which the players are moving around (depicted by the edges) by interacting with the game. One of the features of the tool allows users to visually compare the graph of players who won against the graph of players who lost to see if there are differences in behavior. They demonstrated the usefulness for educational game design by applying the visualization to tracked player data from a grid-based puzzle game about fractions. However, games with a large number of states can cause the graph representation to become cluttered and thus difficult to read. This shortcoming was addressed in a follow-up paper by Liu, Andersen, Snider, Cooper, and Popović (2011) by merging states that share the same preselected features into a single state.

Scarlatos and Scarlatos (2010) proposed a variation on parallel coordinates to create glyphs that reflect the choices made by individual players. Instead of connecting the values of the axes (which represent the different choices available to the player) with line segments, these values form the corners of a closed polygon. By properly ordering the axes, *favorable* behavior in the context of the game can be distinguished from *unfavorable* behavior by the shape of the glyph. These glyphs can, for example, be used to reflect aggregated player choices or can be displayed on a timeline to evaluate player performance over time. In either case, small multiples can be used to compare multiple players with each other.

Govaerts et al. (2010) and Duval (2011) proposed a visualization tool which allows users of personal learning environments to track their progress and compare their performance with their peers. The tool uses multiple views and different visualization techniques, among them, a line chart showing total time spent on the course for each student and a parallel coordinate's plot (Inselberg, 1985). The tool uses different colors to highlight the current user and the average student to facilitate comparison. Desmarais and Lemieux (2013), on the other hand, combined clustering techniques with a timeline visualization to understand the patterns of use of a learning environment. To this end, a sequence of activities is derived from the logged event data for each session. These sequences are then processed by a clustering algorithm to extract common patterns. Finally, each cluster is visualized using a separate timeline visualization with each horizontal row in the timeline representing the sequence of activities of an individual session within a cluster. This enables designers and teachers to understand different usage patterns and make adjustments to the learning environment if necessary. Although both of these approaches are targeted towards learning environments, similar approaches could be used to assess the progress in serious games as well.

Although not directly within the scope of this chapter as it does not make use of player generated data to understand user behavior, the approach by Butler and Banerjee (2014) is also worth mentioning here as it allows designers to compare and reason about progressions already in the design phase and because player data could be incorporated as well as noted by the authors. To that end, a progression is considered to be a sequence of stages, with each stage consisting of a set of different concepts. The two progressions to be compared are then visualized in multiple views. One view overlays graph representations of the progressions over each other. Nodes represent the different stages and the distances between the nodes reflect the similarity of the stages in terms of concepts. A second view juxtaposes two bar charts that show how often different concepts occur per stage within a progression.

Shifting the focus to the second category, examples include O'Rourke et al. (2013) who studied the effects of age on the behavior in two educational games. In-game data was gathered from two websites which target two different age groups. As part of the analysis they used graphs to visualize how players search the space of possible moves to find solutions. Comparison of the graphs of the two age groups revealed that the younger group was searching more broadly and less focused compared to the older group.

Kiili, Ketamo, and Kickmeier-Rust (2014) investigated if high performers and low performers differ in gaze behavior in a game about geography using eye-tracking technology. Beside a statistical analysis of gaze fixations and saccades which showed differences in amount and length, heat maps revealed that low performers—compared to high performers—exhibited a tendency to pay too much attention to areas of little relevance. Similarly, Mehigan, Barry, Kehoe, and Pitt (2011) used eye tracking to investigate the gaze patterns of verbal and visual learners in an eLearning environment. Analysis of the resulting heat maps and gaze plots (gaze plots visualize gaze fixations and the order in which they occur) revealed significant differences in gaze behavior between these two learning styles. In a similar way, Buendía-García, García-Martínez, Navarrete-Ibañez, and Jesús (2013) used heat maps to compare the interaction behavior of novice and expert players in a game about workplace ergonomics.

While these examples illustrate how visualizations can be utilized for comparative analysis, it is also worth mentioning that we could observe that there exist hardly any examples (e.g., Wallner, 2013) that make use of explicit encoding.

4 Case Studies

In this section, we discuss two case studies which explore gender and age-related differences in two educational games. In both cases, the in-game behavior of the players was tracked by instrumenting the source code of the game.

4.1 Case Study: Gender Differences

Internet Hero (Kayali et al., 2014) is a game for children between nine and twelve years to make children aware of the technical and social aspects of Internet use. The game is composed of different mini-games each correlating to a different aspect of the Internet, like spam or malicious software. A preliminary evaluation with 36 children (18 male, 18 female) to assess the playability and appropriateness of the difficulty level of the first two mini-games revealed, as reported in Kayali et al. (2014), that females scored considerable lower in the malicious software mini-game. This mini-game is a tower defense game where players have to fight off incoming waves of viruses by placing different types of towers, resembling security measures like firewalls, spyware scanners, and antivirus software. Points were awarded for successfully destroying viruses. Specifically, males scored 486 points on average while females only received 296 points on average in the first play-through.

In the following, we will use small multiples to compare the game metrics of individual players in order to investigate if these differences in score can be attributed to certain metrics. Figure 7.2 uses star plots to visualize the different game

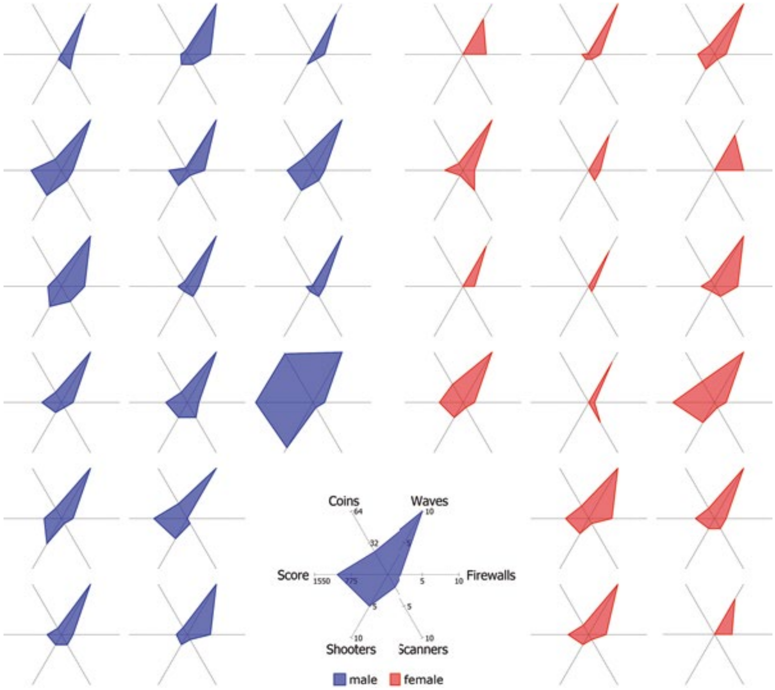


Fig. 7.2 Small multiples of star plots. Each star plot represents the game metrics of an individual player. Star plots colored *dark gray* belong to female players and *light gray* star plots to male players

metrics—collected during the first play-through—separately for each player. Each axis reflects one measured variable and the distance from the center corresponds to the value of the variable. Please note that different scales have been used for different metrics. To aid comparison of gender differences, star plots of female players and male players are depicted in different colors.

First, it is apparent that the plots of many female players are confined to the right half, indicating that they neither collected many coins nor achieved a high score or built many shooters. However, from the charts it appears that the score is related to the number of constructed shooters, as players building more shooters also achieved a higher score. This is not surprising as only shooters are able to kill viruses and points were only rewarded for destroyed viruses. Yet, females rarely used shooters, as pointed out above, but rather firewalls and scanners to keep out the viruses. Boys and girls therefore employed two different strategies, which may be explained by the lower interest of females in violent conflict resolutions (cf. Hartmann & Klimmt, 2006; Peirce & Edwards, 1988; Steiner et al., 2009). However, the scoring scheme favored the strategy of the males over the, similarly successful, play behavior of girls (this is evident from the plots as almost all players survived nearly all waves, even with a small number of shooters).

Looking at the game metrics of subsequent play-throughs revealed an even bigger difference in score, with males rising their score to around 1,000 points on average in the third play-through (by building approximately twice as much shooters as during the initial play-through) compared to females who only increased to around 490 points on average. This may be due to boys displaying stronger competition orientation in games than females (e.g., Hartmann & Klimmt, 2006) and therefore attaching greater importance to the score. However, as both strategies are valid in terms of the learning context (protecting a computer from viruses), we changed the scoring scheme to better accommodate for both genders (cf. Kayali et al., 2014).

4.2 Case Study: Age Differences

DOG*eometry* (Wallner & Kriglstein, 2012a) is an educational puzzle game to teach young children from 8 to 10 years concepts about transformation geometry (translation, rotation, and reflection) and object hierarchy. The game comprises an object editor and a series of puzzles with increasing difficulty. Each of the puzzles requires the player to build a continuous path for a dog to a veterinarian by placing a limited number of road tiles (straight segments, turns) on a grid and arranging these tiles with a limited number of transformations. Obstacles on the grid, like water holes, complicate the task. Some of the puzzles can be solved in different ways with more complicated solutions allowing players to collect bones for the dog. These bones act as a reward system to motivate players to aim for more complicated solutions. Collected bones can also be exchanged for hints in subsequent puzzles.

The game was evaluated using a pretest/posttest control group design. Statistical analysis of the test scores (see Wallner & Kriglstein, 2012a) showed a main effect of age on the treatment effect, with 8-year-olds improving only marginally compared to 9- and 10-year-olds. In such a case, comparative visualizations can help to understand differences in in-game behavior which may influence the learning effect and consequently to implement changes to remedy these problems. To assess the in-game behavior, we will use a graph-based approach first introduced in (Wallner & Kriglstein, 2012b) and later extended to difference analysis (Wallner, 2013) in this case study. Graphs produced by this approach give an aggregated overview of the in-game behavior of multiple players and can, expressed in simplified form, be viewed as weighted directed graph where nodes represent states (e.g., the particular arrangement of the pieces on a *Chess* board or, as in the present case, the arrangement of the road tiles on the grid) and edges depict transitions from one state to another, triggered by a player by interacting with the game (e.g., moving a pawn or placing or translating a road tile). Node weights and edge weights represent how many players arrived at a state or triggered a particular change in state, respectively.

By way of example, Fig. 7.3 gives a side-by-side comparison of the in-game behavior of 8-year-old ($n = 22$, left) and 9-year-old children ($n = 13$, right) for the ninth puzzle of DOG*eometry*. As the states do not contain any spatial information

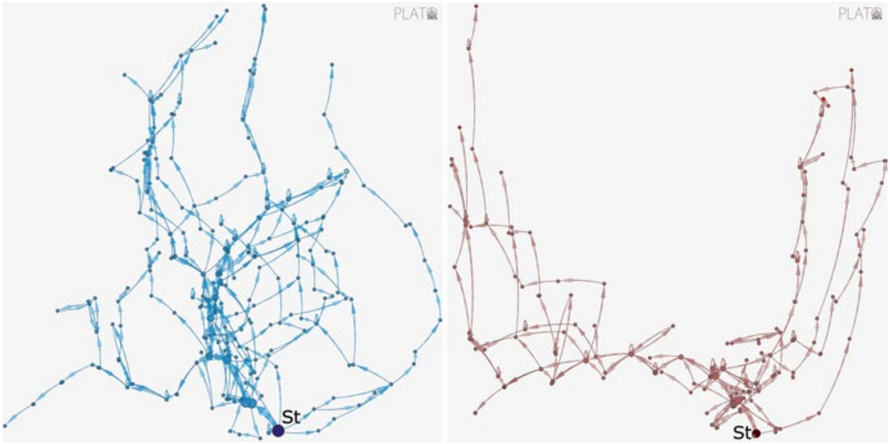


Fig. 7.3 Side-by-side comparison of the in-game behavior between 8-year-old (*left*) and 9-year-old children (*right*)

which can be directly leveraged to obtain a placement of the nodes, the embedding of the graphs was obtained using multidimensional scaling (Kruskal & Wish, 1978) such that nodes corresponding to similar arrangements of road tiles are placed near to each other while nodes with dissimilar arrangements are placed farther away. [Note: multidimensional scaling allows to graphically examine the similarities among objects. In short, given a set of objects and a matrix describing the dissimilarity between pairs of objects, multidimensional scaling aims to find an embedding of these objects such that their distance in the embedding approximates the original dissimilarities.]

Agglomerations of nodes therefore indicate that players were experimenting with similar arrangements but were uncertain on how to best proceed. The thickness of the edges and the radius of the nodes are proportional to the edge and nodes weights. The node, labeled *St*, at the bottom of each graph represents the starting configuration (see Fig. 7.4, bottom, for a graphical representation of this state). The different degree of complexity of these graphs already indicates certain differences in behavior. However, as the different layouts of the graphs make it difficult to assess where the actual differences occur it can be beneficial to derive a single graph that directly encodes the differences between the two.

Given two weighted graphs $G_1 = (N_1, E_1)$ and $G_2 = (N_2, E_2)$ based on different sets of players, P_1 and P_2, the difference between G_1 and G_2 can be computed as $G_d = G_1 - G_2 = (N_d, E_d)$, with $N_d = N_1 \bigcup N_2$ and the weight w_d of a node $n \in N_d$ given by $w_d = w_1 / |P_1| - w_2 / |P_2|$, where w_1 and w_2 are the weights of n in G_1 and G_2 and $|P_1|$ and $|P_2|$ are the number of players on which the graphs are based. E_d is the set of edges with a resulting edge weight $\omega d = \omega_1 / |P_1| - \omega_2 / |P_2|$ unequal to zero, where ω_1 and ω_2 are the edge weights in G_1 and G_2. Please note, that the node and edge weights are related to the number of players do not skew the resulting difference

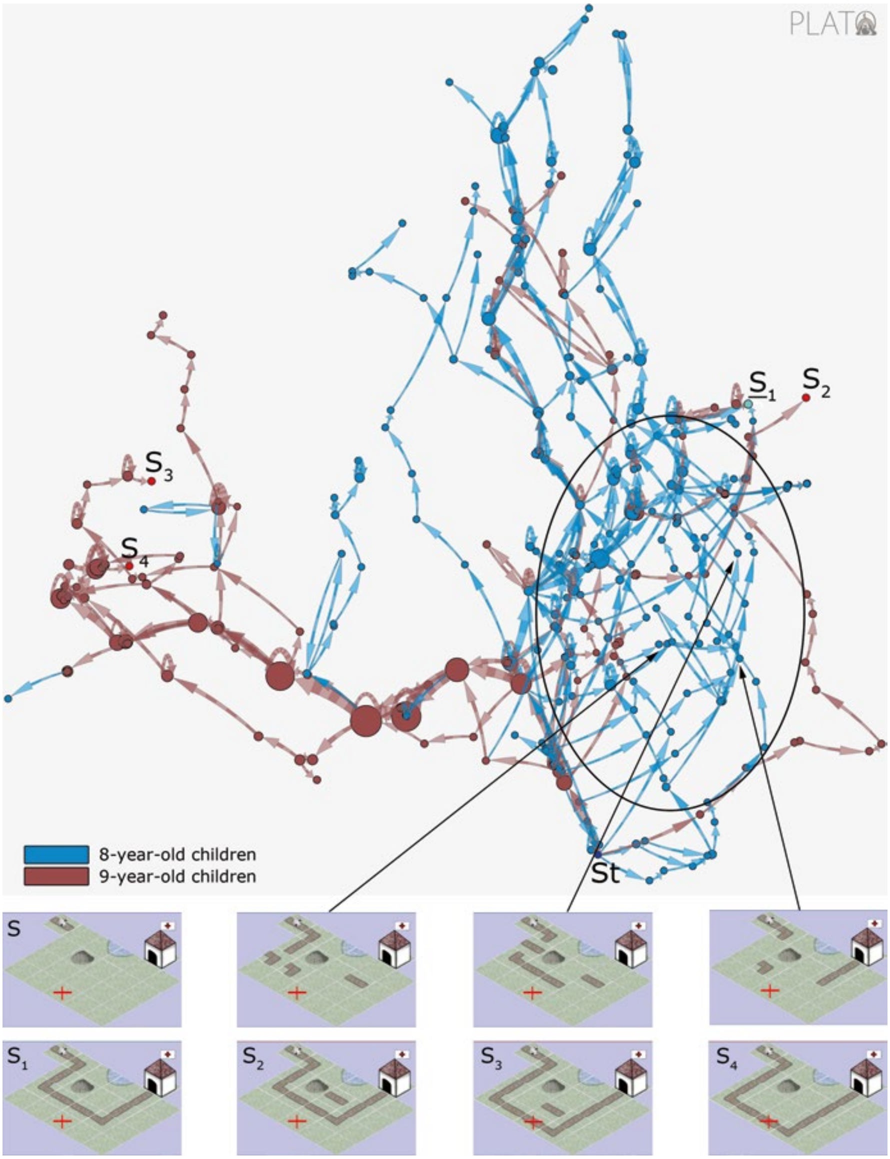


Fig. 7.4 Graph showing the relative differences between the in-game behavior of 8-year-old and 9-year-old children of the ninth puzzle in DOG*eometry*

graph towards one of the two input graphs in case they are based on different sample sizes. A more detailed description can be found in (Wallner, 2013). The resulting difference graph can then be visualized using explicit encoding and color-coding. As an example, Fig. 7.4 shows the resulting difference graph for the ninth puzzle, obtained by subtracting the gameplay graph of 9-year-old children from the graph of 8-year-old children.

Parts more frequented by 9-year-olds are depicted in another color than parts more commonly taken by 8-year-olds (see the color legend in Fig. 7.4). Solutions are labeled with S_1 to S_4. If the label is underlined, then the corresponding solution was more often found by 8-year-olds; otherwise, the solution was more frequently found by 9-year-olds. Screenshots below the graph depict the arrangement of the road tiles for some selected nodes. The *X* in these screenshots marks the location of the bone.

A few interesting insights can be gained from this visualization. First, the part of the graph more frequented by 8-year-olds in relation to 9-year-olds appears more cluttered, especially in the area between the start node *St* and solution S_1 (encircled in Fig. 7.4). A closer look at the arrangement of the road tiles in this area, some of them are depicted in Fig. 7.4, bottom, shows that 8-year-olds were rather uncertain about how to proceed in order to solve the puzzle and were trying out a lot of different arrangements. Second, 8-year-old children also focused mostly on the easier solution (S_1) and did not even attempt the more difficult solution (S_4) as the path towards S_4 is quite thick, indicating that this part has been much more frequently traversed by 9-year-olds compared to 8-year-old children. 9-year-old children were also rather sure in which order to arrange the tiles to reach this solution since the path towards S_4 is rather straight with a low number of branches. In summary, 8-year-olds proceeded less strategically than 9-year-old players and rather followed a trial-and-error approach to solve this particular puzzle.

5 Conclusions

Data collection and analysis of telemetry received increasing attention in the serious games community over the last years because of its ability to continuously and unobtrusively monitor the in-game behavior of players. Telemetry data can thus provide valuable insights into the progress and performance of individual players. Visualizations of the collected data can be of benefit for various stakeholders of serious games, including developers, researchers, instructors, and learners to gain insights and, in turn, guide them in decision-making. To take advantage of this potential it is, however, essential to ensure that the represented data will be interpreted correctly by the target audience. Research on graph comprehension has shown that the interpretation of visualizations is influenced by various factors, including background knowledge and visual characteristics of the representation (cf. Shah & Hoeffner, 2002). Care should also be taken to ensure that the visualizations have the desired impact on the target audience as, for example, inappropriate feedback can have detrimental effects on learners, like a decrease in motivation (cf. Gašević, Dawson, & Siemens, 2015; Westera, Nadolski, & Hummel, 2014). However, studies assessing the impact of visualizations on the learning process in real-life settings are still mostly lacking (Klerkx, Verbert, & Duval, 2014). For example, a recent survey

(Verbert et al., 2014) of learning dashboards showed that only seven out of the 24 surveyed systems have been evaluated in regard to their effectiveness on learning. Only one of the seven systems was assessed as part of a long-term study.

In this chapter, we specifically focused on comparative visualization approaches to understand differences among individuals and demographic subgroups. To this end, we first briefly discussed three common visualization strategies—juxtaposition, superposition, and explicit encoding—which facilitate comparative data analysis. In addition, we presented a range of examples which apply these strategies in the context of serious games or related areas. More specifically, we focused on examples which make use of visualization techniques other than traditional charts. In doing so, we observed that such examples are still rare in the serious games literature (see also Ritsos and Roberts 2014 who made a similar observation in the area of learning analytics). However, we assume that presentation and discussion of examples as well as case studies can help to spark usage and development of comparative visualizations in serious game analytics and thus contribute to the advancement of the field. In that sense, serious game analytics may also benefit by adapting approaches developed for entertainment games (see, Wallner and Kriglstein 2013 for an overview) or by drawing inspiration from fields such as sports visualization where comparison of individual players is also an important factor (e.g., Perin, Vuillemot, & Fekete, 2013; Pileggi, Stolper, Boyle, & Stasko, 2012).

Acknowledgments We would like to thank Helmut Hlavacs from the Entertainment Computing group of the University of Vienna for his permission to use the telemetry data from *Internet Hero* for examples in this chapter. *Internet Hero* was funded by netidee (www.netidee.at, project number 326).

References

Andersen, E., Liu, Y.-E., Apter, E., Boucher-Genesse, F., & Popović, Z. (2010). Gameplay analysis through state projection. In *Proceedings of the 5th International Conference on the Foundations of Digital Games* (pp. 1–8). doi:10.1145/1822348.1822349.

Andrews, K., Wohlfahrt, M., & Wurzinger, G. (2009). Visual graph comparison. In *Proceedings of the 13th International Conference Information Visualisation* (pp. 62–67). doi:10.1109/IV.2009.108.

Anolli, L., & Confalonieri, L. (2011). Learning, dynamic assessment and serious games. In A. Méndez-Vilas (Ed.), *Education in a technological world: Communicating current and emerging research and technological efforts* (pp. 279–287). Badajoz, Spain: Formatex.

Archambault, D., Purchase, H. C., & Pinaud, B. (2011). Animation, small multiples, and the effect of mental map preservation in dynamic graphs. *IEEE Transactions on Visualization and Computer Graphics, 17*(4), 539–552. doi:10.1109/TVCG.2010.78.

Arthur, W., Jr., Strong, M. H., Jordan, J. A., Williamson, J. E., Shebilske, W. L., & Regian, J. W. (1995). Visual attention: Individual differences in training and predicting complex task performance. *Acta Psychologica, 88*(1), 3–23. doi:10.1016/0001-6918(94)E0055-K.

Beck, F., Burch, M., Diehl, S., & Weiskopf, D. (2014). The state of the art in visualizing dynamic graphs. In *EuroVis—STARs* (pp. 83–103). doi:10.2312/eurovisstar.20141174.

Beck, F., Burch, M., & Weiskopf, D. (2013). Visual comparison of time-varying athletes' performance. In *Proceedings of the 1st Workshop on Sports Data Visualization*.

Becker, K., & Parker, J. R. (2011). *The guide to computer simulations and games*. Indianapolis, IN: Wiley.

Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., & Berta, R. (2013). Assessment in and of serious games: An overview. *Advances in Human-Computer Interaction, 2013*. doi:10.1155/2013/136864.

Boyandin, I., Bertini, E., & Lalanne, D. (2012). A qualitative study on the exploration of temporal changes in flow maps with animation and small-multiples. *Computer Graphics Forum, 31*(3pt2), 1005–1014. doi:10.1111/j.1467-8659.2012.03093.x.

Brandes, U., & Corman, S. R. (2003). Visual unrolling of network evolution and the analysis of dynamic discourse. *Information Visualization, 2*(1), 40–50. doi:10.1057/palgrave.ivs.9500037.

Brandes, U., Dwyer, T., & Schreiber, F. (2004). Visualizing related metabolic pathways in two and a half dimensions. In G. Liotta (Ed.), *Graph drawing* (pp. 111–122). Berlin: Springer.

Buendía-García, F., García-Martínez, S., Navarrete-Ibañez, E. M., & Jesús, M. (2013). Designing serious games for getting transferable skills in training settings. *Interaction Design and Architecture(s), 19*, 47–62.

Butler, E., & Banerjee, R. (2014). *Visualizing progressions for education and game design*. Retrieved from http://cse512-14w.github.io/fp-edbutler-piscean/final/paper-edbutler-piscean.pdf

Dawson, S. (2010). "Seeing" the learning community: An exploration of the development of a resource for monitoring online student networking. *British Journal of Educational Technology, 41*(5), 736–752. doi:10.1111/j.1467-8535.2009.00970.x.

de Freitas, S., & Jarvis, S. (2006). A framework for developing serious games to meet learner needs. In *Proceedings of the Interservice/Industry Training, Simulation & Education Conference*.

de Freitas, S., & Oliver, M. (2006). How can exploratory learning with games and simulations within the curriculum be most effectively evaluated? *Computers & Education, 46*(3), 249–264. doi:10.1016/j.compedu.2005.11.007.

Desmarais, M. C., & Lemieux, F. (2013). Clustering and visualizing study state sequences. In *Proceedings of 6th International Conference on Educational Data Mining* (pp. 224–227).

Drachen, A., & Canossa, A. (2011). Evaluating motion: Spatial user behaviour in virtual environments. *International Journal of Arts and Technology, 4*(3), 294–314. doi:10.1504/IJART.2011.041483.

Duval, E. (2011). Attention please!: Learning analytics for visualization and recommendation. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (pp. 9–17). doi:10.1145/2090116.2090118.

Dwyer, T., Hong, S.-H., Koschützki, D., Schreiber, F., & Xu, K. (2006). Visual analysis of network centralities. In *Proceedings of the Asia-Pacific Symposium on Information Visualisation* (pp. 189–197).

Erten, C., Kobourov, S. G., Le, V., & Navabi, A. (2003). Simultaneous graph drawing: Layout algorithms and visualization schemes. In *Proceedings of the 11th Symposium on Graph Drawing* (pp. 437–449).

Fardinpour, A., Reiners, T., & Dreher, H. (2013). Action-based learning assessment method (ALAM) in virtual training environments. In *Proceedings of the 30th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education* (pp. 267–277).

Federico, P., Aigner, W., Miksch, S., Windhager, F., & Zenk, L. (2011). A visual analytics approach to dynamic social networks. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies* (pp. 47:1–47:8). doi:10.1145/2024288.2024344.

Ferster, B., & Shneiderman, B. (2012). *Interactive visualization: Insight through inquiry*. Cambridge, MA: MIT Press.

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends, 59*(1), 64–71. doi:10.1007/s11528-014-0822-x.

Gelderblom, H., & Kotzé, P. (2009). Ten design lessons from the literature on child development and children's use of technology. In *Proceedings of the 8th International Conference on Interaction Design and Children* (pp. 52–60). doi:10.1145/1551788.1551798.

Gleicher, M., Albers, D., Walker, R., Jusufi, I., Hansen, C. D., & Roberts, J. C. (2011). Visual comparison for information visualization. *Information Visualization, 10*(4), 289–309. doi:10.1177/1473871611416549.

Govaerts, S., Verbert, K., Duval, E., & Pardo, A. (2012). The student activity meter for awareness and self-reflection. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems* (pp. 869–884). doi:10.1145/2212776.2212860.

Govaerts, S., Verbert, K., Klerkx, J., & Duval, E. (2010). Visualizing activities for self-reflection and awareness. In X. Luo, M. Spaniol, L. Wang, Q. Li, W. Nejdl, & W. Zhang (Eds.), *Advances in web-based learning—ICWL 2010* (pp. 91–100). Berlin: Springer.

Graham, M., & Kennedy, J. (2010). A survey of multiple tree visualisation. *Information Visualization, 9*(4), 235–252. doi:10.1057/ivs.2009.29.

Griffin, A. L., MacEachren, A. M., Hardisty, F., Steiner, E., & Li, B. (2006). A comparison of animated maps with static small-multiple maps for visually identifying space-time clusters. *Annals of the Association of American Geographers, 96*(4), 740–753. doi:10.1111/j.1467-8306.2006.00514.x.

Guerra-Gómez, J. A., Buck-Coleman, A., Pack, M. L., Plaisant, C., & Shneiderman, B. (2013). TreeVersity: Interactive visualizations for comparing hierarchical data sets. *Transportation Research Record: Journal of the Transportation Research Board, 2392*, 48–58. doi:10.3141/2392-06.

Hartmann, T., & Klimmt, C. (2006). Gender and computer games: Exploring females' dislikes. *Journal of Computer-Mediated Communication, 11*(4), 910–931. doi:10.1111/j.1083-6101.2006.00301.x.

Heeter, C., Lee, Y.-H., Magerko, B., & Medler, B. (2011). Impacts of forced serious game play on vulnerable subgroups. *International Journal of Gaming and Computer-Mediated Simulations, 3*(3), 34–53. doi:10.4018/jgcms.2011070103.

Holten, D., & van Wijk, J. J. (2008). Visual comparison of hierarchically organized data. In *Proceedings of the 10th Joint Eurographics/IEEE—VGTC Conference on Visualization* (pp. 759–766). doi:10.1111/j.1467-8659.2008.01205.x.

Houghton, S. (2011). *Balance and flow maps*. Retrieved April, 2015, from http://www.gamasutra.com/view/news/125213/Opinion_Balance_and_Flow_Maps.php.

Inselberg, A. (1985). The plane with parallel coordinates. *The Visual Computer, 1*(2), 69–91. doi:10.1007/BF01898350.

Javed, W., & Elmqvist, N. (2012). Exploring the design space of composite visualization. In *Proceedings of the IEEE Pacific Visualization Symposium* (pp. 1–8). doi:10.1109/PacificVis.2012.6183556.

Kayali, F., Wallner, G., Kriglstein, S., Bauer, G., Martinek, D., Hlavacs, H., et al. (2014). A case study of a learning game about the Internet. In S. Göbel & J. Wiemeyer (Eds.), *Games for training, education, health and sports* (pp. 47–58). Cham, Switzerland: Springer.

Ketelhut, D. J. (2007). The impact of student self-efficacy on scientific inquiry skills: An exploratory investigation in River City, a multi-user virtual environment. *Journal of Science Education and Technology, 16*(1), 99–111. doi:10.1007/s10956-006-9038-y.

Kiili, K., Ketamo, H., & Kickmeier-Rust, M. D. (2014). Evaluating the usefulness of eye tracking in game-based learning. *International Journal of Serious Games, 1*(2), 51–65.

Kirk, A. (2012). *Data visualization: A successful design process*. Birmingham, England: Packt Publishing.

Klerkx, J., Verbert, K., & Duval, E. (2014). Enhancing learning with visualization techniques. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (pp. 791–807). New York: Springer.

Kriglstein, S., Pohl, M., & Smuc, M. (2014). Pep up your time machine: Recommendations for the design of information visualizations of time-dependent data. In W. Huang (Ed.), *Handbook of human centric visualization* (pp. 203–225). New York: Springer.

Kriglstein, S., Wallner, G., & Rinderle-Ma, S. (2013). A visualization approach for difference analysis of process models and instance traffic. In *Proceedings of the 11th International*

Conference on Business Process Management (pp. 219–226). doi:10.1007/978-3-642-40176-3_18.

Kriz, W. C., & Hense, J. U. (2006). Theory-oriented evaluation for the design of and research in gaming and simulation. *Simulation & Gaming, 37*(2), 268–283. doi:10.1177/1046878106287950.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage.

Linek, S. B., Öttl, G., & Albert, D. (2010). Non-invasive data tracking in educational games: Combination of logfiles and natural language processing. In *Proceeding of the International Technology, Education and Development Conference* (pp. 2977–2988).

Liu, Y.-E., Andersen, E., Snider, R., Cooper, S., & Popović, Z. (2011). Feature-based projections for effective playtrace analysis. In *Proceedings of the 6th International Conference on Foundations of Digital Games* (pp. 69–76). doi:10.1145/2159365.2159375.

Loh, C. S. (2012). Information Trails: In-process assessment of game-based learning. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning* (pp. 123–144). New York: Springer.

Loh, C. S., & Sheng, Y. (2013). Measuring the (dis-)similarity between expert and novice behaviors as serious games analytics. *Education and Information Technologies, 20*, 5–19. doi:10.1007/s10639-013-9263-y.

Magerko, B., Heeter, C., & Medler, B. (2010). Different strokes for different folks: Tapping into the hidden potential of serious games. In R. Van Eck (Ed.), *Gaming and cognition: Theories and practice from the learning sciences* (pp. 255–280). Hershey, PA: IGI Global.

Mehigan, T. J., Barry, M., Kehoe, A., & Pitt, I. (2011). Using eye tracking technology to identify visual and verbal learners. In *Proceedings of the IEEE International Conference on Multimedia and Expo* (pp. 1–6). doi:10.1109/ICME.2011.6012036.

Miller, J. R. (2007). Attribute blocks: Visualizing multiple continuously defined attributes. *IEEE Computer Graphics and Applications, 27*(3), 57–69. doi:10.1109/MCG.2007.54.

Minović, M., & Milovanović, M. (2013). Real-time learning analytics in educational games. In *Proceedings of the 1st International Conference on Technological Ecosystem for Enhancing Multiculturality* (pp. 245–251). doi:10.1145/2536536.2536574.

Mislevy, R. J., & Riconscente, M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology* (PADI Technical Report No. 9). Menlo Park, CA: SRI International.

Moreno-Ger, P., Torrente, J., Hsieh, Y. G., & Lester, W. T. (2012). Usability testing for serious games: Making informed design decisions with user data. *Advances in Human-Computer Interaction, 2012*. doi:10.1155/2012/369637.

Munzner, T., Guimbretière, F., Tasiran, S., Zhang, L., & Zhou, Y. (2003). TreeJuxtaposer: Scalable tree comparison using focus+context with guaranteed visibility. *ACM Transactions on Graphics, 22*(3), 453–462. doi:10.1145/1201775.882291.

O'Rourke, E., Butler, E., Liu, Y.-E., Ballweber, C., & Popović, Z. (2013). The effects of age on player behavior in educational games. In *Proceedings of the 8th International Conference on the Foundations of Digital Games* (pp. 158–165).

Olsen, T., Procci, K., & Bowers, C. (2011). Serious games usability testing: How to ensure proper usability, playability, and effectiveness. In A. Marcus (Ed.), *Design, user experience, and usability. Theory, methods, tools and practice* (pp. 625–634). Berlin: Springer.

Peirce, K., & Edwards, E. D. (1988). Children's construction of fantasy stories: Gender differences in conflict resolution strategies. *Sex Roles, 18*(7–8), 393–404. doi:10.1007/BF00288391.

Perin, C., Vuillemot, R., & Fekete, J.-D. (2013). SoccerStories: A kick-off for visual soccer analysis. *IEEE Transactions on Visualization and Computer Graphics, 19*(12), 2506–2515. doi:10.1109/TVCG.2013.192.

Pileggi, H., Stolper, C. D., Boyle, J. M., & Stasko, J. T. (2012). SnapShot: Visualization to propel ice hockey analytics. *IEEE Transactions on Visualization and Computer Graphics, 18*(12), 2819–2828. doi:10.1109/TVCG.2012.263.

Piringer, H., Pajer, S., Berger, W., & Teichmann, H. (2012). Comparative visual analysis of 2D function ensembles. *Computer Graphics Forum, 31*(3), 1195–1204. doi:10.1111/j.1467-8659.2012.03112.x.

Plass, J. L., Homer, B. D., Kinzer, C. K., Chang, Y. K., Frye, J., Kaczetow, W., et al. (2013). Metrics in simulations and games for learning. In M. S. El-Nasr, A. Drachen, & A. Canossa (Eds.), *Game analytics* (pp. 697–729). London: Springer.

Ritsos, P. D., & Roberts, J. C. (2014). Towards more visual analytics in learning analytics. In *Proceedings of the 5th EuroVis Workshop on Visual Analytics* (pp. 61–65). doi:10.2312/eurova.20141147.

Roberts, J. C. (2005). Exploratory visualization with multiple linked views. In J. Dykes, A. M. MacEachren, & M.-J. Kraak (Eds.), *Exploring geovisualization* (pp. 159–180). Oxford, England: Elsevier.

Robertson, G., Fernandez, R., Fisher, D., Lee, B., & Stasko, J. (2008). Effectiveness of animation in trend visualization. *IEEE Transactions on Visualization and Computer Graphics, 14*(6), 1325–1332. doi:10.1109/TVCG.2008.125.

Rowe, J. P., Shores, L. R., Mott, B. W., & Lester, J. C. (2010). Individual differences in gameplay and learning: A narrative-centered learning perspective. In *Proceedings of the 5th International Conference on the Foundations of Digital Games* (pp. 171–178). doi:10.1145/1822348.1822371.

Scarlatos, L. L., & Scarlatos, T. (2010). Visualizations for the assessment of learning in computer games. In *Proceedings of the 7th International Conference & Expo on Emerging Technologies for a Smarter World*.

Serrano-Laguna, Á., Torrente, J., Moreno-Ger, P., & Fernández-Manjón, B. (2012). Tracing a little for big improvements: Application of learning analytics and videogames for student assessment. *Procedia Computer Science, 15*, 203–209. doi:10.1016/j.procs.2012.10.072.

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review, 14*(1), 47–69. doi:10.1023/A:1013180410169.

Shute, V., Ventura, M., Bauer, M., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfield, M. J. Cody, & P. Vorderer (Eds.), *The social science of serious games: Theories and applications* (pp. 295–321). New York: Routledge.

Steiner, C. M., Kickmeier-Rust, M. D., & Albert, D. (2009). Little big difference: Gender aspects and gender-based adaptation in educational games. In M. Chang, R. Kuo, K. Kinshuk, G.-D. Chen, & M. Hirose (Eds.), *Learning by playing. Game-based education system design and development* (pp. 150–161). Berlin: Springer.

Stewart, C. A., Hart, D., Berry, D. K., Olsen, G. J., Wernert, E. A., & Fischer, W. (2001). Parallel implementation and performance of fastDNAml: A program for maximum likelihood phylogenetic inference. In *Proceedings of the ACM/IEEE Conference on Supercomputing*. doi:10.1145/582034.582054.

Tominski, C., Schulze-Wollgast, P., & Schumann, H. (2005). 3D information visualization for time dependent data on maps. In *Proceedings of the nineth International Conference on Information Visualisation* (pp. 175–181). doi:10.1109/IV.2005.3.

Tufte, E. R. (1990). *Envisioning information*. Cheshire, UK: Graphics Press.

Urness, T., Interrante, V., Marusic, I., Longmire, E., & Ganapathisubramani, B. (2003). Effectively visualizing multi-valued flow data using color and texture. In *Proceedings of the IEEE Visualization* (pp. 115–121). doi:10.1109/VISUAL.2003.1250362.

van Wijk, J. J. (2005). The value of visualization. In *Proceedings of the IEEE Visualization* (pp. 79–86). doi:10.1109/VISUAL.2005.1532781.

Verbert, K., Govaerts, S., Duval, E., Santos, J. L., Assche, F. V., Parra, G., et al. (2014). Learning dashboards: An overview and future research opportunities. *Personal and Ubiquitous Computing, 18*(6), 1499–1514. doi:10.1007/s00779-013-0751-2.

Wallner, G. (2013). Play-Graph: A methodology and visualization approach for the analysis of gameplay data. In *Proceedings of the 8th International Conference on the Foundations of Digital Games* (pp. 253–260).

Wallner, G., & Kriglstein, S. (2012a). DOGeometry: Teaching geometry through play. In *Proceedings of the 4th International Conference on Fun and Games* (pp. 11–18). doi:10.1145/2367616.2367618.

Wallner, G., & Kriglstein, S. (2012b). A spatiotemporal visualization approach for the analysis of gameplay data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1115–1124). doi:10.1145/2207676.2208558.

Wallner, G., & Kriglstein, S. (2013). Visualization-based analysis of gameplay data—A review of literature. *Entertainment Computing, 4*(3), 143–155. doi:10.1016/j.entcom.2013.02.002.

Ward, M. O. (2002). A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization, 1*(3/4), 194–210. doi:10.1057/palgrave.ivs.9500025.

Warren, S., Jones, G., & Lin, L. (2011). Usability and play testing. In L. Annetta & S. C. Bronack (Eds.), *Serious educational game assessment* (pp. 131–146). Rotterdam, The Netherlands: Sense.

Westera, W., Nadolski, R., Hummel, H. (2014). Serious gaming analytics: What students' log files tell us about gaming and learning. *International Journal of Serious Games*, 1(2). doi:10.17083/ijsg.v1i2.9.

Wouters, P., van der Spek, E. D., & Van Oostendorp, H. (2009). Current practices in serious game research: A review from a learning outcomes perspective. In T. M. Connolly, M. Stansfield, & L. Boyle (Eds.), *Games-based learning advancements for multisensory human computer interfaces: Techniques and effective practices*. Hershey, PA: IGI Global.

Yau, N. (2011). *Visualize this: The FlowingData guide to design, visualization, and statistics*. Indianapolis, IN: Wiley.

Chapter 8
Examining Through Visualization What Tools Learners Access as They Play a Serious Game for Middle School Science

Min Liu, Jina Kang, Jaejin Lee, Elena Winzeler, and Sa Liu

Abstract This study intends to use data visualization to examine learners' behaviors in a 3D immersive serious game for middle school science to understand how the players interact with various features to solve the central problem. The analysis combined game log data with measures of in-game performance and learners' goal orientations. The findings indicated students in the high performance and mastery-oriented groups tended to use the tools more appropriately relative to the stage they were at in the problem-solving process, and more productively than students in low performance groups. The use of data visualization with log data in combination with more traditional measures shows visualization as a promising technique in analytics with multiple data sets that can facilitate the interpretation of the relationships among data points at no cost to the complexity of the data. Design implications and future applications of serious games analytics and data visualization to the serious game are discussed.

Keywords Serious games • Problem-based learning • Middle school science • Learner behaviors • Goal orientation

M. Liu (✉) • J. Kang • S. Liu
The University of Texas at Austin, 1912 Speedway Stop D5700,
Austin, TX 78712-1293, USA
e-mail: mliu@austin.utexas.edu; jina.kang@austin.utexas.edu; liusa@utexas.edu

J. Lee
The University of Texas at Austin, 2501 Lake Austin Blvd A205, Austin, TX 78703, USA
e-mail: jaejinlee@utexas.edu

E. Winzeler
The University of Texas at Austin, 2207 Wickersham Lane,
Apartment 1020, Austin, TX 78741, USA
e-mail: emwinzeler@utexas.edu

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based
Learning, DOI 10.1007/978-3-319-05834-4_8

1 Introduction

The popularity of playing games has been increasing. According to a report by the Pew Research Center, digital game industry "takes in about $93 billion a year" (Holcomb & Mitchell, 2014), and playing games continue to be an important of form of how people, young and old, spend their leisure time. A Kaiser Family Foundation report stated, "In a typical day, 8- to 18-year-olds spend an average of 1:13 playing video games on any of several platforms" (Rideout, Foehr, & Roberts, 2010, p. 25). Therefore, it behooves educators to investigate how to employ techniques used in digital games to design digital learning environments.

The goal of this study was to examine learners' behaviors in a 3D immersive serious game environment designed for middle school science to understand how the play-learners interact with various features of the environment to solve the central problem. We used data visualization as a way to represent patterns of learners' behaviors. By applying data visualization techniques to serious games analytics, we hope to acquire insights on how serious game environments should be designed to facilitate learning.

2 Relevant Literature

2.1 Definition and Examples

Serious Games (SGs) are a type of games that include simulated events or virtual processes designed for the purpose of real-world problem-solving (Djaouti, Alvarez, Jessel, & Rampnoux, 2011; Rieber, 1996; Sawyer & Smith, 2008). Abt stated that SGs have "an explicit and carefully thought-out educational purpose and are not intended to be played primarily for amusement" (1970, p. 9). According to the Serious Games Initiative (www.seriousgames.org), SGs leverage game mechanics for training through exer-games, management games, and simulations. Therefore, although serious games can be fun and entertaining, their main purposes are to train, educate, or change users' attitudes in the real-world situations. The applications for SGs are diverse. The term "serious" denotes an alteration of the context of gaming from fun and entertainment to engagement, efficiency, and pedagogical effectiveness for specific purposes such as training and performance enhancement (Djaouti et al., 2011). In this study, we were interested in using SGs to teach science concepts and problem-solving skills and create a fun learning experience for play-learners.

Many commercial games have been integrated into classroom settings for instructional purposes, such as *SimCity* (Tanes & Cemalcilar, 2010), *Civilization* (Squire, 2004), and *Minecraft* (List & Bryant, 2014). Some educational researchers also design and develop SGs themselves. For example, "*Outbreak @ The Institute*" is a role-play science game in which play-learners take on the roles of doctors, medical technicians, and public health experts to discover the cause of and develop a cure for a disease outbreak across a university campus (Rosenbaum, Klopfer, & Perry, 2007). Play-learners can interact with virtual characters and employ virtual diagnostic tests and medicines. In another science SG, *Mad City Mystery*, play-learners develop explanations of scientific phenomena in an inquiry-based learning environment (Squire & Jan, 2007).

2.2 Research Trends in Serious Games

Research on serious games typically focuses on their effects on learners' engagement or effectiveness using traditional intervention studies with experimental designs or qualitative methods. The emergence of serious games analytics (SEGA) makes it possible to investigate beyond traditional research methodologies and focus on the learning processes of individuals as expressed through patterns of in-game behavior and accomplishments (Djaouti et al., 2011; Johnson et al., 2013; Scarlatos & Scarlatos, 2010).

The purpose of using analytics is to illuminate the process of performance improvement via in-game instructional resources (van Barneveld, Arnold, & Campbell, 2012). Studies in the field of SEGA for performance assessment primarily use game logs—unobtrusively saved records—on user activities with chronological and spatial tracking data (Johnson, Adams Becker, Estrada, & Freeman, 2014; Liu, Horton, Kang, Kimmons, & Lee, 2013; Macfadyen & Dawson, 2010; Wallner & Kriglstein, 2013). SEGA, therefore, is inherently an interdisciplinary field that links gaming data and student responses to statistics, computer science, data mining, and visualization (Baker & Yacef, 2009; Romero, Ventura, & García, 2008). The learning models and usage patterns are utilized to predict student knowledge-building trajectories through the categorization of levels of performance, engagement, and resource-processing sequences (U.S. Department of Education, Office of Educational Technology, 2012). Researchers are interested in using analytics to gain insights that can enable the design and validation of pedagogical scaffolding support in online learning environments.

There have been a number of research efforts to produce standardized analysis procedures, from planning the capture of learner activities to analyzing the data to finally visualizing the analysis, so that SEGA techniques can contribute to the field of SG as a solid methodology of learner evaluation (Loh, 2008, 2011; Romero & Ventura, 2010, 2013). Romero's data mining model (2013) provides SG researchers seven steps to follow to conduct a SEGA study with a clear hypothesis: hypothesis formation, raw data gathering, preprocessing, data modification, data mining, finding models and patterns, and interpretation/evaluation. Serrano, Marchiori, del Blanco, Torrente, and Fernández-Manjón (2012) also provided a similar framework containing seven elements: data selection, data capture, aggregation and report, assessment, knowledge creation, knowledge refinement, and knowledge sharing.

In studies involving serious games analytics (Linek, Marte, & Albert, 2008; Loh, 2011; Reese, Tabachnick, & Kosko, 2013; Scarlatos & Scarlatos, 2010), the learning processes of individual students have been tracked using diverse techniques in order to support the personalization of instruction. In these examples, game logs have been regarded as an important metric in examining topics ranging from knowledge domains to tool use (Dede, 2014; Wallner & Kriglstein, 2013).

2.3 Issues in SEGA Evaluation

The efficacy of SGs has often been evaluated using traditional tests (e.g., standardized tests or surveys), which may not sufficiently measure higher learning objectives such as application, analysis, or synthesis (Scarlatos & Scarlatos, 2010). Since most of these tests are collected before or after SG play, the obtained data can merely represent prospective or retrospective views (Linek, Öttl, & Albert, 2010). They cannot be used to assess how learners achieved learning objectives within the game environment or the decision-making processes undertaken to solve a given problem. In addition, Loh (2008) warned of the limitations of computer-based tests since these cannot be used to evaluate opinions of learners, but only to assess the accuracy of their choices. Other methods such as observations or interviews have also been used for evaluating and understanding gameplay (Garzotto, 2007; Sweetser & Wyeth, 2005). Yet, researchers assert that such methods are inefficient in terms of time and lose clarity with large numbers of learners (e.g., Andersen, Liu, Apter, Boucher-Genesse, & Popović, 2010; Drachen & Canossa, 2009).

These challenges highlight the need to use log data to understand the play-learners' behaviors within the environment and examine log data in connection to learners' performance. Game-generated data logs contain records of human behaviors during learning, which can include any interaction between a learner and a game such as mouse click or keystroke. Reese et al. (2013) emphasized that learning objectives align with game objectives; therefore, a player's idiosyncratic trajectory towards the game goal can reveal the dynamics of the learning process. To understand how a learner achieves a learning goal requires the discovery and analysis of patterns of play-learner behaviors (Drachen & Canossa, 2009), and log data can provide insights into play-learner behavior in context (Scarlatos & Scarlatos, 2010). The emerging technology of data visualization allows researchers to present and examine data visually in order to discover patterns relating to what learners are doing in an SG context (Dixit & Youngblood, 2008; Milam & El Nasr, 2010; Scarlatos & Scarlatos, 2010). Therefore, using visualization in combination with more traditional measures should provide more targeted and nuanced information to gain a holistic view of play-learners' behaviors (Linek et al., 2008).

2.4 Background of Research

We have conducted several studies to examine students' usage patterns through statistical procedures such as descriptive analysis and cluster analysis with the same serious game used in this study, *Alien Rescue*. The study by Liu and Bera (2005) applied cluster analysis to sixth-graders' log data to examine what tools were used and at what stages of their problem-solving process. The results showed that tools supporting cognitive processing and tools sharing cognitive load played a more central role early in the problem-solving process whereas tools supporting cognitive activities that would be out of students' reach otherwise and tools supporting hypothesis generation and testing were used more in the later stages of problem-solving. The findings also indicated that the students increasingly used multiple tools in the later stages of the problem-solving process. The various tools appeared to enable students to coordinate multiple cognitive skills in a seamless way and, therefore, facilitated their information processing. Results also suggested that students with higher performance scores seemed to exercise more productive use of the tools than students with lower performance scores.

In a follow-up study in our investigation (Liu et al., 2009), log data were matched with surveys from a group of college students who played *Alien Rescue* in a laboratory setting. A researcher observed each student's activity in the environment and stimulated recall interviews elicited information on students' cognitive processes at specific points in the problem-solving process. Quantitative data–log files–and qualitative data together revealed deliberate and careful use of tools by the students. Students simultaneously used multiple tools while engaged in integrating and evaluating information and different tools predominated during each problem-solving stage. This finding suggested that different types of tools were needed and used by the college students in this study, as they were by sixth graders in the previous research (Liu & Bera, 2005; Liu, Bera, Corliss, Svinicki, & Beth, 2004), but the results did not show evidence that students with higher performance used the tools more consistently or actively than the other groups as in the previous research (Liu et al., 2004; Liu & Bera, 2005).

Given these preliminary findings and especially the technological advancements in our field, the purpose of this study was to further this research line by using data visualization techniques to examine the patterns of how sixth graders played the SG and identify factors contributing to individual variations.

3 Research Questions and Research Context

3.1 Research Questions

The following research questions guided this study:

- How do play-learners access different tools built into the game?
- How do play-learners with different goal orientations access the tools?
- How do play-learners with different performance scores access the tools?

3.2 Description of the Serious Game Environment

The serious game environment under investigation is called *Alien Rescue* (AR, alienrescue.edb.utexas.edu; Liu et al., 2013). AR is designed and developed by a research and development team in the Learning Technologies Program at the University of

Texas at Austin. AR aspires to teach science and complex problem-solving skills to students in fun and interactive ways. Its development is guided by a design-based research framework which aims to generate and refine theories by evaluating iterative enhancements to an instructional innovation within authentic settings (Brown, 1992; Cobb, Confrey, Lehrer, & Schauble, 2003).

AR incorporates problem-based learning pedagogy into a 3D virtual environment to engage middle-school students in solving complex and meaningful scientific problems. Students take on the role of young scientists in a rescue operation to save a group of six distressed alien species displaced from a distant galaxy due to the destruction of their home worlds. The young scientists are challenged to find the most suitable relocation homes for these aliens in our solar system. Each alien species is unique in its characteristics and needs. Upon starting the program, students are not given explicit instructions on how to proceed. They must explore the available tools, discover their capabilities, and develop their own strategies for how and when to effectively use them. Learning occurs as a result of solving a complex, ill-structured problem; there is not one single correct solution, and play-learners must present evidence and justify the rationale for their solutions.

This real-world process of scientific inquiry is transformed into a playful experience and delivered through an immersive, discovery-based, and sensory-rich approach, in line with Salen and Zimmerman's (2004) definition of play as "free movement within a more rigid structure" (p. 304). The element of fantasy evokes uncertainty, mystery, and curiosity, while the quest-based narrative situates students in the role of experts with an urgent mission, motivating them to acquire competence in the language, concepts, tools, and processes of space science in order to succeed. Furthermore, the students must exercise high-level cognitive and metacognitive skills such as goal setting, hypothesis generation, problem-solving, self-regulation, evaluation of various possible solutions, and the effective presentation of evidence. Thus, AR provides a learning experience with real-world authenticity that also accomplishes essential curricular goals, all within an engaging science fiction fantasy context.

3.3 Cognitive Tools and Their Corresponding Conceptual Categories

To assist students' problem-solving, a set of tools are provided. These cognitive tools in the AR environment align with Lajoie's (1993) four conceptual categories (see Table 8.1): tools that (a) share cognitive load, (b) support cognitive and metacognitive processes, (c) support cognitive activities that would otherwise be out of reach, and (d) support hypothesis generation and testing. Table 8.1 outlines the tools according to Lajoie's categorization (1993).

Of the tools that share cognitive load, the Alien Database (see Fig. 8.1c) and Solar System Database are the most central to the problem-solving process.

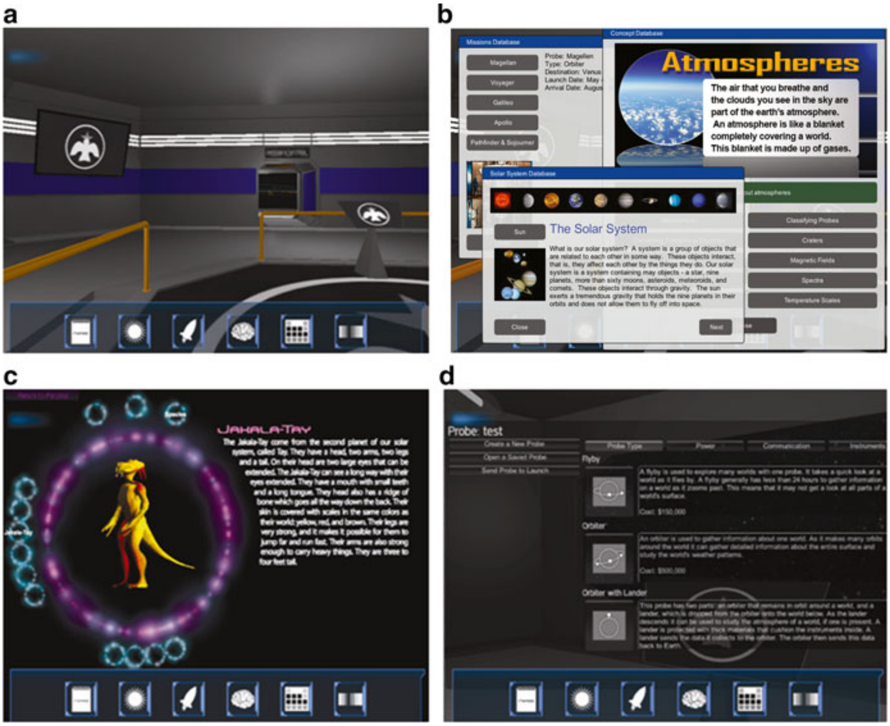
Table 8.1 Descriptions of cognitive tools provided in AR

Tool categories		Tool functions
Tools sharing cognitive load	Alien Database	Presents textual descriptions and 3D visuals of the aliens' home solar system and journey to Earth, as well as the characteristics and needs of each species
	Solar System Database	Provides information on the planets and selected moons in our solar system under consideration as habitats. Intentionally incomplete data ensures the need to generate and test hypotheses
	Missions Database	Presents information on the mission, technology, and findings of historical NASA probe launches
	Concepts Database	Provides interactive and highly visual supplemental instruction on selected scientific concepts presented elsewhere in the environment
	Spectra	Helps students to interpret spectral data encountered in the environment
	Periodic Table	Provides an interactive periodic table of the elements for reference
Tools supporting cognitive process	Notebook	Provides a place for students to record, summarize, and organize data as they engage in solving the central problem
Tools supporting otherwise out-of-reach activities	Probe Design Center	Allows students to design and build probes to send to gather data on worlds in our solar system
	Probe Launch Center	Allows students to review built probes and make launch decisions in consideration of their remaining budget
Tools supporting hypothesis testing	Mission Control Center	Displays data from launched probes
	Message Tool	Allows students to read messages from the Aliens and from the Interstellar Relocation Commission Director. Provides the Solution Form, which allows students to submit their habitat relocation recommendations and rationales for review by teachers

Together all of these tools provide students with a wealth of information to assist them in solving the problem (see Fig. 8.1b). They share cognitive load by reducing the need to memorize facts; the information is always available to the student. Thus, these tools shift the focus of learning from remembering to understanding, applying, and analyzing.

The Notebook supports cognitive processes as students work to solve the problem. As the physical space within the serious game environment where information from disparate sources is integrated, the Notebook facilitates the students' synthesis of knowledge. On a metacognitive level, the Notebook provides a way for students to monitor their own progress towards solving the central problem.

The tools that support cognitive activities that would otherwise be out of reach are the Probe Design Center (see Fig. 8.1d) and Probe Launch Center. Designing and



Probe: Magellen
Type: Orbiter
Destination: Venus
Launch Date: May
Arrival Date: Augus

The air that you breathe and the clouds you see in the sky are part of the earth's atmosphere. An atmosphere is like a blanket completely covering a world. This blanket is made up of gases.

What is our solar system? A system is a group of objects that are related to each other in some way. These objects interact, that is, they affect each other by the things they do. Our solar system is a system containing may objects - a star, nine planets, more than sixty moons, asteroids, meteoroids, and comets. These objects interact through gravity. The sun exerts a tremendous gravity that holds the nine planets in their orbits and does not allow them to fly off into space.

JAKALA-TAY

The Jakala-Tay come from the second planet of our solar system, called Tay. They have a head, two arms, two legs and a tail. On their head are two large eyes that can be extended. The Jakala-Tay can see a long way with their eyes extended. They have a mouth with small teeth and a long tongue. They head also has a ridge of bone which goes all the way down the back. Their skin is covered with scales in the same colors as their world: yellow, red, and brown. Their legs are very strong, and it makes it possible for them to jump far and run fast. Their arms are also strong enough to carry heavy things. They are three to four feet tall.

A flyby is used to explore many worlds with one probe. It takes a quick look at a world as it flies by. A flyby generally has less than 24 hours to gather information on a world as it zooms past. This means that it may not get a look at all parts of a world's surface.

Cost: $150,000

An orbiter is used to gather information about one world. As it makes many orbits around the world it can gather detailed information about the entire surface and study the world's weather patterns.

Cost: $500,000

This probe has two parts: an orbiter that remains in orbit around a world, and a lander, which is dropped from the orbiter into the world below. As the lander descends it can be used to study the atmosphere of a world, if one is present. A lander is protected with thick materials that cushion the instruments inside. A lander sends the data it collects to the orbiter. The orbiter then sends this data back to Earth.

Fig. 8.1 Screenshots of various cognitive tools in AR that support the problem-solving process. (**a**) A view of the space station with tools panel overlay. (**b**) Students can open several tools, such as the Concepts, Solar System and Missions Databases, at a time. (**c**) The Alien Database contains 3D visuals and descriptions of the aliens, their former homes, and their journey to Earth. (**d**) Students can design, launch, and view data collected from their own simulated probes to test their hypotheses

launching probes are activities that most students will only ever experience in a virtual environment such as AR. These tools not only provide an exciting and novel experience to the student, but also preserve the authenticity of the scientific inquiry process and the consequentiality of the serious game environment, since students' probe design decisions directly impact the data available to them (Barab, Gresalfi, & Ingram-Goble, 2010, p. 526).

The Mission Control Center and Message Tool support hypothesis testing. Since the information provided in the research databases is intentionally incomplete, only the data from deployed probes viewed in the Mission Control Center allow students to draw the inferences necessary to generate their own solutions to the central problem. The Solution Form housed in the Message Tool provides students with a mechanism to develop their hypotheses into well-formed rationales to be evaluated by their teacher.

These tools are accessed via a two-layer interface (see Fig. 8.1a). The first layer is the virtual space station itself, which consists of five rooms, each containing an instrument for students to use. The second layer of the interface consists of a collection of persistent tools available at the bottom of the screen. It is possible to have several of these overlay tools open at once, though a student can visit only one room in the navigation layer at a time.

AR is designed for approximately 3 weeks of 50-min class sessions as a sixth-grade science curriculum unit. Depending on specific needs and classroom situations, teachers can adapt and adjust the days accordingly. The open-ended, ill-structured framework of AR gives students the freedom to access any tool(s) they wish at any time.

Our previous research (Liu et al., 2004, 2009) has indicated the problem-solving process in AR can be grouped into four conceptual stages: (a) understanding the problem (roughly days 1–2), (b) identifying, gathering, and organizing information (days 3–7), (c) integrating information (days 8–10), and (d) evaluating the process and outcome (days 11–13). This four-stage process reflects the cognitive processes in the revised version of Bloom's taxonomy (Anderson et al., 2001) and the five components of an IDEAL problem-solver (Bransford & Stein, 1984).

4 Method

4.1 Participants

Participants were sixth graders from a school in a mid-sized southwestern city. The teacher reported that most students were comfortable with computers as computer activities were a common part of classroom instruction. These sixth graders used AR as their science curriculum for approximately 3 weeks in the spring of 2014.

4.2 Data Sources

4.2.1 Log Files

All student actions performed while using the program were logged to a data file, which contained time- and date-stamped entries for each student. The data set consisted of the number of times a student accessed each of the cognitive tools and the amount of time the student used each tool. The participants were introduced to the central problem by watching a video scenario together, and then used the program in their science classes. The log file data presented a view of which tools a student used and for how long during this 3-week period.

Table 8.2 Rubric used for grading solution forms

Description	Points awarded
The student recommends an unsuitable home for the alien species	0
The student recommends a suitable home, but does not provide any reasons to substantiate their choice	1
The student recommends a suitable home and is awarded one additional point for each reason provided to substantiate their choice	2–7

4.2.2 Solution Scores

Students' performance was evaluated by the quality of their solution to the central problem. A student's solution score was determined by how well she solved the problem of finding an appropriate relocation home for each alien species. Variations in pace of work resulted in students submitting different numbers of solutions, in which case we used only one solution score. Assuming the quality of solutions would increase as a student gained more experience in solving the problem, we chose to score the last solution a student submitted.

The assessment of students' performance was evaluated using an 8-point rubric that considers both the suitability of the recommended home and the degree to which students justify their recommendation based upon the evidence they have collected (see Table 8.2).

Two researchers who had recently scored a set of solutions from another school participated in this scoring task. They first reviewed the scoring rubric and scored five solutions together to ensure they applied the same criteria during scoring. Then, the researchers scored the remainder of the solutions independently.

4.2.3 Goal Orientation

Students' goal orientation was measured by the revised *Patterns of Adaptive Learning Scales* (PALS, Midgley et al., 2000), which assesses personal achievement goal orientations through three subscales: mastery ($r = .85$), performance-approach ($r = .89$), and performance-avoidance ($r = .74$) goals with 4 items for each goal orientation and a total of 12 items. Each item was rated on a 5-point scale with 1 being "Not at all true," 3 being "Somewhat true," and 5 being "Very true." Due to this particular learning context, the general term "class" was replaced with "science class" as in these sample statements:

My goal in this science class is to learn as much as I can (mastery).
My goal is to show others that I'm good at my science class work (performance-approach).
It's important to me that I don't look stupid in my science class (performance-avoid).

We looked for natural groupings of the goal orientation scores, which resulted in two groups for mastery and three groups each for performance-approach and performance-avoid (see Table 8.3).

Table 8.3 Grouping based upon students' goal orientation scores and solution scores

Variable			Score	Number of students
Goal orientation (score: 1–5)	Mastery	High	=5	9
		Low	<5	7
	Performance-approach	High	≥3.75	3
		Mid	>2.75 and <3.75	7
		Low	≤2.75	6
	Performance-avoidance	High	≥4	3
		Mid	>3 and <4	7
		Low	≤3	6
Solution score (score: 0–7)		High	≥4	11
		Low	<4	27

4.3 Data Processing and Analysis

4.3.1 Data Cleaning and Processing

Each log file contained: student ID, teacher ID, time stamp including start time, end time, and duration; cognitive tools; and solution texts. After the data was cleaned, students' solution and goal orientations scores were matched with their log files. Only the matched data were included in this study. Since this study was conducted in a real classroom setting, not all students completed all measures, which necessitated dropping the non-matched data and reduced the overall sample size. Students who did not submit any solutions were also removed from the sample.

For research question 1, we examined overall behavior patterns. The log files of 47 students with 7,404 lines of logs were included. To address the second and third research questions, the matched log files with solution scores of 38 students and the matched log files with goal orientation scores of 16 students comprised the respective analyses. Students' solution and mastery goal orientation scores were grouped into high and low (see Table 8.3). Performance-approach and performance-avoid scores were grouped into high, mid, and low.

4.3.2 Analysis

We selected *Tableau Desktop* (tableausoftware.com, Computer software, Seattle, WA) as our visualization tool, since it enables the representation of multidimensional data or multiple layers of information in a single view. To examine overall behavior patterns, we performed descriptive analyses on usage of tools by Lajoie's (1993) four conceptual categories during the entire 3-week period. For log data, we used measures of frequency (number of times a tool was accessed) and duration (total amount of time, in sec., spent with a particular tool) averaged across students for a given time period. We then examined the tool use patterns by different grouping variables (i.e., performance or goal orientation). Specifically we used action shapes (Scarlatos & Scarlatos, 2010) to indicate tool use by each group. For the *X*-axis, we ordered the tools used in each of the four conceptual problem-solving stages or log days to understand different behavior patterns across the stages and over the entire period. The *Y*-axis represents the average frequency or average total duration of tool use by the grouping variable. Among all available tools, we focused on the six most frequently used tools: the Alien Database, Solar System Database, Notebook, Probe Design, Probe Launch, and Mission Control. ANOVAs were performed with grouping variables as the independent variables and frequency and duration of tool use as dependent variables.

5 Findings

For research question one, we examined frequency and duration across all tools for the entire sample. The findings confirmed that play-learners tended to use the tools that were central to the problem-solving process more frequently and for longer. For research questions two and three, we concentrated on six essential tools, looking for patterns according to performance levels and goal orientations. The findings suggested that some patterns of tool use were related to these grouping variables, though at this time the causal mechanism can only be speculated.

5.1 How Do Play-Learners Access Different Tools Built into the Game?

Figure 8.2 presents an overall picture of tool use patterns. The visualization indicates tools in the cognitive load category, especially the Solar System and Alien Databases, were used for significantly longer periods of time than those in the other tool categories ($\text{Mean}_{\text{SolarDB}} = 382.03$, $\text{Mean}_{\text{AlienDB}} = 525.80$, $F(9, 5154) = 154.64$, $p < 0.001$). The cognitive-processing tool, the Notebook, was used for a longer time on day 2 and then again on days 9–12. The Probe Design tool was used frequently, especially on day 8, and for longer on day 5 and often towards the end of the program. Tools for hypothesis testing were used most frequently on days 8–10, coinciding with increased activity with the Probe Design tool. It appears the most active period of overall tool use was around day 8.

Of all the tools, the students used Probe Design (frequency = 3.695) and Mission Control (frequency = 3.804) most often, while they stayed in the Alien Database (525.80 s) and Solar System Database (382.03 s) the longest (see Fig. 8.3). During the problem-solving process, the Alien Database is needed to understand alien characteristics and the Solar System Database is needed to understand what each planet in our solar system can offer. Probably the most fun tool is Probe Design, a simulation allowing students to equip a probe with scientific instruments. Mission Control presents the data from a launched probe. As Fig. 8.3 shows, students accessed these latter tools often, but not for long periods.

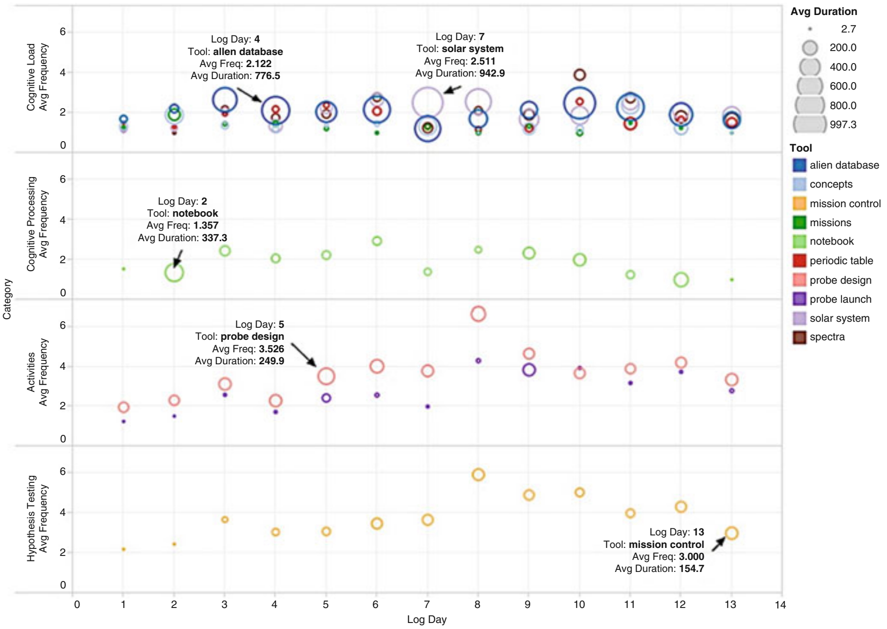


Fig. 8.2 Average frequency and duration of tool use

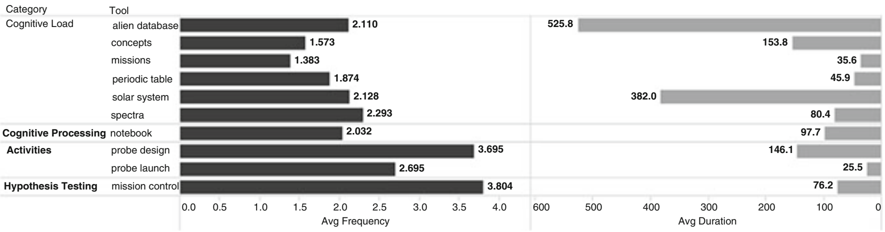


Fig. 8.3 Average frequency and duration of tool use by categories

5.2 *How Do Play-Learners with Different Goal Orientations Access the Tools?*

5.2.1 Mastery Goal Orientation (Mastery GO)

In examining tool use patterns by different goal orientation groups, we focused on six tools the students tended to use the most as shown above: Alien Database, Solar System Database, Notebook, Probe Design, Probe Launch, and Mission Control. In Figs. 8.4 and 8.5, each point in a shape represents the average frequency or duration of tool use according to its value on the *Y*-axis. During Stage 2, the Mastery GO High group used the Alien DB significantly more often ($\text{Mean}_{\text{AlienDB\_High}} = 2.25$,

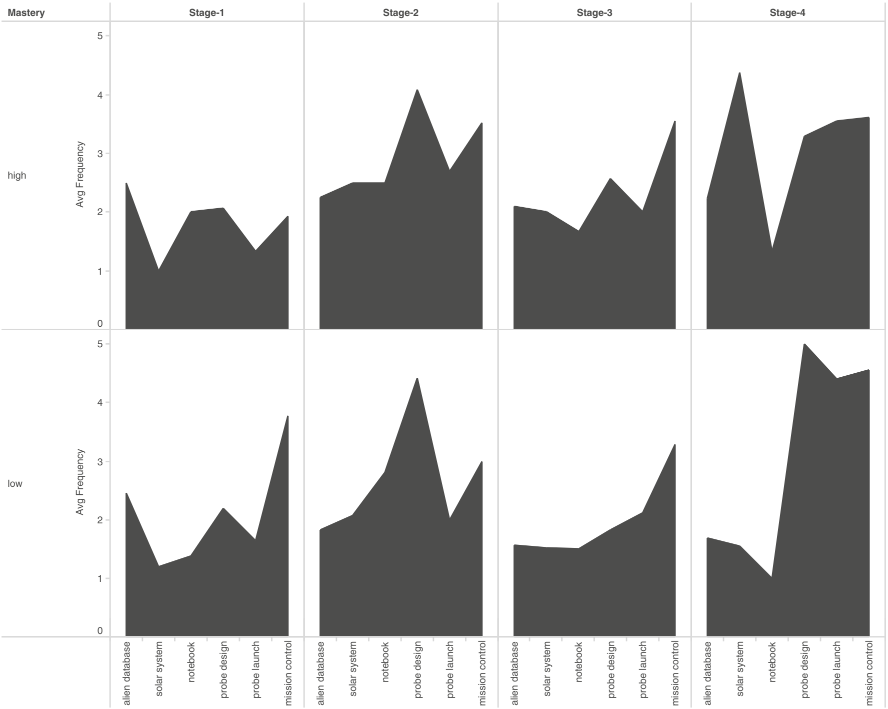


Fig. 8.4 Average frequency of tool use across four stages by mastery goal orientation groups

$\text{Mean}_{\text{AlienDB\_Low}} = 1.83$, $F(1, 110) = 4.135$, $p < 0.05$) and for longer ($\text{Mean}_{\text{AlienDB\_High}} = 727.62$, $\text{Mean}_{\text{AlienDB\_Low}} = 586.80$) than the Mastery GO Low group. They also stayed in the Solar System DB significantly longer ($\text{Mean}_{\text{SolarDB\_High}} = 245.03$, $\text{Mean}_{\text{SolarDB\_Low}} = 84.18$, $F(1, 64) = 5.435$, $p < 0.05$). As discussed above, these two tools are critical for this stage of problem solving. Stage 2 activities center on identifying, gathering, and organizing information in order to further refine the problem.

Therefore, the Alien and Solar Databases are critical to performing these activities. What is interesting, however, is that during Stage 4 the Mastery GO High group also used the Alien Database and Solar Database significantly more: $\text{Mean}_{\text{AlienDB\_High}} = 2.23$, $\text{Mean}_{\text{AlienDB\_Low}} = 1.69$, $F(1, 68) = 5.19$, $p < 0.05$; $\text{Mean}_{\text{SolarDB\_High}} = 4.38$, $\text{Mean}_{\text{SolarDB\_Low}} = 1.56$, $F(1, 42) = 21.46$, $p < 0.01$. In fact, the Mastery GO High group used both the Solar System and Alien Databases consistently more throughout the four stages as compared to the Mastery GO Low group. It is possible they used these two content databases to help verify the information returned from launched probes. The findings also indicate that the Mastery GO Low group used the Probe Design significantly more ($\text{Mean}_{\text{ProbeDesign\_Low}} = 5$, $\text{Mean}_{\text{ProbeDesign\_High}} = 3.29$, $F(1, 54) = 6.93$, $p < 0.01$), which is appropriate to this stage.

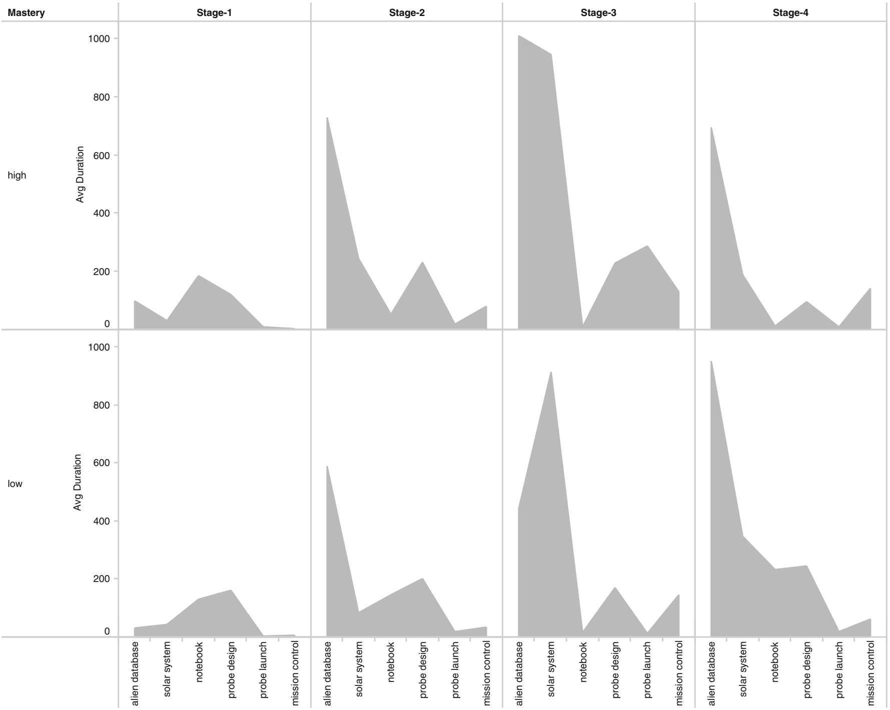


Fig. 8.5 Average duration of tool use across four stages by mastery goal orientation groups

5.2.2 Performance-Approach Goal Orientation (Performance GO)

The Performance GO High group only used Probe Design and little of other tools during Stage 1 and yet, used the Solar Database more during Stage 4 (see Fig. 8.6, $\text{Mean}_{\text{SolarDB\_High}} = 5.33$, $\text{Mean}_{\text{SolarDB\_Mid}} = 2.67$, $\text{Mean}_{\text{SolarDB\_Low}} = 3.1$, $F(2, 41) = 3.05$, $p = 0.06$). Performance GO Mid group showed high usage of Probe Design in Stage 2 (see Fig. 8.6, $\text{Mean}_{\text{ProbeDesign\_High}} = 3.56$, $\text{Mean}_{\text{ProbeDesign\_Mid}} = 4.91$, $\text{Mean}_{\text{ProbeDesign\_Low}} = 3.64$). These patterns indicate inappropriate tool use relative to problem-solving stage. On the other hand, the Performance GO Low group used the Alien Database significantly longer in Stage 3 (see Fig. 8.7, $\text{Mean}_{\text{AlienDB\_High}} = 355.52$, $\text{Mean}_{\text{AlienDB\_Mid}} = 853.18$, $\text{Mean}_{\text{AlienDB\_Low}} = 1042.01$, $F(2, 69) = 3.678$, $p < 0.05$). The Performance GO Mid and Low groups also used the Solar System Database longer in Stage 3 ($\text{Mean}_{\text{SolarDB\_High}} = 689.60$, $\text{Mean}_{\text{SolarDB\_Mid}} = 966.43$, $\text{Mean}_{\text{SolarDB\_Low}} = 993.89$) and used Probe Design significantly more frequently in Stage 4 (see Fig. 8.6, $\text{Mean}_{\text{ProbeDesign\_High}} = 1.75$, $\text{Mean}_{\text{ProbeDesign\_Mid}} = 4.00$, $\text{Mean}_{\text{ProbeDesign\_Low}} = 4.07$, $F(2, 53) = 4.061$, $p < 0.05$). These patterns indicate more appropriate tool use for the problem-solving stages.

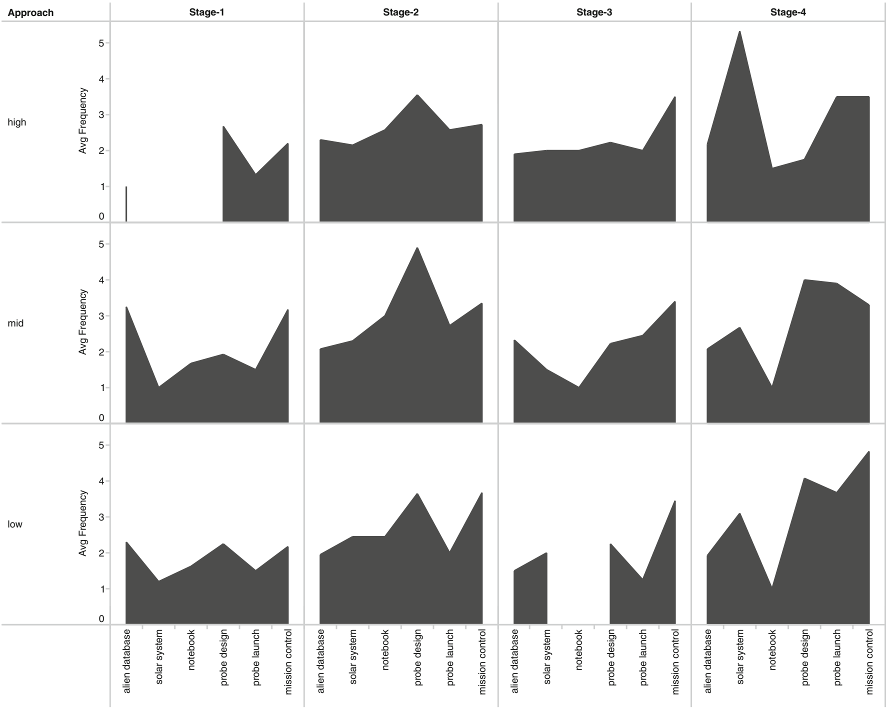


Fig. 8.6 Average frequency of tool use across four stages by performance-approach goal orientation groups

5.2.3 Performance-Avoidance Goal Orientation (Performance-Avoid GO)

Figures 8.8 and 8.9 present tool use patterns by groups according to their degree of performance-avoidance. Since the same students in the Performance GO High group were also in the Performance-Avoidance GO High group, the pattern for this group was the same as above. The Performance-Avoid GO Low group showed significantly high use of the Solar System Database in Stage 2 ($\text{Mean}_{\text{SolarDB\_High}} = 2.14$, $\text{Mean}_{\text{SolarDB\_Mid}} = 1.50$, $\text{Mean}_{\text{SolarDB\_Low}} = 2.94$, $F(2, 63) = 4.991$, $p < 0.05$), while the Performance GO Mid group showed high usage of Probe Design Tool in this stage ($\text{Mean}_{\text{ProbeDesign\_High}} = 3.56$, $\text{Mean}_{\text{ProbeDesign\_Mid}} = 5.11$, $\text{Mean}_{\text{ProbeDesign\_Low}} = 3.65$). The Performance-Avoid GO High group also used the Solar System Database significantly more during the last stage ($\text{Mean}_{\text{SolarDB\_High}} = 5.33$, $\text{Mean}_{\text{SolarDB\_Mid}} = 2.44$, $\text{Mean}_{\text{SolarDB\_Low}} = 3.30$, $F(2, 41) = 3.617$, $p < 0.05$).

Performance-Avoid GO Low group used these tools longer during Stage 3: Probe Design ($\text{Mean}_{\text{ProbeDesign\_Low}} = 405.30$, $\text{Mean}_{\text{ProbeDesign\_Mid}} = 80.53$, $\text{Mean}_{\text{ProbeDesign\_High}} = 216.13$), Probe Launch ($\text{Mean}_{\text{ProbeLaunch\_Low}} = 476.78$, $\text{Mean}_{\text{ProbeLaunch\_Mid}} = 10.24$, $\text{Mean}_{\text{ProbeLaunch\_High}} = 10.79$), and Mission Control Tools ($\text{Mean}_{\text{MissionControl\_Low}} = 196.52$, $\text{Mean}_{\text{MissionControl\_Mid}} = 115.77$, $\text{Mean}_{\text{MissionControl\_High}} = 75.93$). These patterns by the Performance-Avoid GO Low group suggest that students in the Low group used

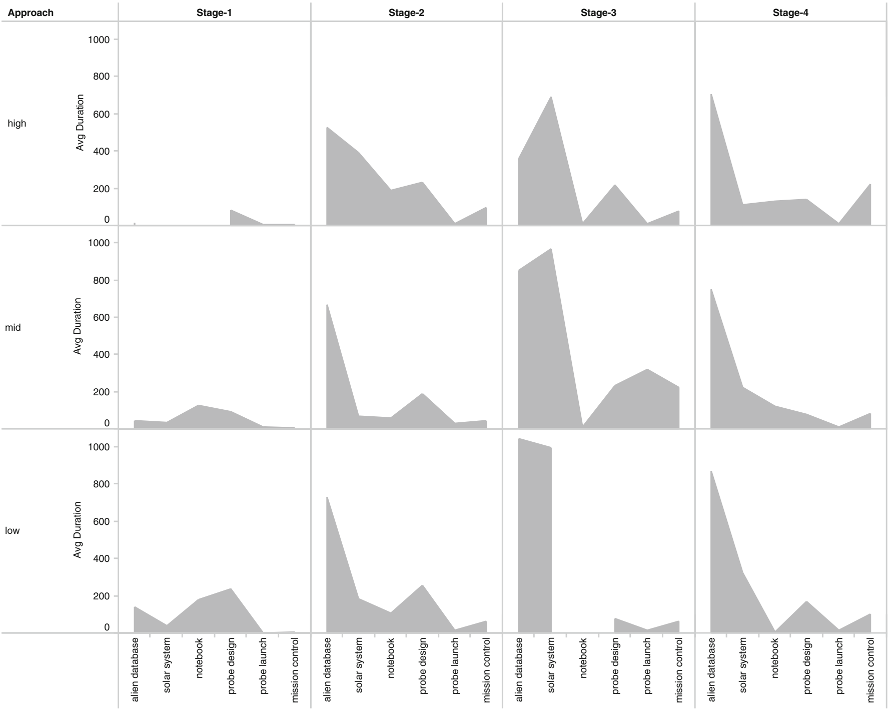


Fig. 8.7 Average duration of tool use across four stages by performance-approach goal orientation groups

tools more appropriate to the problem-solving stages while the Performance-Avoid GO High group seemed to only explore the more fun tools such as Probe Design, Probe Launch, and Mission Control in Stage 1.

5.3 How Do Play-Learners with Different with Performance Scores Access the Tools?

Students in the High Solution (HS) group ($n = 11$ with scores ≥4 out of 7) used the cognitive load tools significantly longer, specifically the Alien and Solar System Databases, than students in the Low Solution (LS) group did ($n = 27$ with scores <4): $\text{Mean}_{\text{SolarDB\_High}} = 492.70$, $\text{Mean}_{\text{SolarDB\_Low}} = 311.71$, $F(1, 490) = 11.94$, $p < 0.01$; $\text{Mean}_{\text{AlienDB\_High}} = 705.31$, $\text{Mean}_{\text{AlienDB\_Low}} = 438.15$, $F(1, 714) = 30.572$, $p < 0.001$ (see Figs. 8.10 and 8.11). Use of activities-out-reach tools increased and peaked on day 8 and use of hypothesis tools increased and peaked on day 10 for HS students, indicating they began to integrate information and test their hypotheses. HS students also utilized the Notebook tool more often and for longer in the initial days than did LS students.

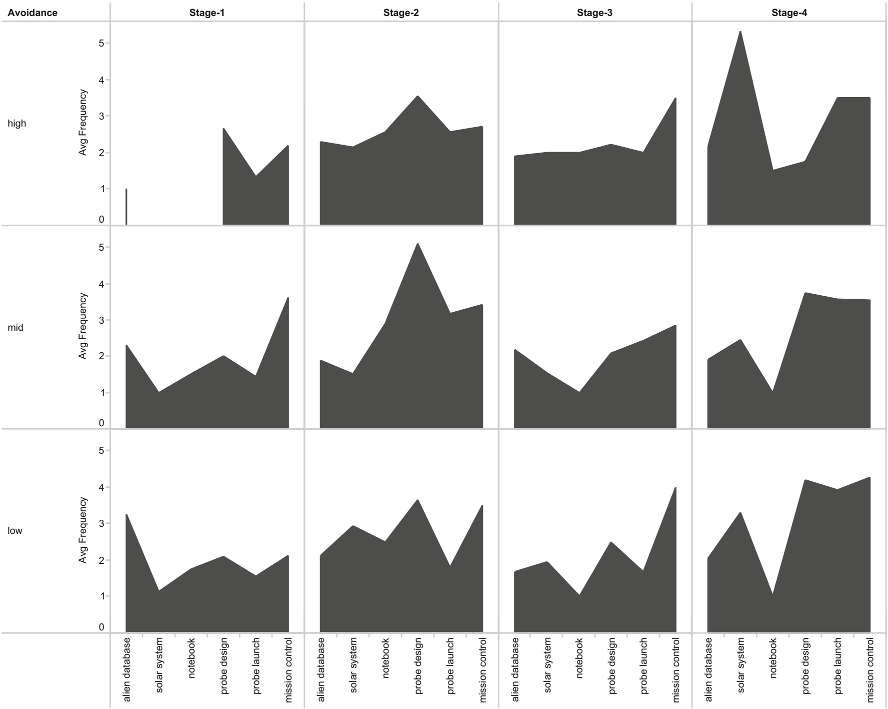


Fig. 8.8 Average frequency of tool use across four stages by performance-avoidance goal orientation groups

Together these patterns indicated more active use of the tools appropriate to the four stages by the HS group. The HS group also used most of the cognitive load tools longer than LS students did. This indicates that these HS students took more advantage of the domain-knowledge scaffolding provided by the serious game.

6 Discussion and Implications

The visualizations revealed several patterns of relevance to our ongoing efforts to design and enhance serious games such as *Alien Rescue*. The ultimate goal is to design effective scaffolds based upon our growing understanding of learner behaviors.

6.1 General Patterns of Tool Use

In general, the results supported our previous research into the four stages of the problem-solving process of AR (Liu & Bera, 2005; Liu et al., 2009). This is significant because play-learners are allowed to move through the process at their own pace and are not guided in how to proceed. In addition, they more frequently

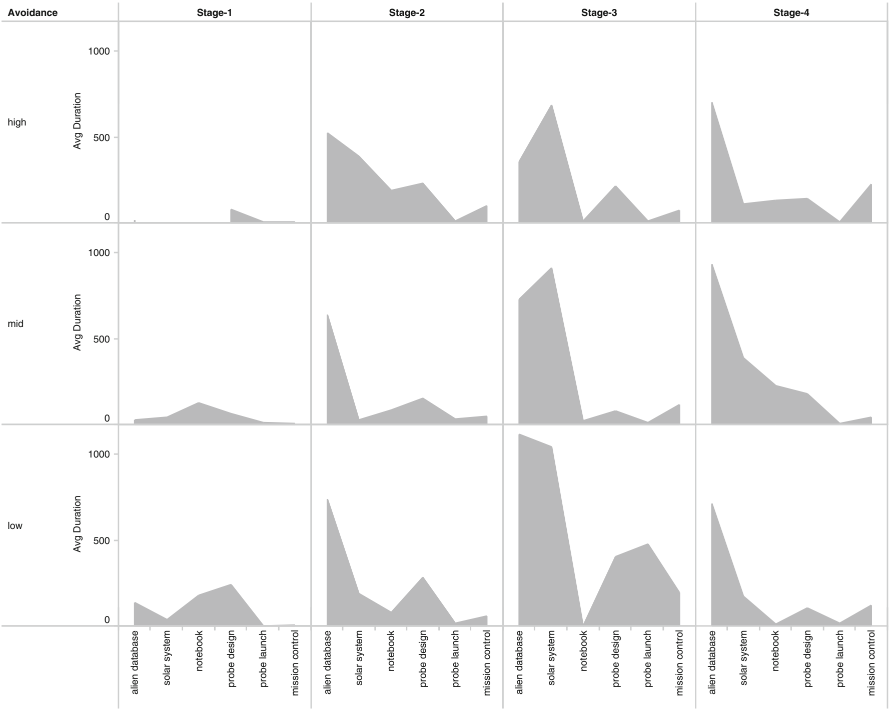


Fig. 8.9 Average duration of tool use across four stages by performance-avoidance goal orientation groups

accessed and spent more time with the six tools that are most vital for solving the central problem. That the play-learners generally play the game "as intended" stands testament to the pedagogical soundness of the design.

The Notebook, which supports cognitive processes related to the synthesis and application of knowledge, was only infrequently accessed by the students. We wondered why since we consider the Notebook to be an integral part of the AR problem-solving process (Liu et al., 2009; Liu, Horton, Toprac, & Yuen, 2012). This finding can possibly be explained by our classroom observations over the years which revealed that teachers often assigned worksheets for students to complete during the AR unit that perform similar functions to the Notebook (Liu, Wivagg, Geurtz, Lee, & Chang, 2012). It is likely that students are doing the work of recording and organizing information on these paper worksheets, rather than with the built-in Notebook tool, thereby achieving the same end by different means. However, such paper worksheets may or may not be designed with the problem-based learning pedagogical approach that is the foundation of this serious game, and they may take away from the immersive experience of the play-learners. For future improvements to AR, we hope to address this undesired outcome by making the content of students' notes available to the teacher, thereby eliminating the impetus to assign paper-and-pencil work.

The Alien and Solar System Databases represent two critical tools for gathering information and were therefore frequently accessed, yet students tended to stay in

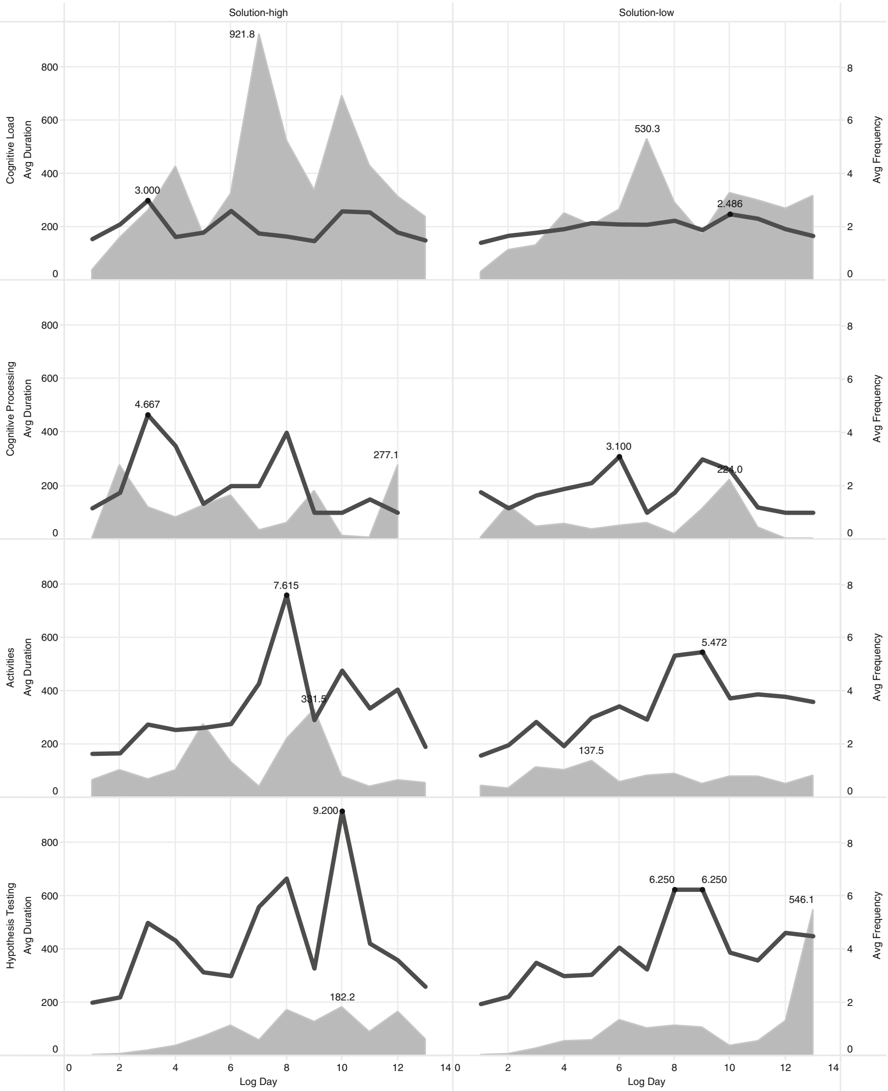


Fig. 8.10 Average frequency and duration of four categories tool use by solution groups (*lines* representing frequency and *areas* representing duration)

the Alien Database much longer than the Solar System Database as the visualization showed. This finding can possibly be explained by the fact that the Solar System Database can be accessed at any time via a pop-up window, whereas students must navigate to the Research Lab to view the Alien Database (see Fig. 8.1a). So students need to navigate to the Alien Database first and then access the Solar Database concurrently. Another possible explanation is that Alien Database with 3D models and animations may just be more engaging for students, as our previous research has indicated (Liu et al., 2013).

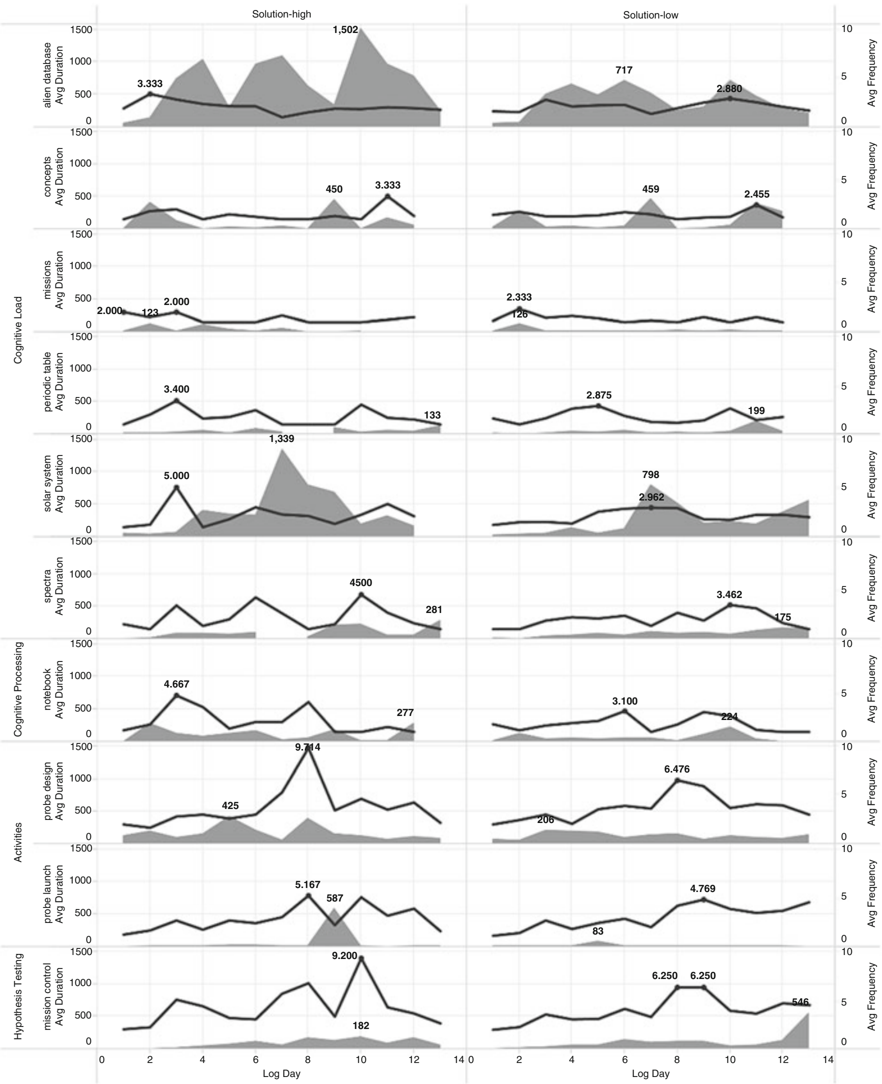


Fig. 8.11 Average frequency and duration of individual tool use by solution groups (*lines* representing frequency and *areas* representing duration)

6.2 Productive Tool Use by High-Performance and Mastery Goal Orientation Groups

Our previous research has indicated that high-performing and low-performing students differed in their patterns of tool use (Liu & Bera, 2005). The present study confirmed this finding and additionally linked the similar pattern of productive tool use shown by high-performing students to those with a mastery goal orientation, as might be expected from the previously established connection between goal orientation and performance (Hsieh, Cho, Liu, & Schallert, 2008). Students in the High Solution and Mastery GO High groups tended to use the tools more appropriately according to the problem-solving stages. HS students used cognitive load and processing tools more and longer during Stages 2 and 3 and Probe Design and Launch centers during Stages 3 and 4, exactly when these tools are most pertinent. Since all students in a class are generally given the same amount of time to solve the central problem, less productive tool use can affect performance scores, as shown by the findings.

Concerning the other two goal orientation groups related to performance, the patterns are less straightforward. In our sample, the same students appeared in both the Performance GO High and the Performance-avoid GO High groups. This puzzling result is perhaps due to the small sample size and therefore limits the conclusions to be drawn. What is more, although the performance-related goal orientation groups showed active use of tools at times, they did not show a clear pattern on in-game productivity in contrast to the high-performing and mastery-oriented groups.

Goal orientation indicates a student's motivations for completing an academic task, which play an influential role on behaviors and performance (Ames, 1992; Dweck, 1986). Students with a mastery goal orientation tend to focus more on mastering a task and acquiring new skills, and less on how competent they look in front of others (performance-approach goal) or on avoiding unfavorable judgments of capabilities and embarrassment in front of peers (performance-avoidance goal) (Elliot, 1999; Elliot & Harackiewicz, 1996). The findings from this study offered some evidence in support of the literature on goal orientations (Middleton & Midgley, 1997; Midgley & Urdan, 1995; Pajares, Britner, & Valiante, 2000) in that students with a mastery goal orientation tend to show more positive patterns of learning, while students with performance-approach or performance-avoidance goals appear to try to find a quick way to solve the complex problem and do not exhibit purposeful learning patterns.

6.3 *Visualization as a Promising Technique for Serious Games Analytics*

Our experience of visually exploring log data in combination with data from traditional sources indicates visualization as a promising technique in serious games analytics, especially with multidimensional data sets. Visualization facilitates interpretation of the relationships among multiple data points at no cost to the complexity of the data (Milam & El Nasr, 2010; Scarlatos & Scarlatos, 2010). In our study, by displaying data points (tool use frequency, duration) over days and across stages according to different grouping variables (performance levels and goal orientations) in a multidimensional way, visualization helped present the data and reveal findings not easily detected using traditional measures. The findings confirmed some of our previous research findings and more importantly, also revealed areas that call for further research. For examples, the Mastery GO High group used both the Solar System and Alien Databases consistently more throughout the four stages as compared to the Mastery GO Low group. Why? Could this be attributed to their mastery goal orientation or to other factors? The Performance GO High group only used Probe Design and little of other tools during Stage 1. Does their goal orientation have anything to do with this finding? The findings showed the potential of using visualization to facilitate the interpretation of how multiple data points may contribute to the patterns of play-learners' behaviors as they engage in an SG environment, and provided empirical support for the use of multifaceted approaches to visually represent complex and sophisticated information (Drachen & Canossa, 2009; Linek et al., 2008; Wallner & Kriglstein, 2013).

6.4 Limitations and Future Directions

This study involved discovering patterns of play-learner behavior among students grouped by performance levels and goal orientations. Therefore, we limited the log data to the students who completed at least one of the measures, which reduced the overall sample size. The small size of the matched data used in this analysis is a limitation. For the log data, it was first necessary to manually compare the time stamp and the school calendar to calculate how long a class used AR while eliminating school holidays and testing days. As a part of our future work, we intend to develop code to parse log data into a more useable format. An attempt to make the processing of the log data automatic is a logical next step for our future research on this topic.

Our research group plans to continue this line of inquiry in several ways. First, we are designing an interactive dashboard for teachers, which will enable them to more closely monitor students' work and thereby facilitate and intervene in a play-learner's activity as needed. Visualizations, including some presented in this chapter, will allow teachers to monitor activities at the level of the classroom and an individual student, thereby facilitating both classroom management and grading. As we continue to refine our analytics and visualization techniques, we hope to replace the paper-and-pencil worksheets with more empirically tested analytics. We consider the exploration into visualization reported in this chapter as an important initial step in our application of serious games analytics to AR.

A second application of SEGA to AR will involve the provision of cognitive feedback to play-learners in the environment through visualizations. Thus far, in-game scaffolding and teacher support have been, for practical reasons, restricted to information about the task itself. The introduction of analytics-based visual feedback to play-learners can provide feedback on their decision-making processes and the effectiveness of those decisions, thus increasing the potential of success for all students (Balzer, Doherty, & O'Connor, 1989).

We used the commercial data visualization software, *Tableau* for data analysis in this study. The ready-made visualizations created using this software have facilitated

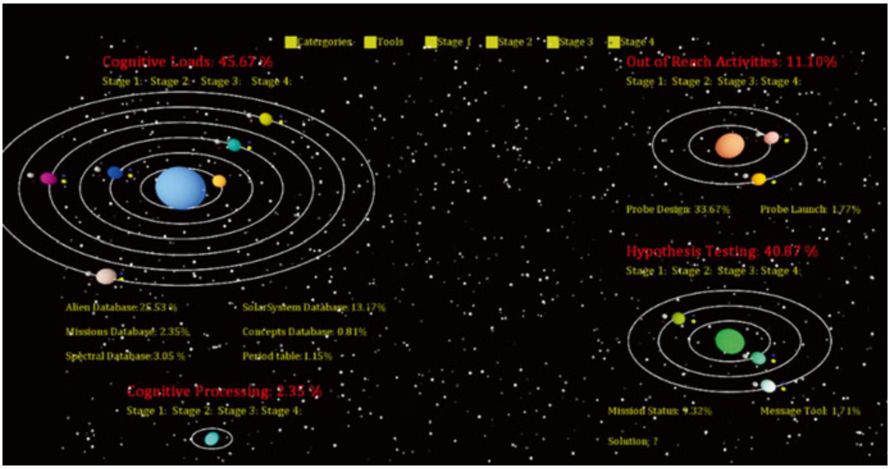


Fig. 8.12 Average frequency of tool use by four categories

our exploratory analysis in this study, but the output cannot be fully customized to fit our future needs for displaying just-in-time visualizations within the context of this SG. We are therefore also exploring in-house development of visualizations that can convey the data in forms consistent with the serious game context. Figure 8.12 represents our initial effort: We used *Processing* (Software, retrieved from https://www.processing.org/download/, 2001) to visualize the overall tool use patterns using a solar system metaphor as aligned with the theme of AR. There are four solar systems, in which each sun represents a tool category and each revolving planet signifies a tool in that category. The size of every object including sun, planets, and moons indicates average frequency of tool use. This is our preliminary attempt to situate data visualization within the specific serious game context. We will continue to pursue this endeavor, particularly in conjunction with efforts to provide gameplay data to teachers and students, as outlined above.

7 Conclusion

We have reported on a study using serious games analytics and data visualizations to discover patterns of play-learner behaviors in *Alien Rescue*, a serious game for sixth-grade space science. Play-learners' use of built-in cognitive tools was visually presented in multiple formats and discussed according to trends among all students in the sample and between groups that differed according to performance levels and goal orientations. The results showed that specific patterns of tool use do indeed correlate with successful performance. The results were discussed in terms of the pedagogical implications for the design of the serious game and the integral role that serious games analytics and data visualization will play in that effort.

Acknowledgments We would like to acknowledge the help by Damilola Shonaike in creating the image in Figure 14 as part of her 2014 summer CERT REU internship program. We also appreciate the help from Divya Thakur and Kelly Gaither from the Texas Advanced Computing Center at the University of Texas at Austin in exploring the use of the Processing language to create visualizations in the specific game environment.

References

Abt, C. C. (1970). *Serious games*. New York: The Viking Press.

Ames, C. (1992). Achievement goals and classroom motivational climate. In J. Meece & D. Schunk (Eds.), *Students' perceptions in the classroom* (pp. 327–348). Hillsdale, NJ: Erlbaum.

Andersen, E., Liu, Y. E., Apter, E., Boucher-Genesse, F., & Popović, Z. (2010). Gameplay analysis through state projection. In *Proceedings from The Fifth International Conference on the Foundations of Digital Games,* Pacific Grove, CA (pp. 1–8). doi:10.1145/1822348.1822349.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., et al. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3–17.

Balzer, W. K., Doherty, M. E., & O'Connor, R. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin, 106*(3), 410.

Barab, S. A., Gresalfi, M., & Ingram-Goble, A. (2010). Transformational play using games to position person, content, and context. *Educational Researcher, 39*(7), 525–536.

Bransford, J. D., & Stein, B. S. (1984). *The IDEAL problem solver*. New York: W.H. Freeman and Company.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences, 2*(2), 141–178. doi:10.1207/s15327809jls0202_2.

Cobb, P., Confrey, J., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9–13. doi:10.3102/0013189X032001009.

Dede, C. (2014, May 6). *Data visualizations in immersive, authentic simulations for learning* [Flash slides]. Retrieved from http://www.edvis.org/tuesday-presentations/

Dixit, P. N., & Youngblood, G. M. (2008). Understanding playtest data through visual data mining in interactive 3d environments. In Proceedings from *12th International Conference on Computer Games: AI, Animation, Mobile, Interactive Multimedia and Serious Games (CGAMES)* (pp. 34–42).

Djaouti, D., Alvarez, J., Jessel, J. P., & Rampnoux, O. (2011). Origins of serious games. In M. Ma, A. Oikonomou, & L. C. Jain (Eds.), *Serious games and edutainment applications* (pp. 25–43). Berlin, Germany: Springer.

Drachen, A., & Canossa, A. (2009). Towards gameplay analysis via gameplay metrics. In *Proceedings from the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era* (pp. 202–209). ACM. doi:10.1145/1621841.1621878.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41*, 1040–1048.

Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist, 34*, 169–189.

Elliot, A. J., & Harackiewicz, J. M. (1996). Approach and avoidance achievement goals and intrinsic motivation: A mediational analysis. *Journal of Personality and Social Psychology, 70*, 461–475.

Garzotto, F. (2007). Investigating the educational effectiveness of multiplayer online games for children. In *Proceedings from the 6th International Conference on Interaction Design and Children,* Aalborg, Denmark (pp. 29–36). doi:10.1145/1297277.1297284.

Holcomb, J., & Mitchell, A. (2014, March). *The revenue picture for American journalism and how it is changing*. Retrieved from http://www.journalism.org/2014/03/26/the-revenue-picture-for-american-journalism-and-how-it-is-changing/

Hsieh, P., Cho, Y., Liu, M., & Schallert, D. (2008). Examining the interplay between middle school students' achievement goals and self-efficacy in a technology-enhanced learning environment. *American Secondary Education, 36*(3), 33–50.

Johnson, L., Adams Becker, S., Cummins, M., Estrada, V., Freeman, A., & Ludgate, H. (2013). *NMC horizon report: 2013 Higher Education Edition*. Austin, TX: The New Media Consortium.

Johnson, L., Adams Becker, S., Estrada, V., & Freeman, A. (2014). *NMC horizon report: 2014 Higher Education Edition*. Austin, TX: The New Media Consortium.

Lajoie, S. P. (1993). Computer environments as cognitive tools for enhancing learning. In S. P. Lajoie & S. J. Derry (Eds.), *Computers as cognitive tools* (pp. 261–288). Hillsdale, NJ: Lawrence Erlbaum Associates.

Linek, S. B., Marte, B., & Albert, D. (2008). The differential use and effective combination of questionnaires and logfiles. In *Computer-Based Knowledge & Skill Assessment and Feedback in Learning Settings (CAF), Proceedings from The International Conference on Interactive Computer Aided Learning (ICL)*, Villach, Austria.

Linek, S. B., Öttl, G., & Albert, D. (2010). Non-invasive data tracking in educational games: Combination of logfiles and natural language processing. In L. G. Chova, D. M. Belenguer (Eds.), *Proceedings from INTED 2010: International Technology, Education and Development Conference*, Spain, Valenica.

List, J., & Bryant, B. (2014, March). Using Minecraft to encourage critical engagement of geography concepts. In *Society for Information Technology & Teacher Education International* [*Conference Proceedings*] (pp. 2384–2388). Jacksonville, FL.

Liu, M., & Bera, S. (2005). An analysis of cognitive tool use patterns in a hypermedia learning environment. *Educational Technology Research and Development, 53*(1), 5–21. doi:10.1007/BF02504854.

Liu, M., Bera, S., Corliss, S., Svinicki, M., & Beth, A. (2004). Understanding the connection between cognitive tool use and cognitive processes as used by sixth graders in a problem-based hypermedia learning environment. *Journal of Educational Computing Research, 31*(3), 309–334.

Liu, M., Horton, L. R., Corliss, S. B., Svinicki, M. D., Bogard, T., Kim, J., et al. (2009). Students' problem solving as mediated by their cognitive tool use: A study of tool use patterns. *Journal of Educational Computing Research, 40*(1), 111–139.

Liu, M., Horton, L., Kang, J., Kimmons, R., & Lee, J. (2013). Using a ludic simulation to make learning of middle school space science fun. *The International Journal of Gaming and Computer-Mediated Simulations, 5*(1), 66–86. doi:10.4018/jgcms.2013010105.

Liu, M., Horton, L., Toprac, P., & Yuen, T. T. (2012). Examining the design of media-rich cognitive tools as scaffolds in a multimedia problem-based learning environment. In *Educational media and technology yearbook* (pp. 113–125). New York: Springer.

Liu, M., Wivagg, J., Geurtz, R., Lee, S.-T., & Chang, H. M. (2012). Examining how middle school science teachers implement a multimedia-enriched problem-based learning environment. *Interdisciplinary Journal of Problem-Based Learning, 6*(2), 46–84.

Loh, C. S. (2008). Designing online games assessment as "Information Trails". In V. Sugumaran (Ed.), *Intelligent information technologies: Concepts, methodologies, tools, and applications* (pp. 553–574). Hershey, PA: Information Science Reference. doi:10.4018/978-1-59904-941-0.ch032.

Loh, C. S. (2011). Using *in situ* data collection to improve the impact and return of investment of game-based learning. In *Old Meets New: Media in Education—Proceedings of the 61*st *International Council for Educational Media and the XIII International Symposium on Computers in Education (ICEM & SIIE'2011) Joint Conference* (pp. 801–811). doi: 10.4018/jvple.2013010101.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an “early warning system” for educators: A proof of concept. *Computers & Education, 54*(2), 588–599. doi:10.1016/j.compedu.2009.09.008.

Middleton, M. J., & Midgley, C. (1997). Avoiding the demonstration of lack of ability: An under-explored aspect of goal theory. *Journal of Educational Psychology, 89*, 710–718.

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., et al. (2000). *Patterns of adaptive learning scales (PALS)*. Ann Arbor, MI: University of Michigan.

Midgley, C., & Urdan, T. (1995). Predictors of middle school students’ use of self-handicapping strategies. *The Journal of Early Adolescence, 15*, 389–411.

Milam, D., & El Nasr, M. S. (2010, July). Design patterns to guide player movement in 3D games. In *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games* (pp. 37–42). ACM. doi:10.1145/1836135.1836141.

Pajares, F., Britner, S., & Valiante, G. (2000). Relation between achievement goals and self-beliefs of middle school students in writing and science. *Contemporary Educational Psychology, 25*, 406–422.

Reese, D. D., Tabachnick, B. G., & Kosko, R. E. (2013). Video game learning dynamics: Actionable measures of multidimensional learning trajectories. *British Journal of Educational Technology*. doi:10.1111/bjet.12128.

Rideout, V. J., Foehr, U. G., & Roberts, D.F. (2010, January). *Generation M2: Media in the lives of 8- to 18-year-olds*. Kaiser Family Foundation. Retrieved from http://kff.org/other/poll-finding/report-generation-m2-media-in-the-lives/

Rieber, L. (1996). Seriously considering play: Designing interactive learning environments based on the blending of microworlds, simulations, and games. *Educational Technology Research and Development, 44*(2), 43–58. doi:10.1007/BF02300540.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 40*(6), 601–618. doi:10.1109/TSMCC.2010.2053532.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3*(1), 12–27. doi:10.1002/widm.1075.

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*(1), 368–384. doi:10.1016/j.compedu.2007.05.016.

Rosenbaum, E., Klopfer, E., & Perry, J. (2007). On location learning: Authentic applied science with networked augmented realities. *Journal of Science Education and Technology, 16*(1), 31–45. doi:10.1007/sl0956-006-9036-0.

Salen, K., & Zimmerman, E. (2004). *Rules of play: Game design fundamentals*. Cambridge, MA: MIT Press.

Sawyer, B., & Smith, P. (2008). *Serious games taxonomy*. [PDF document]. Retrieved from http://www.dmill.com/presentations/serious-games-taxonomy-2008.pdf

Scarlatos, L. L., & Scarlatos, T. (2010). Visualizations for the assessment of learning in computer games. In *7th International Conference & Expo on Emerging Technologies for a Smarter World (CEWIT 2010), September 27–29 2010*, Incheon, Korea.

Serrano, A., Marchiori, E. J., del Blanco, A., Torrente, J., & Fernández-Manjón, B. (2012, April). A framework to improve evaluation in educational games. In *Proceedings from Global Engineering Education Conference (EDUCON),* 2012 IEEE (pp. 1–8). IEEE. doi:10.1109/EDUCON.2012.6201154.

Squire, K. D. (2004). Review. *Simulation & Gaming, 35*(1), 135–140. doi:10.1177/1046878103255490.

Squire, K. D., & Jan, M. (2007). Mad City Mystery: Developing scientific argumentation skills with a place-based augmented reality game on handheld computers. *Journal of Science Education and Technology, 16*(1), 5–29. doi:10.1007/s10956-006-9037-z.

Sweetser, P., & Wyeth, P. (2005). GameFlow: A model for evaluating player enjoyment in games. *Computers in Entertainment, 3*(3), 3–3.

Tanes, Z., & Cemalcilar, Z. (2010). Learning from SimCity: An empirical study of Turkish adolescents. *Journal of Adolescence, 33*(5), 731–739.

U.S. Department of Education, Office of Educational Technology (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief.* Washington, DC.

van Barneveld, A., Arnold, K. E., & Campbell, J. P. (2012). *Analytics in higher education: Establishing a common language*. EDUCAUSE Learning Initiative. Retrieved from https://qa.itap.purdue.edu/learning/docs/research/ELI3026.pdf

Wallner, G., & Kriglstein, S. (2013). Visualization-based analysis of gameplay data—A review of literature. *Entertainment Computing, 4*(3), 143–155. doi:10.1016/j.entcom.2013.02.002.

Part IV
Serious Games Analytics for Medical Learning

Chapter 9
Using Visual Analytics to Inform Rheumatoid Arthritis Patient Choices

Radu P. Mihail, Nathan Jacobs, Judy Goldsmith, and Kristine Lohr

Abstract Individuals diagnosed with chronic diseases often face difficult, potentially life-altering, treatment decisions. Without sufficient knowledge, it can be difficult for a patient to make an informed decision. An essential element of medical care is educating the patient regarding disease outcomes and treatment options, thereby reducing feelings of uncertainty and increasing confidence in the resulting decision. It has been shown that incorporating decision aids (DAs) based on serious computer games into medical care can increase health knowledge and assist decision-making of patients with a variety of diseases. We discuss the benefits and challenges of using serious games as patient decision aids. We focus on rheumatoid arthritis (RA), a chronic disease that primarily affects the joints of the hand. We propose the use of serious games to enable RA patients to safely explore the uncertain effects of treatments through an avatar that performs common daily tasks, which are known to cause problems for RA patients, and experiences the side effects. We discuss the engineering challenges in building such a system and propose a data-driven approach using medical imagery to communicate the effects of the disease.

Keywords Game-based decision aids • Rheumatoid arthritis • Chronic diseases

R.P. Mihail (✉)
Valdosta State University, 1500 N Patterson Street, Valdosta, GA 31698, USA
e-mail: r.p.mihail@valdosta.edu

N. Jacobs • J. Goldsmith
University of Kentucky, 319 Davis Marksbury Building, 329 Rose Street,
Lexington, KY 40506, USA
e-mail: Jacobs@cs.uky.edu; goldsmit@cs.uky.edu

K. Lohr
University of Kentucky, Room J515 Kentucky Clinic, 740 S. Limestone Street,
Lexington, KY 40536-0284, USA
e-mail: kmlohr2@email.uky.edu

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based
Learning, DOI 10.1007/978-3-319-05834-4_9

1 Introduction

We face many choices every day. For example, we decide what color shirt to wear on a given day, which gas station to stop at for fuel, which route to take to work, how to distribute our income to satisfy several goals and objectives. This list can be made significantly longer, so it is natural to ask if some decisions we make are more important than others (i.e., we may have many defaults and a few decisions we take time to consider carefully) and if our decisions are the best. Surely, choosing a shirt is easier than investing in mutual funds or carefully considering the benefits and side effects of long-term medical treatments.

In this chapter, we consider the decisions patients have to make together with their healthcare provider, a paradigm called "shared decision-making." Studies have shown that patient involvement is directly correlated to decision satisfaction. The shared decision-making paradigm in the medical encounter expects patients to be active participants in their own health care. This model contrasts with traditional paternalistic care, in which the healthcare provider is expected to make the best decision for the patient. For shared decision-making to be effective, patients have to understand their choice set. Understanding the choice set depends on patients' literacy and numeracy skills, level of involvement in the process (active vs. passive) and face-to-face time spent with the medical practitioner. Because of healthcare economics, time with the practitioner is limited to approximately 15 min, which often leaves patients with incomplete and incoherent information. Face-to-face time with the practitioner is routinely augmented with printed informational materials such as pamphlets and leaflets, but due to the varied levels of patients' literacy, such printed material may not contribute to feelings of understanding and empowerment, nor to good decision-making. Surely, there are better ways to communicate information about the disease, the possible treatments, and the certain and possible consequences of those treatments.

In this chapter, we explore both the use and the construction of serious games as patient decision aids (DAs). We discuss some of the technical difficulties in the designing and development of such games. We also discuss how clinical data can be leveraged in the context of a game-based decision tool. We review a few relevant articles from recent research on game-based approaches used in the shared medical decision-making paradigm and discuss the applicability to game-based decision aids of our own work in medical image processing and machine learning. We focus on rheumatoid arthritis (RA), a chronic inflammatory autoimmune disease that primarily affects the synovial joints in the hands, because RA pain and functional impairments affect every aspect of life.

Interactive patient decision aids have been shown to be effective. Patient decision aid effectiveness can be assessed through several measures and instruments, discussed later in this chapter. The "gamification" of decision aids (Bandura, 1977; Beale, Kato, Marin-Bowling, Guthrie, & Cole, 2007; Bosworth, Gustafson, Hawkins, Chewning, & Day, 1983; Lieberman, 2001; Reichlin et al., 2011) led to a new class of serious games, with a purpose to educate and inform patients about their choice sets, while also providing an entertainment value. We will discuss current practice in RA care and argue how serious games can be a valuable tool for patients.

Playing video and computer games is now a popular pastime in most countries. Noticeably, this is not just limited to teens and adolescents. According to the Entertainment Software Association (ESA), about half of US households own an average of two dedicated game consoles; meanwhile, the average age of game players is 30 years old, with 37 % of them aged 36 or older (http://www.theesa.com). In less than a decade, casual and social games, partly triggered by the phenomenal Wii® success in attracting non-hardcore players, have significantly reshaped the video game culture (Juul, 2012). Video games create unique advantages for health-related interventions in a number of areas: massive reach (accessible to virtually all demographics in society); highly appealing (in packaging otherwise dull information in an easy-to-understand and possibly entertaining mode); enhanced learning/ communication outcomes (through delivering complex health-relevant messages to a large base of lay audience). An additional advantage is that, once a computer or game application is proven to be effective, it becomes rather inexpensive to mass produce and disseminate. In the following section, we discuss current practice in health care and decision-making for RA patients, and motivate the need for game-based approaches. We then discuss the challenges of designing and building a data-driven game-based DA for RA using medical image processing and machine learning.

2 Rheumatoid Arthritis Care

Patients with RA have multiple treatment options available, each associated with distinct risk profiles and, thus, tradeoffs. The prescription of a disease modifying anti-rheumatic drug (DMARD) has been designated as the first quality measure in the 2008 Physician Quality Reporting Initiative. There are many DMARD options available, including traditional synthetic drugs like methotrexate as well as an increasing number of targeted biologic molecules like etanercept and rituximab (Martin, Brower, Geralds, Gallagher, & Tellinghuisen, 2012). Each DMARD has attributes that can influence a clinician's and patient's decision to initiate treatment. Each has different effects on the progression of RA, from no impact on the baseline rate of progression to nearly halting all progression. For example, the most widely used DMARD, methotrexate, has been shown to slow the rate of progression of structural joint damage by 85 % in early RA patients (Martin et al., 2012). However, each DMARD also has different safety concerns (e.g., risk of serious infection) and costs associated with administration and monitoring.

Today's patients are faced with difficult decisions to choose one treatment from several available options with probabilistic outcomes. Their reasoning is biased by emotion, difficulty processing numerical data and misconceptions after reading literature that they consider pertinent to their condition (Schwab, 2008).

2.1 Patient–Physician Communication

Communication of treatment benefits and side effects is essential to medical care. Currently, RA patients are typically informed about treatment options through a brief consultation with the clinician, and/or leaflets, video and audio formats, and websites.

Patients with RA have multiple options from which to make complex decisions. Each drug offers potential benefits and risk profiles that patients may value differently (Grove, Hassell, Hay, & Shadforth, 2001). For example, biologic agents may induce remission, but may incur significant out-of-pocket costs and risk of serious infection or malignancy. When patients participate in shared decision-making, they want information about alternatives and the ability to assess risk and ask, "What if?"

Effective shared healthcare decision-making requires a quality patient–clinician interaction, which is often hindered by time constraints. To partially overcome this, patients receive verbal explanations, often combined with static decision aids (DAs) (e.g., printed materials) (Fagerlin, Wang, & Ubel, 2005; O'Connor et al., 2009). They also seek information from other patients and search the Internet to find relevant information that may not be scientifically sound. Patients are expected to understand their disease, treatment options, and associated risks, and to be competent to participate in treatment decisions. However, patients may not fully understand treatment options, risks, and benefits from written material. Conversely, some patients may fully understand the material, but some existing DAs fail to provide citations, leaving the burden of research fully on the patient if they wish to learn more (Feldman-Stewart et al., 2007). Moreover, patients have to carefully weigh cost of treatment, insurance, improvement in symptoms, and risk of harms. There are complex interactions, e.g., improvement in symptoms has monetary and, more generally, functionality implications. There is much possible harm with potentially independent risks, and those risks affect the patient's employment ability.

Schwab (2008) claims that patients or their families are unlikely to have a fundamentally different mental construct of the decision than the physician or practitioner. Disagreements appear when a decision is made about which option should be chosen (Schwab, 2008). Here, a DA can arguably act as a moderator, biased by the patient's preferences, and find the best fit with the patient's priorities. A fundamental concern for any DA is potentially intractable demands on decision-makers, leading to general misunderstandings conducive to uninformed decisions (Schwab, 2008). This may result in decreased longevity and reduced quality of life for the patient.

Patient participation is particularly relevant for many rheumatic diseases (Daltroy, 1993), compared to other conditions in which a treatment decision is urgent. In situations in which there is no time pressure associated with selecting a treatment, the patient can be given information to digest at home and potentially become an active participant in shared decision-making where it is not immediately clear which tradeoffs of effect and side effect likelihoods is most desirable for the patient (Daltroy, 1993). Ideally, patients comprehend their role in the process as seen by physicians; however, this is not always the case.

Patients can be divided roughly into two groups based on their interaction style with the physician (Daltroy, 1993): those who are more passive and accept a paternalistic approach by a clinician, and those who are more autonomous because they understand they are active participants. Haugli, Strand, and Finset (2004) found that patients with RA wished to be seen holistically by their physicians. Thus, interaction style becomes a factor in the decision-making process (Haugli et al., 2004). RA affects every aspect of patients' lives due to consistent pain, inability to perform routine tasks and deformities that develop in later stages of the disease. The feeling of being understood by physicians provides patients with a sense of security and emotional support during times of hardship and vulnerability (Daltroy, 1993).

Patient involvement is directly correlated to decision satisfaction. O'Connor, Légaré, and Stacey (2003) studied the impact that DAs have on risk communication prior to interaction with practitioners. They determined that, if a DA was used, the quality of the time spent with the provider was better, and decision satisfaction was increased. Thus, patient–clinician style has been shown to affect patient decision-making. In addition, DAs can help patients, whether or not the patients are otherwise involved in their own healthcare decision-making. Ishikawa, Hashimoto, and Yano (2006) address the relationship between participation style and the feeling of being understood by the physician. They conducted a qualitative study and found that patients who perceived themselves as having actively participated in the visit felt they were better understood by the physician. Conversely, those who were less active in the decision-making process during the visit felt less understood. Thus, giving patients inappropriate forms of information (for instance, pamphlets way above their literacy level) will not encourage participation nor feelings of being understood.

Another shortcoming of current practice is that written and verbal explanations may frame information such that the final decision is biased towards a specific outcome (Elwyn et al., 2006; Feldman-Stewart et al., 2007).

We discuss the uncertainties associated with clinical trials, treatment selection and personal values and preferences. We suggest that the critical ingredients which would improve patient DAs are dynamicity or interactivity using a computer or gaming console.

2.2 Decision-Making for RA Patients

DAs are tools to increase patients' knowledge of options and facilitate their involvement in the health decision-making process, while taking into consideration cognitive biases, possible information overload, vocabulary, and avoidant coping (Protheroe, Bower, Chew-Graham, Peters, & Fahey, 2007). O'Connor, Graham, and Visser (2005) state that DAs are an invaluable addition to usual clinical care. However, DAs often present incomplete and uncertain information, so care must be taken to create DAs that reduce uncertainty without increasing patient anxiety (Protheroe et al., 2007).

DAs can be extremely helpful, but a DA's effect on decision-making is only as good as the DA itself. Quality of DAs can be measured in terms of completeness of information, clarity, and correctness of presentations, the ability to personalize the DA with the individual patient's condition and preferences, the weighting of short and long-term quality-of-life issues, and the appropriate use of probabilities in computing expected outcomes. We discuss how to evaluate individual DAs later, but look here at effectiveness of different types of DAs.

Many studies have shown that the use of written materials and other static tools to present treatment option information about RA and educate patients is ineffective for many reasons, the most significant of which is low health literacy and numeracy (Berkman et al., 2011; Gigerenzer & Edwards, 2003; Nielsen-Bohlman, Panzer, & Kindig, 2004; Reyna, Nelson, Han, & Dieckmann, 2009; Trevena, Barratt, Butow, & Caldwell, 2006). Health literacy is defined as "the degree to which individuals can obtain, process, and understand the basic health information and services they need to make appropriate health decisions" (Nielsen-Bohlman et al., 2004). Numeracy refers to the degree of one's competency to use numerical information in one or a few short calculations in order to solve a problem (Gigerenzer & Edwards, 2003).

Walker et al. (2007) explored the relationship between health literacy and knowledge gain. They gave patients an arthritis information booklet, accompanied by a "Mind Map" (dramatic words and images to aid cognitive processing). One group received the booklet, while the other received the booklet and the Mind Map. Both groups gained some knowledge, but there was no evidence of the Mind Map helping, regardless of health literacy assessment. Such evidence suggests that we need a different approach to communicate uncertainty and information about treatment effectiveness from clinical trials.

RA patients are faced with decisions for which they often lack information and understanding of the options. The use of DAs can, potentially, help patients make more informed decisions, and thus increase adherence to those decisions. A good DA should convey the possible positive and negative outcomes of treatment, and their likelihoods, in ways that patients can understand. It should enable patients to learn more, perhaps through the use of "What if?" scenarios or other interactions in a serious game. Although the development of DAs is a fast-growing area of medical research, the creation of game-based DAs for arthritis patients, and the study of effective risk communication through interactive tools for patients remain limited.

2.3 Patient Risk Perception

It is generally difficult to process and understand probabilities (Gigerenzer, 1996). Gigerenzer (1996) suggests that human evolution led to the development of cognitive inference machinery in order to adapt to risky situations; however, the format of the information that we use naturally does not come in probabilities or percentages, but absolute frequencies (Gigerenzer, 1996). As an example, it is often easier for people to understand that during a clinical trial, 2 out of 1,000 patients experienced severe side effects. It is harder for them to understand that 0.2 % of the patients will experience those side effects. Inferences are reasoning processes on the basis of sometimes incomplete, circumstantial evidence and prior conclusions. Patients and healthcare practitioners have to make inferences based on numerical data from evidence-based clinical trials. The trial results reveal a need for clinicians to pay closer attention to the educational materials they distribute to patients. Problems can arise when patients must perform inferences from the numerical information provided and some additional knowledge. Often, such inferences depend on Bayes' rule for updating conditional probabilities based on new evidence; this sort of reasoning is hard enough for students and harder for those not used to such calculations.

Research has shown that a salient issue in RA treatment decisions appears to be an extreme focus by patients on some low-probability adverse outcomes. Research participants whose treatment decisions are otherwise sensitive to manipulations of the probability of various adverse outcomes are relatively unmoved by reductions in the reported probability of these "deal breakers"; the risk of cancer, for example, falls in this category (Fraenkel, Bogardus, Concato, & Felson, 2003).

Researchers have tried to calibrate patients' risk perception by explaining probabilities in terms of more familiar events, for example, explicitly comparing an adverse event that occurs to one person in 100,000 as equal to "the probability of dying in a car accident in the next year if you drive 10 miles per week" (Fraenkel et al., 2003). However, such interventions are unlikely to help because patients' estimation of more familiar events can themselves be biased. Car accidents are relatively easy for people to bring to mind, a memory characteristic associated with the overestimation of event probabilities (i.e., the "availability heuristic"), (Tversky & Kahneman, 1973). This example highlights the challenge faced by those trying to help patients develop accurate mental representations of event probabilities.

Patient DAs can increase understanding, but are more effective if structured, tailored, and/or interactive (Schapira, Nattinger, & McAuliffe, 2006; Walker et al., 2007). Walker et al. (2007) have shown that the Arthritis Research Campaign (ARC) booklet increases knowledge in functionally literate patients and the mind maps had no effect. This was contrary to expectations, since the mind map was intended to aid cognition visually through diagrams and images, thus suggesting different approaches need to be considered. Visual graphics that display risk information can aid understanding and supplement counseling by providing information about options and outcomes and by clarifying personal values related to benefits and harms (Schapira et al., 2006). For example, pictographs can be used to represent probabilities in a format that allows one to count icons in a grid, where different colors or versions of the icon represent a probability class and the total number of icons is known. Using pictographs limits biases based on anecdotal information from other patients, effectively communicates medication side effects, and reduces side effect aversion in decision-making (Schapira et al., 2006; Zikmund-Fisher, Fagerlin, Roberts, Derry, & Ubel, 2008). To date, interactive, accessible DAs for RA remain undeveloped.

In short, DAs can be extremely helpful, but a DA's effect on decision-making is only as good as the DA itself. Quality of DAs can be measured in terms of completeness of information, clarity, and correctness of presentations, the ability to personalize the DA with the individual patient's condition and preferences, the weighting of short- and long-term quality-of-life issues, and the appropriate use of probabilities in computing expected outcomes.

3 The Case for Game-Based Decision AIDS

Using entertainment as a motivating factor in health-related education is not new. Roughly 30 years ago, Bosworth et al. (1983) created the Body Awareness Resource Network (BARN), an early computer system designed to provide adolescents with health information on diverse topics such as alcohol use, illegal drugs, sexuality, smoking, and stress management. BARN included several simple games, which the authors enthusiastically describe as "… fast-paced challenges, wild colors and zany sounds…" BARN included a quiz game and a space game. In both cases, players were rewarded based on their choices; "unhealthy" choices resulted in loss of points or aborted missions. Information on how to do better was given after each game session. The authors report users played repeatedly to improve their scores. In a later study using BARN (Bosworth et al., 1983), the authors reported that users who interacted with the system had reduced risk-taking behavior as compared to the control group.

Children and adolescents who have chronic health conditions often feel different from their peers and have low self-esteem due to daily self-care routines and self-monitoring. Lieberman (2001) presents randomized clinical trials of three commercial serious games designed to motivate behavioral changes. Based on social learning theory (Bandura, 1977), children relate to, and are attentive to the game characters when they perceive self-similarities. The three games developed by Click Health Inc. are targeted asthma, diabetes, and smoking prevention. The author found several positive impacts, among which are an impressive drop of 77 % in visits to the emergency room for diabetes care, improved self-efficacy (belief in one's own capability of bringing about specific desirable events and avoidance of undesirable ones (Bandura, 1994) and improved health-related discussions with peers, family, and clinicians.

In a randomized controlled study by Beale et al. (2007), a serious game for adolescents and young adults called "Re-Mission" was used to teach about cancer and its treatment. The game was evaluated with respect to knowledge gained post-intervention. The authors reported an effect of increase in knowledge that is attributable to gameplay. We hope that this finding would extend to game-based RA DAs.

Reichlin et al. (2011) discuss an interactive serious game-based DA titled "Time After Time," intended to help patients with newly diagnosed localized prostate cancer. The goal of the game was to translate evidence-based treatment outcomes into accessible formats that men can incorporate in their decision processes.

The authors identified the importance of a preference elicitation component to help patients build a more accurate representation of their preferences, in order to rank the alternative treatments. Playing cards were used to convey uncertainty of side effects for four treatment options: radical prostatectomy, brachytherapy, external radiotherapy, and watchful waiting. When a card would show up, the user was asked to rank the side-effect on a Likert scale from 1, “no problem” to 5, “big problem.” A “spinner” screen (similar to slot machines) would graphically convey probabilities of side effects. Patients reported positive outcomes after playing the game: clarification of values and preferences, and generating new questions for the healthcare team.

3.1 Evaluating DAs

In order to think about building a good DA, we need to define “goodness” criteria. In fact, such criteria already exist, in the work on evaluating DAs.

Evaluating DAs is a complex process. Most trials focus on the short-term impact of the decision and knowledge enhancements they provide (Protheroe et al., 2007). Protheroe suggests a new approach to evaluating DAs; they focus on the long-term effects in terms of patients’ quality of life. They argue that while a DA might make the decision process longer and more complex, if long-term quality of life is improved, then it can be considered successful as opposed to one that makes the decision process quicker, simpler, and more rewarding at the time when the decision is made, but leads to a decrease in quality of life. Thus, DAs should be designed with long-term effects in mind, with the goal of minimizing decisional conflict and maximizing quality of life in the short, medium, and long term.

Until 2003, there was no initiative to support effective development criteria for patient DAs. In 2003, a group of researchers established International Patient Decision Aid Standards (IPDAS) to enhance the quality and effectiveness of patient DAs through evidence reviews. This group developed an instrument called “International Patient Decision Aid Standards instrument” (IPDASi) (Elwyn et al., 2009) to measure the quality of patient DA interventions and technologies.

In order to evaluate the efficacy and impact of a DA, several patient outcomes should be evaluated across patients in the control and intervention groups. More specifically, we identified the following: (1) patient adherence to prescribed treatment, (2) knowledge about the disease, (3) confidence in decision-making and decisional conflict, and (4) satisfaction with treatment decision. Specific tools that have been used in RA and DA research can also be used for game-based DAs.

Medication adherence is the extent to which patients take medication as prescribed by their healthcare providers. Some factors include: getting prescriptions filled, remembering to take medications on time, and an understanding of directions. A validated instrument called the Medication Adherence Report Scale (Horne & Hankins, 2004) has been shown to be an effective measure of effectiveness in a randomized trial of a DA for osteoporosis (Montori et al., 2011).

An effective DA will increase knowledge about the disease and patients in the intervention group will demonstrate higher levels of knowledge. A tool called the Patient Knowledge Questionnaire (PKQ) (Hill, Bird, Hopkins, Lawton, & Wright, 1991) has been developed and used to measure knowledge acquisition in pre- and post-education programs for RA patients. The 12-item PKQ has established psychometric quality and has been found to be sensitive to change with a statistically significant improvement following patient education sessions.

Decisional conflict refers to personal perceptions of uncertainty in choosing options, feeling uniformed and having a lack of clarity about personal values. The Decisional Conflict Scale (DCS) (O'Connor, 1995) is a 16-item instrument measuring difficulties in decision-making. The total scale has 16 items that measure five subscales. Each item is answered on a 5-point Likert scale. The scale has been used successfully with more than 1,000 adults with different acute and chronic diseases. The DCS has discriminated between patients who accepted or declined treatment for various chronic diseases. Furthermore, the DCS has been shown to be an easily administered instrument with good validity and reliability, with Chronbach's alpha coefficients ranging from 0.78 to 0.92 and test-retest reliability coefficients >0.80 (O'Connor, 1995).

The Satisfaction with Decision Scale (SWD) (Holmes-Rovner et al., 1996) is used to measure patients' satisfaction with treatment decisions. The scale consists of seven items rated on a 5-point Likert scale. The SWD measures satisfaction with the decision and is different from related aspects of satisfaction, including decisional conflict as measured by the DCS scale.

Thus, we argue that an effective RA DA must be designed with these criteria (adherence, knowledge, confidence, and satisfaction) in mind.

4 The Case for a Data-Driven Game

RA treatments have improved with the introduction of methotrexate and biologics. Improvements came with the added risk of potentially severe side effects (Singh et al., 2012). An informed and rational patient's objective should be to reduce pain and improve functional abilities. RA effects functionality in part because it affects joints in the hands of approximately 60–80 % of patients (King & Tomaino, 2001). (Effects on other joints also affect functionality, but we are focused on hands in this chapter.)

Deformities of the hands are typically classified into thumb deformities (Stein & Terrono, 1996), swan-neck, boutonnière, and ulnar deviation deformities. These deformities can occur independently of each other, although some combinations are more likely. Patients typically visit a rheumatologist after radiographic damage (changes visible in radiographs) and functional damage have already occurred. In order to answer their questions about future functionality, we want to know what is the most likely deformity that will develop in a few years, and its likely progression. How will the damage impact their functional abilities to perform daily tasks? Which side effects are the "lesser of the evils"? More research needs to be done before we can claim good answers to these questions.

Until recently (Toyama et al., 2014), there were few longitudinal studies tracking the evolution of morphological changes in the hand (deformities and erosions typically detected in radiographs), and none that systematically evaluated functional abilities. Toyama et al. (2014) conducted a 5-year longitudinal study regarding deformity evolution over time. Their analyses consisted of radiographic evaluation of deformities using the Sharp/van der Heijde scoring system (van der Heijde, 1996), functional evaluation using a goniometer (which measures angles of joints), and several questionnaires. They concluded that, despite treatments, there was a marked progression for patients ($N = 52$) over 5 years. The authors provide details about the frequency of various deformity types and functional losses for their sample population. The radiographic analysis performed in this study was done manually by two trained specialists, a process that takes time and expertise.

While Toyama et al. provide a systematic approach to identify long-term damage in hands due to RA, the time and cost required for a significantly larger sample is prohibitive. A much larger sample is needed to answer questions such as: "How will my hand most likely look like in 2 years if I take drug option A? What if I take drug option B?" Particularly relevant to serious games and simulations was Toyama et al.'s (2014) use of an instrument called JSSH (Japanese Disability of the Arm Shoulder and Hand). The JSSH score is computed based on answers to 20 Likert (0 to 3) questions regarding daily living activities (e.g., Can you hold a glass? Can you turn on a faucet?). We conjecture that predicting JSSH likelihoods given a patient's current state would be a powerful tool for a DA to communicate the effects of the disease as well as treatment uncertainty.

Interactive virtual game-worlds provide a safe arena for patients to investigate possible futures. Patients could control avatars who perform mundane activities (e.g., picking up a glass) with success rates given by predictive models derived from clinical data. Game elements can also be incorporated in simulations. If we can elicit patients' preferences and values, it then becomes possible to display the patient's own decision problem objective function (whatever it is that they are trying to optimize, represented as a numerical function) as the game score in the simulation. During gameplay, patients try to maximize the score, and in the process learn their choice set and objective function. We now present a couple of use cases for a serious game-based DA for RA.

4.1 Use Cases

Use case 1: A hypothetical patient M is a 43-year-old woman who has undiagnosed RA. Her hands have weakened, and her joints frequently ache. Her symptoms make many everyday tasks painful; she often has trouble turning door knobs and buttoning clothing. At the rheumatology clinic, M is diagnosed. The doctor names several treatment options and quickly describes the negative side effects of each medication. M has trouble deciding on a medication since she is unclear on the repercussions of refusing treatment, and the risks, such as hair loss and death, are frightening.

Here, current practice ends. M is left to weigh her alternatives with a pamphlet and her recollection of the office visit.

With a game-based data-driven DA, M could explore the risks and benefits of the proposed treatments given her specific disease stage and its most likely course. With the aid of software that processes information (e.g., radiographs or camera phone images) about how RA has affected M's hands thus far, the DA quickly customizes a hand model. The system can then predict future possible deformities and losses of functionality that might occur given M's specific condition at the time of diagnosis and the likely effects of alternative treatments (or no treatment). The DA can also use the conditional probabilities of adverse events as a function of factors such as age and gender.

M is allowed to explore the different treatments by interacting with by interacting with an on-screen time line that shows the expected decline of hand functionality over time. The graph, in isolation, might mean very little to a patient who has little sense of the outcome scales being represented. The purpose of the interaction is to cognitively calibrate the scales to examples of the impact of the disease on M's life. A secondary, but important, purpose is to provide implicit training in the comprehension of a graphical format that is frequently seen in medical literature. Thus, if M chooses on the line representing treatment at a point 2 years in the future, she will see a virtual hand representing her hand as our model predicts it will appear under these circumstances. Vignettes will depict modal performance, i.e., the statistically average. However, M will also have the option to see best case and worst case scenarios (e.g., one patient in ten who chooses this treatment course will do worse than this; one patient in ten will do better than this). M will be allowed to explore as many of the points on the timelines for each treatment as she wants.

Use case 2: A hypothetical patient X is a 50-year-old man diagnosed with RA 3 years ago, currently on a treatment plan that he believes seems to have little positive effect on slowing the progression or pain. Having had the disease for a number of years, X is fairly well informed about the possible treatment options and is self-describing himself as medically literate. X asks his doctor about the television ads he had seen recently about a class of drugs he recalls are called "biologics." X hopes a new treatment will be more effective, but is also worried about the long list of side effects mentioned in the ads and what he had read on the Internet.

As part of routine care, X has had radiographs of his hands taken at a regular 6 month interval. A game-based data-driven DA has a wealth of example radiographs from patients at various stages of disease progression annotated with the treatment at the time of the radiograph. The tool then computes a most-likely path of progression visualized as a 3D hand model, conditioned on X's radiographs and his choice of new treatment.

X can now explore, at his own pace, possible futures that are predicted based on his own data. He can explore the functional impact of the new treatment by directing the customized hand to perform tasks, such as picking up a cup of coffee or shuffling cards. He is engaged in the simulation, in part because the input "device" consists of his own hands; the hands on screen mimic his actual gestures, except they reflect the effects of the disease in the future. In one instance when he directs the on-screen hand to pick up a cup of coffee, the simulation shows the hand failing by dropping the cup and noting that some percentage of patients at that stage, with his selected choice of new treatment, experience such mundane failures.

5 Technical Challenges and Analytics

Several challenges need be addressed in order to make an effective serious game for RA decision-making. RA patients suffer from pain and hand dexterity, thus standard interaction modalities are often not practical. We address this issue in Sect. 5.1. Game characters or avatars have been shown to be more effective when they physically resemble the player. For RA patients, we believe custom and anatomically accurate hand models are critical for a visceral experience. We address this problem using radiographs in Sect. 5.2. In Sect. 5.3, we address the problem of disease deformity discovery and prediction from patient data.

5.1 Human–Computer Interaction

Since RA causes deformities and pain in the hands, standard interaction modalities used in games (e.g., keyboard and mouse) may not be feasible. We addressed this problem (Mihail, Jacobs, & Goldsmith, 2012) through a gesture recognition system using two Kinect sensors. Our design is robust and easy to set up in a doctor's office environment. The Kinect sensors and hardware required to run the software are inexpensive. The system can be used to perform actions as simple as selecting an item from a menu to navigating an avatar through a 3D world. We describe the system in more detail below.

We use two Kinect sensors that observe a user's hand from two perspectives. This configuration was motivated by self-occlusions that are often alleviated by a multi-view setup. Like standard cameras, the depth-capable Kinect is also likely to observe hands in self-occluding configurations. The two Kinect setup is shown in Fig. 9.1. This setup allows hand configurations such as a palm oriented vertically, to be observed from an angle by two devices (i.e., both sides of the palm are observed), in contrast with a single devices observing only the fingertips.

The two Kinects are calibrated and aligned automatically (simple rigid body transformation) to a world coordinate system where the depth axis points in the middle of the two sensors. Under the assumption that the hand is the nearest observable object, we extract a point cloud that fully contains the hand, as observed by the sensors. See Fig. 9.2 for an example.

Our system achieves rotation invariance by automatically aligning the volume containing the hand with the world depth axis. This is done using principal component

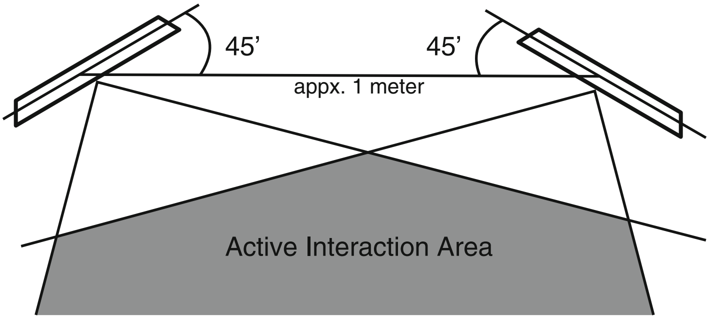
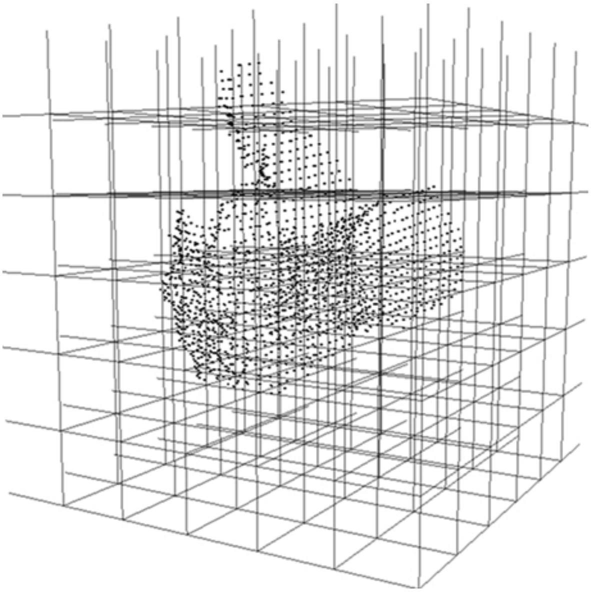


Fig. 9.1 In the above configuration, the sensors observe a user’s hand (*shaded area*) from two perspectives during gesture interaction

Fig. 9.2 The segmented point cloud extracted for a sample “thumbs up” hand configuration



analysis (PCA). We collect a useful by-product of this automatic alignment: the pitch and the yaw of the hand with respect to the world depth axis. In the chapter (Mihail et al., 2012), we show how this information can be used in combination with the detected hand configuration (e.g., pointing the index finger) to guide an avatar in a 3D world.

We demonstrated how a set of hand configurations can be compactly described by dividing the volume containing the hand into a fixed set of voxels and counting the number of points contained in each voxel. Due to the constraints on patient hand configurations from RA deformities, the system is custom trained for each user by asking them to perform a few gestures (e.g., point straight, thumbs-up, etc.)

Analytics: We evaluated the gesture-based interaction system described above with respect to classification accuracy in real time under different motion and rotation conditions for a limited set of users (the authors). The results were highly encouraging, achieving over 90 % accuracy under challenging rotation and motion.

In the future, we are planning to assess the usability of the system in a cohort of RA patients with varying levels of hand deformity and functional ability. Of particular interest is to determine a set of easy to perform gestures for RA patients. This subject is of interest to interdisciplinary research in human–computer interaction (HCI) research and biomechanics.

We perform complex activities using our hands on a daily basis, but most of us give little thought to breaking down each activity into a discrete set of smaller actions, or hand configurations in motion. In serious games using gesture-based interaction (and in augmented reality applications), the way commands are expected and interpreted is important to the overall experience. The example given in the second use case was grabbing a cup of coffee. While significantly less involved than making a cup of coffee, this action can be performed in multiple sequences of gestures (e.g., grabbing the handle with one's index finger and thumb, or the mouth of the cup with all five fingers). Analysis of the actions in a focus group is needed to establish feasibility and usability.

Data collection will consist of Kinect output, namely its 3D point clouds at a real-time frame rate, gesture recognition algorithm results (classification of volumes) augmented by video of patients interacting with the system, and usability questionnaires.

5.2 Arthritic Hand Models

Deformities due to RA are a function of disease activity as well as environmental factors; hence, patients can have different progressions with significant variation in appearance. Using annotated radiographs, we show that it is possible to simulate the deformities on a virtual hand model (Burton, Hallock, & Mihail, 2013).

In a serious game, we believe that a realistically animated avatar hand will be taken more seriously by patients, thus improving the impact of a game-based intervention. In order to accomplish that, we created a hand model using a simplified rig (in this context, rig is a technical term used in computer graphics animations) mapped from bones easily visible in postero-anterior (PA) view radiographs.

In Fig. 9.3, we show a sample model derived from a radiograph of a patient in late stage RA. The simple rig can successfully model subluxation (displacement of bones) and joint space width. This rest pose is applied to off-the-shelf animations, thus simulating a diseased hand at a particular stage learned from the radiograph.

Manually annotating and analyzing hand radiographs is time consuming and requires medical expertise. We developed a method to automatically label the bones in hand radiographs (Mihail, Blomquist, & Jacobs, 2014). The labeled set of bones (hereby referred to as a point distribution model, or more generally, shape) contains

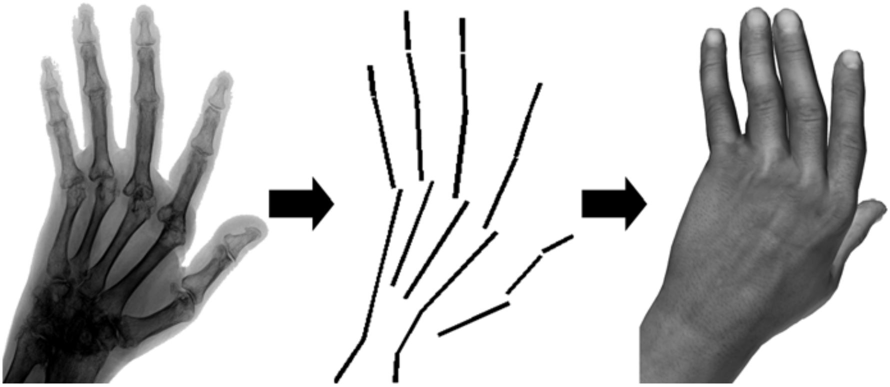


Fig. 9.3 Anatomically inspired rig (*middle*) as applied to a 3D hand model (*right*) from a PA radiograph (*left*)

the positions and orientations of the long bones in a hand, and is robust to capture specific deformities caused by RA, namely joint space narrowing and subluxation. The processing time required for one image is on the order of a few minutes, thus the analysis of a large set becomes feasible.

5.3 Automatic Deformation Discovery and Prediction

Given many RA hand radiographs of different patients and at various points in the progression time, we have shown a mathematical formulation for discovering trends of variance due to the disease, while minimizing variance from anatomical differences in subjects. The machine learning technique we developed is called Disease Stage Metric Learning (DSML) (Mihail, 2014). We show that this model can be used to predict appearance in the future conditioned on one or a series of radiographs. We summarize the mathematical formulation and try to give the reader intuition below.

We consider each radiograph a sample in RA hand shape space, after the transformation using the algorithm outlined by Mihail et al. (2014). Hand shape is represented as a vector s (set of $\{x, y\}$ coordinates, in some order, that specify where each bone is located on a radiograph. Hand shapes of patients with RA differ from healthy hands in nontrivial ways; therefore, we seek a way to model variance due to RA and separate it from healthy, inter-patient variance. Since subluxations and joint space changes can be thought of as translations on a flat radiograph, we assume a linear generative model, where each shape s has the form:

$$s = \mu + \alpha \boldsymbol{I} + \gamma\beta E + \in$$

In the above equation, μ is the average hand, $\boldsymbol{I}$ is a matrix whose columns define an intrinsic variability subspace (e.g., genetic differences lead to anatomical shape variability), the columns of $\boldsymbol{E}$ define an extrinsic variability subspace (e.g., deformations due to disease processes) and ϵ is observational noise. We assume Gaussian distributions for both intrinsic, extrinsic, and noise components (i.e., α, β, ϵ are normally distributed). Here, γ is a positive scalar that controls the disease stage.

Our sample set of RA and non-RA radiographs are annotated by an expert with stage γ as: healthy (numeric stage 0), early (stage 1), moderate (stage 2), and late (stage 3). Our contribution for learning a disease model is to solve for a linear transformation matrix W, that maps samples with stage $\gamma = 0$ to the origin of a latent space (called disease space metric learning, or DSML embedding), and maps samples with $\gamma > 0$ to vectors in latent space with magnitude directly proportional to γ. Previously unseen radiographs can be mapped to the DSML embedding, where the vector magnitude predicts the disease stage. Furthermore, sampling this space can reveal modes of deformation, i.e., answers to questions such as: "what other deformations are likely for patients who exhibit a strong trend for ulnar deviation?"

Analytics: We evaluated DSML on synthetic and patient data. Our evaluation suggests that, in addition to being used to predict disease stage, the resulting embedding can be used to generate samples of future deformities that are likely to occur. Moreover, the DSML embedding can be conditioned on a series of radiographs for a single patient. The implication for serious games is that we can now customize anatomically derived hand models for use in simulations such as a DA, where visceral interactions are critical.

In the future, we will augment this model with treatment data, i.e., when patients are playing the simulation and select treatment X, they will get a more accurate prediction of deformity progression, conditioned on the selected treatment.

6 Conclusions and Future Work

Serious games have found many applications in health care. In this chapter, we described a new category of serious games, game-based decision aids for patient decision support. In the shared decision-making paradigm, patients who suffer from chronic diseases often face difficult treatment decisions. Clinicians try to distill and present patients evidence from clinical drug trials under severe time constraints. This short interaction is augmented with leaflets and pamphlets that are often not enough and leave patients anxious and misinformed. Game-based decision aids have the potential to bridge this communication gap by allowing patients to safely learn more about their choice set, clarify their own preferences and values, and safely explore possible futures in data-drive simulations.

Rheumatoid arthritis is a chronic disease with a relatively slow progression rate. This puts RA patients at a slight "advantage" over sufferers of most other diseases due to the lack of urgency in selecting a treatment that is in tune with their values and preferences. Patients' decision processes take time, and the decisions can have long lasting impact on the patient's long-term quality of life. It is therefore important for providers to make available a quality DA for patients, complementary to traditional patient–clinician interaction. Few RA DAs are available at the time of this writing. Those that do exist are lacking in key areas of effective probability communication and evidence citing. As noted above, DAs are more effective if structured, tailored, and/or interactive (Schapira et al., 2006; Walker et al., 2007). Very few RA DAs now are interactive.

Medical data in the form of imagery and clinical trials can be leveraged to build powerful models of disease progression. These models can be incorporated into serious game-based DAs that allow patients to ask difficult questions, e.g., "What will my hand look like in X years on treatment plan Y?" While significant research efforts have led to improved risk communication strategies, more work is needed to embed them into game-based DAs.

We proposed a game-based decision tool for patients with RA that uses medical imagery to generate visualizations and predictions useful for patients contemplating difficult decisions with uncertain future. This work is based on existing research in medical decision-making under uncertainty. When patients plan, their understanding of the disease and treatment options is critical to a successful treatment (e.g., compliance and adherence in light of negative side effects or long onset time of positive effects). We strongly believe that thorough understanding and testing of each component used in the DA is important to the success of our long-term plan to build a functional game-based DA.

We have described our work in human–computer interaction (HCI) that helps patients with deformities of the hands interact with a machine. We presented the analytics and plans for future evaluation. In the area of hand-based interaction, there are still core challenges to be addressed. For example, muscle fatigue and the effects of system response time can drastically affect usability, which has the potential to negatively impact the medical decision-making outcomes. We will explore these questions in future studies.

We have shown an application of our work in medical image analysis to create anatomically realistic diseased hand models for use as a visual aid in a game-based DA. We described the analytics performed and acknowledged the limitations of our approach. In particular, radiographs are projections of volume onto a plane, but are attractive because they are routinely ordered and inexpensive. While standard graphics animation techniques (e.g., skeletal subspace deformation) when constrained by our disease model can produce satisfactory results, this is still an area of open inquiry. For example, one common symptom of RA is swelling of the joints. Swelling is not easily visible in radiographs. We will continue to work with radiologists in an effort to develop automated methods to detect and quantify swelling.

Finally, we described a method that uses radiographs from different patients to build a model of appearance variation in radiographs due to RA stage. We believe that the ability to show patients predictions conditioned on their own hands is a powerful tool. More work has to be done to quantify exactly how beneficial this approach is to patients. The risk communication literature suggests that personalization (of avatars, characters, etc.) is positively correlated with level of involvement. A patient who is actively involved in their health care will likely benefit more than a passive observer.

References

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (1994). *Self-efficacy*. New York: Wiley.

Beale, I. L., Kato, P. M., Marin-Bowling, V. M., Guthrie, N., & Cole, S. W. (2007). Improvement in cancer-related knowledge following use of a psychoeducational video game for adolescents and young adults with cancer. *Journal of Adolescent Health, 41*(3), 263–270.

Berkman, N. D., Sheridan, S. L., Donahue, K. E., Halpern, D. J., Viera, A., Crotty, K., et al. (2011). *Health literacy interventions and outcomes: An updated systematic review*. Prepared by RTI International-University of North Carolina Evidence-Based Practice Center Under Contract No. 290-2007-10056-I. AHRQ Publication Number 11-E006. Rockville, MD: Agency for Healthcare Research and Quality.

Bosworth, K., Gustafson, D. H., Hawkins, R. P., Chewning, B., & Day, T. (1983). Adolescents, health education, and computers: The body awareness resource network (BARN). *Health Education, 14*(6), 58–60.

Burton, K., Hallock, F., & Mihail, R. P. (2013, July). A data-driven approach to visualize the effects of rheumatoid arthritis on hands. In *2013 18th International Conference on Computer Games: AI, Animation, Mobile, Interactive Multimedia, Educational & Serious Games (CGAMES)* (pp. 188–196). IEEE.

Daltroy, L. H. (1993). Doctor–patient communication in rheumatological disorders. *Baillière's Clinical Rheumatology, 7*(2), 221–239.

Elwyn, G., O'Connor, A. M., Bennett, C., Newcombe, R. G., Politi, M., Durand, M. A., et al. (2009). Assessing the quality of decision support technologies using the International Patient Decision Aid Standards instrument (IPDASi). *PloS One, 4*(3), e4705.

Elwyn, G., O'Connor, A., Stacey, D., Volk, R., Edwards, A., Coulter, A., et al. (2006). Developing a quality criteria framework for patient decision aids: Online international Delphi consensus process. *British Medical Journal, 333*(7565), 417.

Fagerlin, A., Wang, C., & Ubel, P. A. (2005). Reducing the influence of anecdotal reasoning on people's health care decisions: Is a picture worth a thousand statistics? *Medical Decision Making, 25*(4), 398–405.

Feldman-Stewart, D., Brennenstuhl, S., McIssac, K., Austoker, J., Charvet, A., Hewitson, P., et al. (2007). A systematic review of information in decision aids. *Health Expectations, 10*(1), 46–61.

Fraenkel, L., Bogardus, S., Concato, J., & Felson, D. (2003). Risk communication in rheumatoid arthritis. *The Journal of Rheumatology, 30*(3), 443–448.

Gigerenzer, G. (1996). The psychology of good judgment frequency formats and simple algorithms. *Medical Decision Making, 16*(3), 273–280.

Gigerenzer, G., & Edwards, A. (2003). Simple tools for understanding risks: From innumeracy to insight. *British Medical Journal, 327*(7417), 741.

Grove, M. L., Hassell, A. B., Hay, E. M., & Shadforth, M. F. (2001). Adverse reactions to disease-modifying anti-rheumatic drugs in clinical practice. *Quarterly Journal of Medicine, 94*(6), 309–319.

Haugli, L., Strand, E., & Finset, A. (2004). How do patients with rheumatic disease experience their relationship with their doctors?: A qualitative study of experiences of stress and support in the doctor–patient relationship. *Patient Education and Counseling, 52*(2), 169–174.

Hill, J., Bird, H. A., Hopkins, R., Lawton, C., & Wright, V. (1991). The development and use of a patient knowledge questionnaire in rheumatoid arthritis. *Rheumatology, 30*(1), 45–49.

Holmes-Rovner, M., Kroll, J., Schmitt, N., Rovner, D. R., Breer, M. L., Rothert, M. L., et al. (1996). Patient satisfaction with health care decisions the satisfaction with decision scale. *Medical Decision Making, 16*(1), 58–64.

Horne, R., & Hankins, M. (2004). *The medication adherence report scale*. Brighton, UK: University of Brighton.

Ishikawa, H., Hashimoto, H., & Yano, E. (2006). Patients' preferences for decision making and the feeling of being understood in the medical encounter among patients with rheumatoid arthritis. *Arthritis Care & Research, 55*(6), 878–883.

Juul, J. (2012). *A casual revolution: Reinventing video games and their players*. Cambridge, MA: The MIT Press.

King, J. A., & Tomaino, M. M. (2001). Surgical treatment of the rheumatoid thumb. *Hand Clinics, 17*(2), 275–289.

Lieberman, D. A. (2001). Management of chronic pediatric diseases with interactive health games: Theory and research findings. *The Journal of Ambulatory Care Management, 24*(1), 26–38.

Martin, R. W., Brower, M. E., Geralds, A., Gallagher, P. J., & Tellinghuisen, D. J. (2012). An experimental evaluation of patient decision aid design to communicate the effects of medications on the rate of progression of structural joint damage in rheumatoid arthritis. *Patient Education and Counseling, 86*(3), 329–334.

Mihail, R. P. (2014) *Visualizing and predicting the effects of rheumatoid arthritis on hands* (doctoral dissertation). Retrieved from http://uknowledge.uky.edu/cs_etds/19

Mihail, R. P., Blomquist, G., & Jacobs, N. (2014). A CRF approach to fitting a generalized hand skeleton model. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Mihail, R. P., Jacobs, N., & Goldsmith, J. (2012). Static hand gesture recognition with 2 Kinect sensors. In *Proceedings of the 2012 International Conference on Image Processing, Computer Vision, and Pattern Recognition*.

Montori, V. M., Shah, N. D., Pencille, L. J., Branda, M. E., Van Houten, H. K., Swiglo, B. A., et al. (2011). Use of a decision aid to improve treatment decisions in osteoporosis: The osteoporosis choice randomized trial. *The American Journal of Medicine, 124*(6), 549–556.

Nielsen-Bohlman, L., Panzer, A. M., & Kindig, D. A. (2004). *Health literacy: A prescription to end confusion*. Committee on Health Literacy, Board on Neuroscience and Behavioral Health, Institute of Medicine of the National Academies.

O'Connor, A. M. (1995). Validation of a decisional conflict scale. *Medical Decision Making, 15*(1), 25–30.

O'Connor, A. M., Bennett, C. L., Stacey, D., Barry, M., Col, N. F., Eden, K. B., et al. (2009). Decision aids for people facing health treatment or screening decisions. *Cochrane Database of Systematic Review, 3*(3), CD001431.

O'Connor, A. M., Graham, I. D., & Visser, A. (2005). Implementing shared decision making in diverse health care systems: The role of patient decision aids. *Patient Education and Counseling, 57*(3), 247–249.

O'Connor, A. M., Légaré, F., & Stacey, D. (2003). Risk communication in practice: The contribution of decision aids. *British Medical Journal, 327*(7417), 736–740.

Protheroe, J., Bower, P., Chew-Graham, C., Peters, T. J., & Fahey, T. (2007). Effectiveness of a computerized decision aid in primary care on decision making and quality of life in menorrhagia: Results of the MENTIP randomized controlled trial. *Medical Decision Making, 27*(5), 575–584.

Reichlin, L., Mani, N., McArthur, K., Harris, A. M., Rajan, N., & Dacso, C. C. (2011). Assessing the acceptability and usability of an interactive serious game in aiding treatment decisions for patients with localized prostate cancer. *Journal of Medical Internet Research, 13*(1).

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135*(6), 943.

Schapira, M. M., Nattinger, A. B., & McAuliffe, T. L. (2006). The influence of graphic format on breast cancer risk communication. *Journal of Health Communication, 11*(6), 569–582.

Schwab, A. P. (2008). Putting cognitive psychology to work: Improving decision-making in the medical encounter. *Social Science & Medicine, 67*(11), 1861–1869.

Singh, J. A., Furst, D. E., Bharat, A., Curtis, J. R., Kavanaugh, A. F., Kremer, J. M., et al. (2012). 2012 Update of the 2008 American College of Rheumatology recommendations for the use of disease-modifying antirheumatic drugs and biologic agents in the treatment of rheumatoid arthritis. *Arthritis Care & Research, 64*(5), 625–639.

Stein, A. B., & Terrono, A. L. (1996). The rheumatoid thumb. *Hand Clinics, 12*(3), 541–550.

Toyama, S., Tokunaga, D., Fujiwara, H., Oda, R., Kobashi, H., Okumura, H., et al. (2014). Rheumatoid arthritis of the hand: A five-year longitudinal analysis of clinical and radiographic findings. *Modern Rheumatology, 24*(1), 69–77.

Trevena, L. J., Barratt, A., Butow, P., & Caldwell, P. (2006). A systematic review on communicating with patients about evidence. *Journal of Evaluation in Clinical Practice, 12*(1), 13–23.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 207–232.

van der Heijde, D. M. (1996). Plain X-rays in rheumatoid arthritis: Overview of scoring methods, their reliability and applicability. *Baillière's Clinical Rheumatology, 10*(3), 435–453.

Walker, D., Adebajo, A., Heslop, P., Hill, J., Firth, J., Bishop, P., et al. (2007). Patient education in rheumatoid arthritis: The effectiveness of the ARC booklet and the mind map. *Rheumatology, 46*(10), 1593–1596.

Zikmund-Fisher, B. J., Fagerlin, A., Roberts, T. R., Derry, H. A., & Ubel, P. A. (2008). Alternate methods of framing information about medication side effects: Incremental risk versus total risk of occurrence. *Journal of Health Communication, 13*(2), 107–124.

Chapter 10
The Role of Serious Games in Robot Exoskeleton-Assisted Rehabilitation of Stroke Patients

David J. Cornforth, Alexander Koenig, Robert Riener, Katherine August, Ahsan H. Khandoker, Chandan Karmakar, Marimuthu Palaniswami, and Herbert F. Jelinek

Abstract This chapter describes how serious games can be used to improve the rehabilitation of stroke patients. Determining ideal training conditions for rehabilitation is difficult, as no objective measures exist and the psychological state of patients during therapy is often neglected. What is missing is a way to vary the difficulty of the tasks during a therapy session in response to the patient needs, in order to adapt the training specifically to the individual. In this chapter, we describe such a method. A serious game is used to present challenges to the patient, including motor and cognitive tasks. The psychological state of the patient is inferred from measures computed from heart rate variability (HRV) as well as breathing frequency, skin conductance response, and skin temperature. Once the psychological state of the patient can be determined from these measures, it is possible to vary the tasks in real time by adjusting parameters of the game. The serious game aspect of the training allows the virtual environment to become adaptive in real time, leading to improved

D.J. Cornforth (✉)
University of Newcastle, Australia, University Drive, ICT3.12, Callaghan,
NSW 2308, Australia
e-mail: david.cornforth@newcastle.edu.au

A. Koenig • R. Riener • K. August
ETH Zurich, Tannenstrasse 1, E4, Zurich 8092, Switzerland
e-mail: alexander.c.koenig@gmx.de; robert.riener@hest.ethz.ch; kit.august@gmail.com

A.H. Khandoker • C. Karmakar
The University of Melbourne, Electrical Engineering Building, Level 4,
Parkville, VIC 3010, Australia
e-mail: ahsank@unimelb.edu.au; karmakar@unimelb.edu.au

M. Palaniswami
The University of Melbourne, Level 5, 161 Barry Street, Parkville, VIC 3010, Australia
e-mail: palani@unimelb.edu.au

H.F. Jelinek
Charles Sturt University, Building 673, Room 419, Albury, NSW, Australia
e-mail: hjelinek@csu.edu.au

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based
Learning, DOI 10.1007/978-3-319-05834-4_10

matching of the activity to the needs of the patient. This is likely to lead to improved training outcomes and has the potential to lead to faster and more complete recovery, as it enables training that is challenging yet does not overstress the patient.

Keywords Robot exoskeleton • Stroke rehabilitation • Physiological measurements • Closed loop difficulty

1 Background

This chapter describes an application of serious games to stroke rehabilitation. Serious games have been defined in terms of their function as achieving a specific goal by using entertainment as a means to reach that goal rather than forming the primary goal of the game (Rego et al., 2011). Games are now recognized as one of the technologies that can be used to assist with aged care (Vichitvanichphong et al., 2014). Serious games have been applied to rehabilitation as they can provide motivation for patients, to encourage compliance with exercises, and therefore assist recovery (Wiemeyer & Kliem, 2012). When applied to stroke recovery, serious games have been shown to improve engagement (Burke et al., 2009).

Tasks that form part of a game can be finely tuned in terms of difficulty, duration, and repetition. When the game is adopted for a serious or therapeutic purpose, it becomes a serious game. Such serious games incorporating virtual task environments have been used in neurorehabilitation with good outcomes and can provide the basis for improvement or neurorehabilitation practice. The benefits of serious games have shown as transferable to the requirements of activities of daily living in patients with stroke (Adamovich, Fluet, Tunik, & Merians, 2009; Holden, 2005). The degree of active participation and motivation, as a function of task difficulty, has been explored by several authors. This degree of participation is difficult to assess and is often dependent on the interpretation of the physical therapist (Maclean, Pound, Wolfe, & Rudd, 2002). Serious gaming can be a useful adjunct as an edutainment device, which can provide different training scenarios and set an appropriate task difficulty level. Numerous methods exist that try to achieve this in real time. How to provide this adaptation to the patients' needs is the subject of this chapter.

1.1 Rehabilitation

Stroke rehabilitation aims to return patients to an optimal level of activity through a combination of progressively improving motor and cognitive tasks. The goal of rehabilitation is often to reduce motor-related impairments, increase participation in activities of daily living, and improve quality of life. Spontaneous recovery of motor skills after a stroke plateaus at approximately 3 months. However, rehabilitation-based improvements beyond spontaneous recovery have been demonstrated even in chronic stroke patients, inspiring research, and application of long-term therapies.

Determining ideal training conditions for rehabilitation that provides optimal active mental engagement and a physical challenge to individual patients is difficult, as no objective measures exist. Technology-assisted rehabilitation including robot-assisted training and the use of virtual reality (VR) is becoming a practical component of rehabilitation, with increasing availability and ease of use. Combining robot-assisted technology with VR can address mental engagement and increasing repetitions, as well as greatly increasing the reach and effectiveness of today's healthcare system (Saposnik et al., 2010).

Gait is one area often affected by stroke, requiring robot-assisted devices linked with a treadmill where patients retrain to walk. Robot-assisted treadmill training is an established intervention to improve motor function and walking ability in neurologically impaired patients. Treadmill training is an established treatment for gait rehabilitation in neurological patients such as stroke survivors, or people with spinal cord or traumatic brain injury (Dietz & Duysens, 2000; Harro et al., 2014). To improve the rehabilitation outcome in those patients, an increasing number of Driven Gait Orthosis (DGO) devices are available, which automate gait training; among them are the Lokomat, the Autoambulator, the LOPES, and the Gaittrainer (Arazpour et al., 2014; Colombo, Joerg, Schreier, & Dietz, 2000; Stauffer et al., 2009; Veneman et al., 2007; Winchester & Querry, 2006).

The work described in this chapter used the Lokomat (Hocoma, Switzerland). Parts of this work have previously been published in Koenig et al. (2011). The Lokomat assists locomotion at the knee and ankle joint as well as providing support for foot drop. The device allows assisted locomotion on a treadmill by guiding the participant's legs along a predefined trajectory. The training is assisted by providing the participant with an interactive serious game facilitated by video screens placed in front of the Lokomat, as shown in Fig. 10.1.

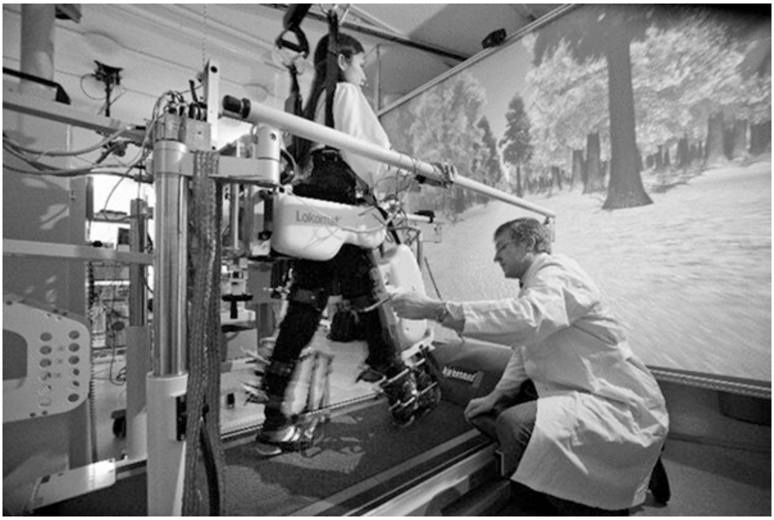


Fig. 10.1 The experimental setup described in this chapter, showing the Lokomat and video screens used for the interactive serious game

Training involves learning motor skills associated with balance and walking. Ideally, these activities should be challenging to the patients, yet not too challenging, as this can lead to increased stress and despair. Learning and exercise can be exceptionally challenging for stroke patients (Jelinek et al., 2011). It is important to ensure that rehabilitation and exercise are safe for the patient and at the same time, it is essential to select a suitable training protocol that brings the most benefit in terms of improved functionality. Measures that can inform the clinician of any risk for the individual, such as a sluggish or rapid physiological or psychological response to abrupt changes or specific challenges, would improve safety for the patient. Active biomechanical engagement of the patients in rehabilitation training has been shown to be an important factor for successful rehabilitation results (Lotze, Braun, Birbaumer, Anders, & Cohen, 2003); the patient's biomechanical effort can be quantified by torque and force sensors. This information is used to assess the patient's level of physical activity (Banz, Bolliger, Colombo, Dietz, & Lunenburger, 2008).

1.2 Psychological State

Despite the fact that attention has been shown to play a role in training success, psychological responsiveness to task difficulty and motivational levels at task onset have not been routinely measured. Active mental engagement and a positive emotional state are the prerequisites for optimal learning in rehabilitation programs of stroke patients. In order for patients to realize the full potential of rehabilitation, it is important to establish and maintain an ideal motor learning situation combining appropriate cognitive, emotional, and physical aspects of the training (Adamovich et al., 2009; Meyer, Peters, Zander, Scholkopf, & Grosse-Wentrup, 2014).

This chapter describes an approach to automatically estimate and classify a patient's psychological state, i.e., his/her mental engagement, in real time, during gait training. In this work, participants engaged in a virtual task with varying difficulty levels that was shown to induce a feeling of being bored, excited, and overstressed.

1.3 Measurement of Psychological State

To obtain an objective measure of the current psychological state, psychophysiological measurements of heart rate (HR), breathing frequency, skin conductance, and skin temperature may be used. Skin conductance responses and skin temperature can be used as markers for psychological states in the presence of physical effort induced by walking. In stroke patients, there are often impaired responses in the involuntary nervous systems (also known as the autonomic nervous system), which are responsible for regulation of breathing, heart rate, swallowing, and other bodily functions. The sympathetic branch of the system is responsible for arousal or the "fight-or-flight" response, while the parasympathetic or vagal is responsible for resting and feeding. Generally, sympathetic activity increases heart rate and decreases variability, whereas parasympathetic activity decreases heart rate and increases variability (Berntson et al., 1997). These can be assessed using heart rate variability (HRV) and interpreted to provide information on the relative emotional state of the patient (Gunther, Witte, & Hoyer, 2010). Acute stroke leads to increased HR but lower variability, with some improvement over time post-stroke (Lakusic, Mahovic, Babic, & Sporis, 2003). Although the magnitude of HRV is influenced by differences in underlying illness, injury, or a result of therapies including medications, HRV reflects adaptation of the organism to physical, cognitive, and emotional conditions (Thayer, Hansen, Saus-Rose, & Johnsen, 2009). HRV has been shown to correlate well with EEG measures of cognitive involvement as well as to motivation levels determined using a psychological test battery (Andreassi, 2007).

In practice, there is always some variability in the HR, due to imbalances in the activity levels of the sympathetic and parasympathetic nervous systems. Hence, any HR cannot increase or decrease indefinitely but instead will be followed by an opposite trend. The speed at which the HR increases or decreases is variable, which implies that the periods of increasing or decreasing inter-beat intervals are also not equal. As a result, heart rate asymmetry (HRA) should be a common phenomenon present in the healthy heart (Jelinek et al., 2014; Porta et al., 2008). The asymmetry in consecutive beat-to-beat intervals can be represented by three different asymmetry indices, namely Guzik's, Porta's, and Ehlers' index (Porta et al., 2008) and indicates sympatho-vagal balance (Jelinek et al., 2011).

In this work, HRA was calculated from the percentage index (PI, defined in Sect. 2.2, Eq. (10.1)) of the normalized probability of the accelerations and decelerations within the time series. Heart periods (inverse of HR) are either shorter or longer on a beat-by-beat basis following acceleration associated with sympathetic activation and inhibition due to vagal activation.

Signals from the Autonomic Nervous System (ANS) that could indicate mental engagement are primarily signals that respond to mental stress or relaxation (Andreassi, 2007). However, in addition to psychological processes, physical effort, such as walking on a treadmill, can influence the psychophysiological measurements. Physiological effort and psychological stress have an influence on the short-term variation of HR. HRV was shown to decrease during physical effort (de la Cruz Torres, Lopez, & Orellana, 2008) and mental stress (Delaney & Brodie, 2000). Galvanic skin response is used as a direct measure for arousal (Hirshfield et al., 2014). From the galvanic skin response, skin conductance response (SCR) and skin conductance level (SCL) are computed. The SCR, measured as a number, is a sensitive indicator for emotional strain (Brown & Macefield, 2014). In recent research, SCL was found to increase during demanding tasks compared to a rest period (Dawson, Schell, & Filion, 2008). The breathing frequency was found to increase during stress (Suess, Alexander, Smith, Sweeney, & Marion, 1980) and mental effort (Carroll, Turner, & Prasad, 1986) and also during physical activity (Doberenz, Roth, Wollburg, Maslowski, & Kim, 2011; Mackersie & Cones, 2011). Skin temperature decreased during mental work stress in a study by Ohsuga, Shimono, and Genno (2001) but increased with physical activity (Mancuso & Knight, 1992).

The challenge is how to combine these various indicators in order to provide one variable that describes mental engagement. To solve this problem, one can turn to the methods offered by machine learning.

1.4 Machine Learning

Machine learning has the potential to assist with the identification of mental engagement in a training scenario. Machine learning treats a variety of problems including supervised learning, or classification, which is the concern of this work. Here, the key is to determine some relationship between a set of input vectors that represent measurements, and a corresponding set of values on a nominal scale that represent category or class. The relationship is obtained by applying an algorithm to training samples that are 2-tuples $\langle \boldsymbol{u}, z \rangle$, consisting of an input vector $\boldsymbol{u}$ and a class label z. The learned relationship can then be applied to instances of $\boldsymbol{u}$ not included in the training set, in order to discover the corresponding class label z (Dietterich & Bakiri, 1995). A variety of techniques, including Artificial Neural Networks (Duda, Hart, & Stork, 2012), have been shown to be very effective for solving such problems. A related problem is to determine the optimum set of measures, by selecting from those available, in order to maximize the performance of the classifier.

A neural network is a mathematical abstraction of some function of neural tissue, but greatly simplified and used to form a complex model relating dependent and independent variables. It can be regarded as a kind of piecewise linear regression model (Nguyen & Widrow, 1990). The problem described in this chapter that requires the use of machine learning is how to determine the current state of mental engagement, without having to continuously administer a questionnaire. This is difficult within a training environment because of the movement required by the participant, as well as the concentration required for the task. The inputs that are available are measurements of HR, breathing frequency, skin conductance, and skin temperature. Is it possible from these to determine a classification of engagement as baseline, under-challenged, challenged, or over-challenged? In the work described here, a neural network was used to investigate the possibility of identifying the current state of mental engagement directly from physiology. This is necessary in order that the video game can be adjusted to vary the difficulty of the task without stressing the patient.

1.5 Implementation

Experiments used a definition of three different levels of mental engagement according to the circumplex model of affect (Koenig, 2011; Russell, 1980) (Fig. 10.2), in which emotions are defined by two dimensions: valence (ranging from unpleasant

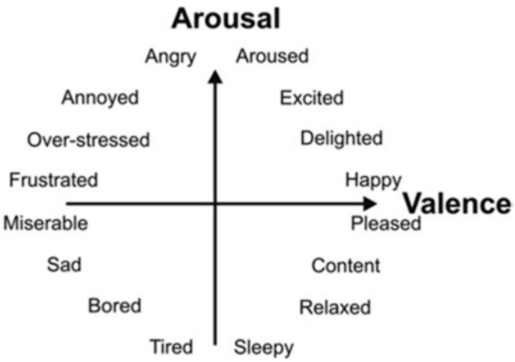


Fig. 10.2 Levels of mental engagement according to the circumplex model of affect

to pleasant) and arousal (ranging from deactivation to activation). The virtual environments were used during robot-assisted gait training to induce different levels of mental engagement in participants.

In the present state of the art, mental engagement of participants is quantified via questionnaires—motivation for example can be quantified via the "Intrinsic Motivation Inventory" (McAuley, Duncan, & Tammen, 1989). During gait rehabilitation, questionnaires are not appropriate for continuous, objective assessment of the psychological state of the patient. In addition, neurological patients with severe cognitive deficits or aphasia (language disorder) might not be able to understand and respond appropriately to the questions.

Here, the goal is to determine if a patient is mentally engaged during the training in order to maximize motor learning during rehabilitation. From motor learning theory, it is known that the learning rate is maximal at a task difficulty level that positively challenges and excites participants while not being too stressful or boring (Guadagnoli & Lee, 2004). A task which is too easy for the participant will be perceived as boring, a task which is too difficult will overstress the participant, while an optimally challenging task should induce maximal mental engagement and optimal physical participation.

The new approach described here focused on measuring the activity of the ANS, as real-time measurement and analysis of signals from the Central Nervous System (CNS) during walking in a robotic device are in general not feasible due to noise and motion artifacts.

2 Experimental Protocol

The studies described in this chapter were conducted at two locations. Measurements with healthy participants were conducted at the Spinal Cord Injury Center Balgrist, Zurich, Switzerland. Measurements with participants who experienced stroke were conducted at the Neurologische Klinik Bad Aibling, Germany. The latter group suffered from neurological gait impairment due to their illness. All participants were selected and approved for participation in the study by a clinical expert to ensure that the participants were able to follow the instructions and respond accordingly. Approval for both studies was obtained from local ethics committees, and all participants or their legal representative gave written informed consent before data collection.

All participants were made secure in the machine and given 30 % body weight support. The feet of the participants were passively lifted by elastic foot straps to prevent foot drop. The speed of the system was kept at a constant 2 km/h and cadence adjusted to suit each participant. The virtual environment was projected on to screens and auditory signals projected to an appropriate speaker system. Participants were challenged to differing degrees during navigation through the virtual environment.

2.1 Task

Participants were given serious game objectives that required a simultaneous mechanical and cognitive response. The mechanical task was to pick up items by walking to them, which involved a change of walking direction in the virtual environment. To change the walking direction, participants had to perform an active push-off. To turn left, the participants had to increase activity in the right leg during stance (Zimmerli, Duschau-Wicke, Mayr, Riener, & Lunenburger, 2009). As the required physical effort to change walking direction was set individually, the challenge was to navigate through the virtual environment and collect items. In order to provide for different task difficulty levels, the distance between virtual items involved in the task was adjustable. Furthermore, the distance between the barrels as well as their speed was adjustable. The cognitive task was to jump over barrels which rolled towards subjects as they walked. As it was not possible for stroke patients to physically jump, barrels were negotiated by clicking a computer mouse button. Points were scored for each collected item and subtracted if items were not collected or not jumped over.

Participants were given an initial training session, then completed five levels of the game: standing with harness; walking with harness; and three levels of difficulty/challenge while walking. The difficulty was related to the distance between barrels and the speed at which they were moving (Koenig et al., 2011). In the first difficulty level participants were under-challenged, as all objects were easily collected without major changes in the walking direction. In the second difficulty level, participants were challenged by varying the distance settings between items and their distribution, so that only 80–90 % of objects could be collected. In the third difficulty level, participants were over-challenged as objects were distributed to reduce score to less than 10 % of the maximum possible score. These three challenge levels correlated with three levels of mental engagement and cortical response: under-challenge is equivalent to boring, correctly challenged is equivalent to excitement and over-challenged is synonymous with the feeling of being overstressed.

Task difficulty was specific to each participant and determined during the initial training session using the Self-Assessment Manikin questionnaire (SAM) (Bradley & Lang, 1994). The SAM is used to measure emotional response to different stimuli (Bradley & Lang, 1994), in particular the emotional response arousal and valence. This allows an individualized program that increases in task difficulty to be developed and the extent of engagement/emotion/stress experienced by the participant to be determined using HRV analysis (Koenig et al., 2011; Morris, 1995). In this work, SAM was used to verify the hypothesis that the three conditions in the virtual task really resulted in a feeling of boredom, excitement, and of being overstressed.

Participants were asked to respond to a 5-point scale by selecting a number which best represented their current emotion. The value of 1 represented the lowest valence ("unhappy") and arousal ("sleepy") and 5 represented the highest valence ("very happy") and arousal ("excited") (Morris, 1995). The SAM was administered after each task. This nonverbal, pictorial questionnaire was chosen so as not to disturb the breathing frequency analysis by speaking, and also to reduce the complexity of responding for aphasic stroke patients or patients with cognitive impairments. Results of the SAM were compared to the predictions made by the neural network, in order to assess the accuracy of the latter. Statistical methods are described at the end of Sect. 2.2.

Each subsequent level lasted 5 min. The last minute of each level was regarded as a steady state for purposes of data analysis, and the difference between the last minute of one level and the first minute of the following level was regarded as a transition period.

The sequence of one training session was:

- Practice time: participants became acquainted with the effects of their movements upon the system (controlling the system). The walking speed for the baseline and the balance of the measurement interval was maintained at 2 km/h. The task difficulty levels were set individually for each participant as described above.
- Five minute walking baseline: physiological signals were recorded in the Lokomat with body weight support of 30 % and without the virtual environment tasks.
- Three task conditions in the virtual environment: the three task conditions were arranged in increasing levels of difficulty, each with duration of 5 min. Five minutes was determined as a trade-off between the time required to reach a steady state in the physiological signals and also to keep the exercise portion of the experiment time below 45 min for the participants who are patients, since it has been informally reported by physiotherapists to be the maximum time for patients to exercise in the Lokomat. After the walking baseline and after each scenario, the participants were requested to respond to the SAM. During the questionnaire response time, the virtual environment was turned off.

The entire experiment can therefore be summarized according to these steps:

1. Practice
2. Baseline walking
   SAM

3. Under-challenging task
   SAM
4. Challenging task
   SAM
5. Over-challenging task
   SAM

2.2 *Measurement*

Objective measures (ECG, breathing, skin conductance, and skin temperature) were used to assess different levels of mental engagement. Using a thermistor flow sensor placed underneath the nose, breathing of participants was measured and breathing frequency was computed using a peak detection algorithm. Changes in SCR and skin temperature were also measured.

HR was computed from ECG using a real-time R-wave detection algorithm (adapted from Christov 2004), R wave peaks were determined using the algorithm first suggested by Tomkins (Hamilton & Tompkins, 1986). HRV was computed as a discrete time series of consecutive RR intervals.

Inter-beat variation and complexity was determined from the ECG using time domain measures such as the Root Mean Square of Successive Differences in the RR interval (RMSSD) and the Tone-Entropy (Karmakar, Khandoker, Jelinek, & Palaniswami, 2013). Frequency domain measures included the number of RR intervals in the high frequency band (HFn) (Malik & Camm, 1995), and parameters of the Poincaré plot. The Poincaré plot of the HRV signal is constructed by plotting consecutive points of RR interval time series against the previous RR interval. Figure 10.3

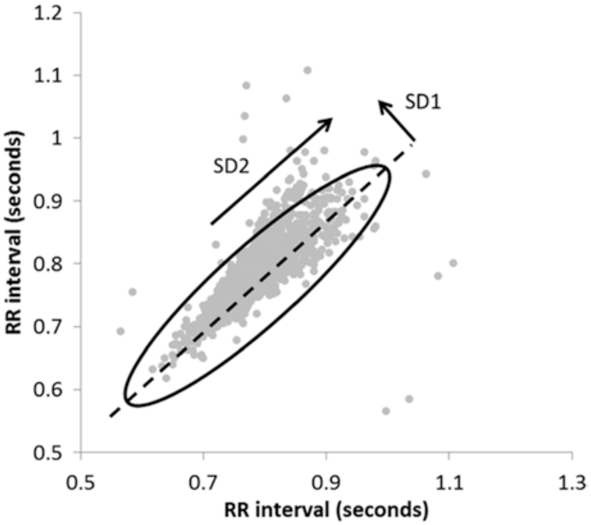


Fig. 10.3 Poincaré plot for a sequence of RR intervals allows the estimation of SD1 and SD2

illustrates this. The technique to quantify the Poincaré plot is to fit an ellipse to the shape of the plot and measure the dispersion along the major and minor axis of the ellipse. SD1 (minor axis) provides a numeric expression of the parasympathetic, that is short-term correlation between inter-beat intervals, whereas SD2 describes the sympatho-vagal balance (Brennan, Palaniswami, & Kamen, 2001).

The Complex Correlation Measure (CCM) measures the variability in the temporal structure of the Poincaré plot, which can characterize or distinguish plots with similar shapes (Karmakar, Khandoker, Gubbi, & Palaniswami, 2011). The CCM measures the point-to-point variation of the signal rather than the gross description of the Poincaré plot. It is computed in a windowed manner, which embeds the temporal information of the signal. A moving window of three consecutive points from the Poincaré plot is considered and the temporal variation of the points is measured. CCM is more sensitive than SD1 and SD2 to changes of parasympathetic activity (Karmakar, Khandoker, Voss, & Palaniswami, 2011).

HRA is determined from the RR intervals as the probability index of accelerations and deceleration of the HR and derived from the calculation of tone and entropy (Khandoker, Jelinek, Moritani, & Palaniswami, 2010). Heart period data or RR intervals are first transformed into the percentage index (PI) time series by:

$$\mathrm{PI}(n) = \left[H(n) - H(n-1)\right] \times 100 / H(n) \tag{10.1}$$

where $H(n)$ is a heart period time series, and n a serial number of heart beats. The tone is defined as a first order moment (arithmetic average) of this PI time series as:

$$\frac{1}{N}\sum_{n} \mathrm{PI}(n) \tag{10.2}$$

where N is a total number of PI terms. The tone represents the balance between accelerations ($\mathrm{PI} > 0$) and inhibitions ($\mathrm{PI} < 0$) of the heart rhythm and represents the sympatho-vagal balance (Amano, Oida, & Moritani, 2005). The entropy is defined from the PI probability distribution by using Shannon's formula:

$$-\sum_{n} p(i)\log_2 p(i) \tag{10.3}$$

where $p(i)$ is a probability that $\mathrm{PI}(n)$ has a value in the range, $i < \mathrm{PI}(n) < i + 1$, where i is an integer. The entropy evaluates total acceleration–inhibition activities, or total heart period variations (Rosenblueth & Simeone, 1984). Acceleration of the HR is, therefore, expressed as a plus difference, inhibition as a minus difference.

From the PI series the positive and negative difference periods were separated. The calculation $\mathrm{PI}(n) = 0$ was omitted because it is neither acceleration nor inhibition. Then, the entropy of the positive and negative differences of the PI time series was calculated. The HR asymmetry is given by the following formula:

$$\mathrm{HRA} = \frac{\text{Entropy of positive difference part of PI time series}}{\text{Total entropy of PI time series}} \tag{10.4}$$

HRA has a value of 0.5 when there is symmetry between the positive and negative parts of the PI time series.

The neural network was used to investigate the possibility of identifying the current state of mental engagement directly from physiology. For automatic classification of mental engagement from physiological recordings, the effectiveness of a neural network was evaluated. As a classifier, a data fitting neural network was trained using the Neural Network Fitting Tool in Matlab (Mathworks, www.mathworks.com), containing 30 hidden layer neurons. The neural network provided an estimation of the current state of mental engagement, based on the physiological recordings. Twenty percent of the data was taken as training data, 20 % as validation and 60 % as testing data. This ensures that the neural network was always tested on unseen data, in order to avoid over-fitting bias in the model, which could lead to overly optimistic results. As neural networks require labeled data during the training phase, the training data was labeled as 1 = "baseline," 2 = "under-challenged," 3 = "challenged," or 4 = "over-challenged." Learning was performed with the Levenberg-Marquardt back-propagation algorithm (Moré, 1977).

2.3 Feature Selection

The attachment of sensors for physiological recordings on the participant's body is time consuming for a clinical application, demanding resources of the therapist, and also reducing the time a participant can exercise in the Lokomat. To improve clinical applicability of this approach, it is important to determine whether all recorded physiological signals are necessary to perform classification of mental engagement or if the recorded data contained information from dependent variables. It could, for example, be possible that HR and breathing frequency would show a strong correlation. In this case, one of these signals could then be omitted in future recordings without degrading classification performance. Principal component analysis (PCA) allows identification of those signals that explain most of the variance in the data (Pearson, 1901). PCA is a mathematical technique that takes a set of measures, some of which could be correlated, and converts them to another, usually smaller, set of measures that are guaranteed to not be correlated. In the process of doing this, the original measures have scores attached to them that describe how much variance they contribute to the overall data set. In the context of feature detection, such measures can be used to select those features that represent less variance so can be omitted with little consequence.

PCA was applied to each participant individually using 5 min of data for each of the four training conditions. Inputs to the PCA were HR, a discrete time series of HRV, a discrete time series of the number of SCR events, SCL, skin temperature, and a discrete time series of breathing frequency. PCA provides combinations of the inputs, where the first axis (or first principal component) explains most of the variance. The second PC explains the second most important variance, etc.

The number of factors k that were necessary to explain more than 80 % of the variance in all participants was computed ($k \in [1, n]$, where n is the dimensionality of original data, i.e., 6). A factor rotation on these first *k* PCs was performed to obtain a clearer picture of which input signals provided the largest variance. Factor rotation is a mathematical transformation that does not alter the subspace spanned by the PCs, but shifts the weight of an input, e.g., from the first PC to the second, while maintaining the orthogonality between the components.

Descriptive statistics were used to investigate which physiological signals changed significantly between the different task level conditions. Only the last minute of each 5 min condition was analyzed to ensure that steady state had been reached. All conditions were tested using the Friedman test followed by a Wilcoxon test for paired comparison. Bonferroni correction corrected multiple errors caused by the paired comparison. The significance level was set at $p<0.05$. The performance of the classifier was assessed by means of the mean squared classification error for two sets of input data. The first set was the six raw physiological data streams, while the second was a reduced data set that contained only the physiological recordings that were dominant in the first *k* PCs ($k<n$).

3 Results

Results for HRA indicated that at rest, stroke patients are anxious. This is indicated by a higher value for HRA, which indicates that the entropy of positive changes in HR has higher entropy than that of the negative changes. The HRA inverts in patients when the tasks increase in complexity, leading to a value less than 0.5, indicating that the entropy of the negative changes in HR dominates during the challenged condition. This high probability of deceleration (0.45 ± 0.07) returns towards a more balanced response during the over-challenged condition (0.47 ± 0.04). The control group showed a steady increase in acceleration, peaking at the challenged condition (0.62 ± 0.17), then returning towards baseline during the over-challenged condition as well.

These differences are shown graphically in Fig. 10.4. It can be seen that the differences between stroke patients and healthy participants are not statistically significant.

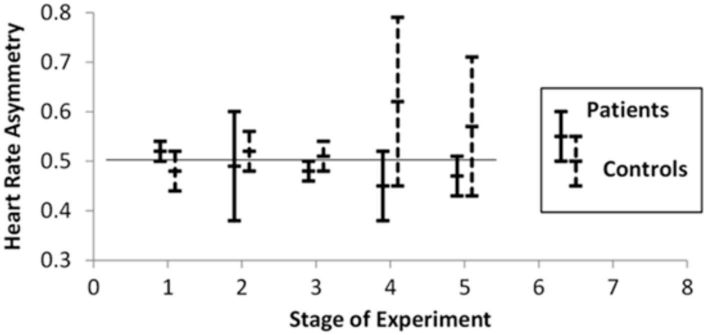


Fig. 10.4 Heart rate variability recorded at different stages of the experiment

Table 10.1 Statistical results of physiological recordings in healthy participants

	HR (bpm)		BF (cpm)		SCR (SCR/min)	
Baseline	73.3^{b,c,d}	CI: 60.3–81.8	21.6^{b,c,d}	CI: 20.3–24.5	0.2^{b,c,d}	CI: 0.0–0.6
Under-challenged	81.3^{a,c,d}	CI: 68.0–91.2	23.0^{a,c,d}	CI: 22.2–26.1	1.0^{a,d}	CI: 0.2–3.7
Challenged	94.1^{a,b}	CI: 77.8–103.0	27.5^{a,b}	CI: 24.9–30.3	3.1^{a}	CI: 0.6–5.3
Over-challenged	96.4^{a,b}	CI: 76.3–102.9	27.8^{a,b}	CI: 25.0–29.4	3.3^{a,b}	CI: 0.4–6.2
	RMSSD (ms)		Skin temperature (°C)			
Baseline	30.0^{c,d}	CI: 19.4–49.9	32.5^{b,c}	CI: 31.7–32.9		
Under-challenged	27.5^{d}	CI: 6.8–37.3	30.8^{a,d}	CI: 29.1–32.1		
Challenged	25.5^{a}	CI: 5.0–60.2	31.5^{a,d}	CI: 30.5–32.5		
Over-challenged	15.3^{a,b}	CI: 4.7–63.0	32.0^{b,c}	CI: 31.3–32.7		

The table shows median and 95 % confidence interval (CI) of heart rate (HR), breathing frequency (BF), skin conductance response (SCR), square root of the mean squared differences of successive normal-to-normal intervals (RMSSD), and skin temperature. Data previously presented in Koenig et al. (2011)

[a] Significant different from the baseline ($p < 0.05$)

[b] Significant different from the under-challenged condition ($p < 0.05$)

[c] Significant different from the challenged condition ($p < 0.05$)

[d] Significant different from the over-challenged condition ($p < 0.05$)

However, the data suggest the difference in control of heart rhythm between the two groups with respect to task condition. In Fig. 10.4, the 0.5 line indicator is the reference value of HRA. Values greater than 0.5 indicate an accelerating influence, whereas values below 0.5 indicate slowing of the HR.

Healthy participants revealed statistically significant differences in several physiological signals (Table 10.1). HR increased significantly from baseline for all conditions with the virtual task and for the conditions challenged and over-challenged compared to the condition under-challenged. The same significant changes were found for breathing frequency. Similar results were also found for the SCR. For all virtual task conditions, the SCR increased significantly compared to baseline. In addition, the SCR increased significantly for the over-challenged condition compared with the under-challenged condition.

RMSSD decreased significantly from baseline for the conditions challenged and over-challenged. Furthermore, a significant decrease was found for the over-challenged condition compared with the under-challenged condition. In the frequency domain, no significant changes were found. A significant decrease was also found in skin temperature. The skin temperature during conditions under-challenged and challenged were significantly decreased when compared with the baseline and the over-challenged conditions.

The results for stroke patients were very different. Compared with the very robust and variable physiological signals in healthy participants, only three significant changes were found in patients. HR increased significantly for the challenged condition (+7.6 %) and for the over-challenged condition (+6.2 %) compared with baseline (median 89.7 bpm, CI: 77.5–103.2). Breathing frequency decreased significantly for the over-challenged condition (−5 %) compared to the challenged condition (median 27.7 cpm, CI: 24.8–29.8). This highlights the need for Principal Components Analysis (PCA) to discover variables that can discriminate these mental states.

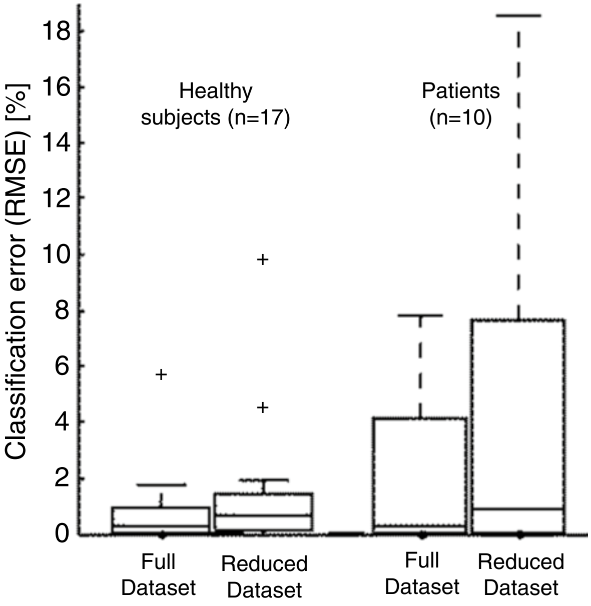


Fig. 10.5 Classification error by means of the root mean square error (RMSE) for all 17 healthy subjects (*left*) and all ten patients (*right*), using all six physiological signals (full data set) and using only the three signals that were dominant in the first three PCs (reduced data set)

PCA was able to directly separate the four conditions baseline, under-challenged, challenged, and over-challenged, in spite of the fact that the statistical analysis of the physiological data was unclear.

Evaluation of the results of the PCA revealed the importance of skin temperature and SCL in both groups. HRV played a lesser role. Breathing frequency was not dominant in healthy participants, but it was for patients.

Although only two significant differences were found in all physiological recordings over all conditions of patient data, the classification of the different psychological states using a neural network was possible for healthy participants and also for those who had suffered from stroke. The classification results were evaluated for a neural network for two different sets of input data: on the one side with six physiological parameters extracted, on the other side using only the physiological signals dominant in the first three PCs. Mean classification error was 1.4 % for the full and 2.5 % for the reduced data set in healthy participants and 2.1 % for the full and 4.7 % for the reduced data set for patients (Fig. 10.5).

3.1 Changes During the Training Session

Turning to an examination of how some of these measures change during the training session, we examine the four transitions in healthy participants:

- Transition 1—standing to walking
- Transition 2—walking to under-challenged
- Transition 3—under-challenged to challenged
- Transition 4—challenged to over-challenged

Table 10.2 Adaptation to transitions, as manifested in changes of different HRV parameters

Stages	SDNN	RMSSD	HFn	SD1	CCM
Transition 1	0.40 ± 0.07	0.57 ± 0.12	0.5 ± 0.05	0.24 ± 0.08*	0.71 ± 0.12*
Transition 2	0.44 ± 0.07	0.51 ± 0.13	0.57 ± 0.08	0.27 ± 0.08	0.70 ± 0.09
Transition 3	0.59 ± 0.07	0.56 ± 0.08	0.67 ± 0.1	0.39 ± 0.10*	0.81 ± 0.07*
Transition 4	0.46 ± 0.06	0.61 ± 0.09	0.49 ± 0.09	0.37 ± 0.09	0.74 ± 0.13

All values are mean ± standard error. Transitions between standing to walking and between under-challenged to challenged show significant difference at $p < 0.05$, and so are indicated by an asterisk

Table 10.2 shows the result of investigating the level of adaptation, which is an important component in task performance. The HRV parameters shown are:

- SDNN—standard deviation of RR intervals
- RMSSD—square root of the mean squared differences of successive RR intervals
- HFn—normalized high frequency power
- SD1—Poincare short-time correlation parameter
- CCM—complex correlation measure of Poincare plot

Only the nonlinear measures, SD1 and CCM differentiated between the level of adaptation. Specifically, only transition 1 and transition 3 were significant, which is found from calculating the gradient of changes for those parameters.

4 Discussion

Experiments using a serious game to enhance robot-assisted treadmill training in healthy participants and stroke patients, yielded data, and objective measurements that were fully sufficient to detect the current psychological state of a participant. This is an important result, since one would question whether it is possible to detect psychological state from such simple physiological measures as those used. This work suggests that this is indeed the case. The difficulty of a virtual task during the rehabilitation training was set as under-challenging, challenging, and over-challenging to induce a feeling of boredom, challenge, or stress. During training the ECG, breathing frequency, skin temperature, and the galvanic skin response were recorded. The psychological state of a participant could be classified using a combination of PCA with a neural network. Variables chosen were HR, SCL, and skin temperature, which can be used as markers for psychological states in the presence of physical effort.

Evaluation of questionnaires from healthy participants confirmed that virtual tasks of different difficulty levels can indeed induce, or result in, different levels of mental engagement, i.e., of being bored, challenged, or overstressed. Also in healthy participants, descriptive statistics suffice to distinguish between different levels of mental engagement. The automatic classification worked in all but two participants with less than 2 % classification error. This is a promising result that provides confidence in the ability to deduce mental engagement from simple measures.

In patients, neither questionnaires nor physiological signals showed a picture as clearly as in healthy participants. While therapists anecdotally reported that the virtual task bored, challenged, or overstressed the participant, the questionnaires did not confirm this observation. A possible explanation of the results showing lower clarity from questionnaires is that patients with cerebral lesions might suffer from cognitive deficits, which might prevent them from assessing, expressing, and verbalizing their level of mental engagement during rehabilitation. This was consistent with reports by the therapists. In addition, it was reported that patients did not usually admit if they experienced the task as too difficult because they were very ambitious and determined to solve the task successfully. Although the questionnaire showed only few changes, the predefined scores (success rate of 100 % for the under-challenged condition, of 80–90 % for the challenged condition and of 10–20 % for the over-challenged condition) were achieved by every patient. Another explanation might be the fact that walking with the help of the DGO was an experience that is positively perceived by patients who may be otherwise unable to walk well on their own. The significance of this finding is that the difficulty in identifying mental engagement by questionnaire might be overcome if the appropriate numerical measures are used.

Despite the heterogeneous and unclear nature of the picture in the descriptive statistical analysis of the physiological data of patients, the classification of the various conditions in the virtual environment was possible with less than 8 % classification error in all patients. In this context, a real-time automatic classification algorithm applied to physiological recordings seems to allow an objective estimation of mental engagement for the benefit of the patient in clinical applications, and in particular, when compared with the sometimes conflicting and often unreliable subjective information obtained from other sources. This provides a pathway to move beyond subjective measures and to allow reliable assessment of mental engagement through numerical means.

PCA was used to identify the minimum set of physiological signals that would be necessary to perform classification of mental engagement while not degrading the performance of the classifier. In both, healthy participants and participants who suffered from neurological injuries, skin temperature, and SCL were the main psychophysiological responders to this intervention. Also, HRV did not contribute significantly to the first three PCs. As discussed above, HRV might have been reduced due to the physical effort involved in walking. It might be possible that more advanced analysis based on HRV will provide measures that will have more success in the automatic classification of mental engagement.

Previously, virtual environments in rehabilitation did not provide patient-specific and adaptive features, with the ability to adjust levels of difficulty that correlated with mental engagement. The long-term goal of this work is to perform closed loop control of psychological states during robot-assisted rehabilitation—to objectively determine the psychological state of the patient, and then by automatically or by enabling the therapist to adjust the attributes of the system, and to stimulate a desired level of mental engagement during rehabilitation. The goal of approaching optimal mental engagement during exercise in rehabilitation is consistent with evidence that attention to task and mental engagement improve outcomes of rehabilitation efforts for patients in the training of motor skills. Previously, researchers have used closed loop control of psychophysiological measurements to control stress.

This approach is not limited to a particular gait orthosis, and not even to rehabilitation of the lower limbs. In robot-assisted arm rehabilitation, as performed with the ARMin (Nef, Guidali, & Riener, 2009; Nef, Mihelj, & Riener, 2007), the HapticMaster (Houtsma & Van Houten, 2006), or the MIT Manus (Aisen, Krebs, Hogan, McDowell, & Volpe, 1997), the absence of physical effort induced by walking might even improve the results obtained from healthy as well as neurologically impaired participants.

5 Conclusion

A stroke can be a devastating event in a person's life, leading to severe loss of mobility, cognitive impairment, the inability to participate in various activities of daily living, and associated loss of independence. The goal of rehabilitation is often to reduce motor-related impairments beyond spontaneous recovery and to increase participation in activities of daily living and improve quality of life. Effective rehabilitation relies upon an appropriate level of physical and emotional engagement during task practice involving learning a motor skill. The use of serious games in this context can be shown to assist in stroke rehabilitation, by providing an environment which can be tailored to meet the needs of each patient. The challenge in the past, to provide a system that is responsive to the changing needs of each patient, has been how to easily and in real time objectively assess physical ability and psychological state of the patient, to dynamically adapt the task in order to suit the current state of the patient, and to do this in a way that does not complicate the scene. The use of real-time physiological measurement has been demonstrated to provide feedback indicating the psychological state in real time, during the training and while the patient is engaged by the serious game. This means that cumbersome questionnaires can be omitted and replaced with feedback from an effective objective measure taken from patients in real time that does not interrupt the training activity or burden the therapist. Consequently, the virtual environment can be adapted in real time to meet the needs of the patient, in order to provide training that is challenging yet does not overstress the patient. In this way, the serious game would enable optimal personalized rehabilitation conditions.

Future work will focus on improving the classifier system so that the patient psychological state can be identified with even more accuracy. There is a large number of classifier algorithms available, and these should be investigated to determine their relative performance. In order to validate the findings, the system will be tested on a larger number of patients. In addition, the use of multiple measurements to determine the psychological state may be unnecessary. The ability to eliminate the measures that are less useful should allow the environment to be simplified, leading to a more accessible training environment that could be adopted more widely in stroke rehabilitation and extended to other areas of physical rehabilitation.

It is our aim that through these advances, a generic training framework using a serious game would become mainstream and enhance rehabilitation options available, meeting the needs of the largest possible number of patients with safe, effective, personalized therapies.

References

Adamovich, S. V., Fluet, G. G., Tunik, E., & Merians, A. S. (2009). Sensorimotor training in virtual reality: A review. *NeuroRehabilitation, 25*(1), 29–44.

Aisen, M. L., Krebs, H. I., Hogan, N., McDowell, F., & Volpe, B. T. (1997). The effect of robot-assisted therapy and rehabilitative training on motor recovery following stroke. *Archives of Neurology, 54*, 443–446.

Amano, M., Oida, E., & Moritani, T. (2005). Age-associated alteration of sympatho-vagal balance in a female population assessed through the tone-entropy analysis. *European Journal of Applied Physiology, 94*, 602–610.

Andreassi, J. L. (2007). *Psychophysiology: Human behavior and physiological response* 5th ed. Mahwah, NJ: Lawrence Erlbaum.

Arazpour, M., Mehrpour, S. R., Bani, M. A., Hutchins, S. W., Bahramizadeh, M., & Rahgozar, M. (2014). Comparison of gait between healthy participants and persons with spinal cord injury when using a powered gait orthosis—A pilot study. *Spinal Cord, 52*(1), 44–48.

Banz, R., Bolliger, M., Colombo, G., Dietz, V., & Lunenburger, L. (2008). Computerized visual feedback: An adjunct to robotic-assisted gait training. *Physical Therapy, 88*, 1135–1145.

Berntson, G. G., Bigger, J. T., Jr., Eckberg, D. L., Grossman, P., Kaufmann, P. G., Malik, M., et al. (1997). Heart rate variability: Origins, methods, and interpretive caveats. *Psychophysiology, 34*(6), 623–648.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry, 25*(1), 49–59.

Brennan, M., Palaniswami, M., & Kamen, P. (2001). New insights into the relationship between Poincare plot geometry and linear measures of heart rate variability. In *Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Vol. 1, pp. 526–529).

Brown, R., & Macefield, V. G. (2014). Skin sympathetic nerve activity in humans during exposure to emotionally-charged images: Sex differences. *Frontiers in Physiology, 5*, 111.

Burke, J. W., McNeill, M. D. J., Charles, D. K., Morrow, P. J., Crosbie, J. H., & McDonough, S. M. (2009). Optimising engagement for stroke rehabilitation using serious games. *The Visual Computer, 25*(12), 1085–1099.

Carroll, D., Turner, J. R., & Prasad, R. (1986). The effects of level of difficulty of mental arithmetic challenge on heart rate and oxygen-consumption. *International Journal of Psychophysiology, 4*, 167–173.

Christov, I. (2004). Real time electrocardiogram QRS detection using combined adaptive threshold. *Biomedical Engineering Online, 3*, 8.

Colombo, G., Joerg, M., Schreier, R., & Dietz, V. (2000). Treadmill training of paraplegic patients using a robotic orthosis. *Journal of Rehabilitation Research and Development, 37*, 693–700.

Dawson, M. E., Schell, A. M., & Filion, D. L. (2008). *Handbook of psychophysiology* (3rd ed.). New York: Cambridge University Press.

de la Cruz Torres, B., Lopez, C. L., & Orellana, J. N. (2008). Analysis of heart rate variability at rest and during aerobic exercise: A study in healthy people and cardiac patients. *British Journal of Sports Medicine, 42*(9), 715–720.

Delaney, J. P. A., & Brodie, D. A. (2000). Effects of short-term psychological stress on the time and frequency domains of heart-rate variability. *Perceptual and Motor Skills, 91*, 515–524.

Dietterich, T. G., & Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research, 2*, 263–286.

Dietz, V., & Duysens, J. (2000). Significance of load receptor input during locomotion: A review. *Gait & Posture, 11*(2), 102–110.

Doberenz, S., Roth, W. T., Wollburg, E., Maslowski, N. I., & Kim, S. (2011). Methodological considerations in ambulatory skin conductance monitoring. *International Journal of Psychophysiology, 80*(2), 87–95.

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification and scene analysis*. London: Wiley.

Guadagnoli, M. A., & Lee, T. D. (2004). Challenge point: A framework for conceptualizing the effects of various practice conditions in motor learning. *Journal of Motor Behavior, 36*, 212–224.

Gunther, A., Witte, O. W., & Hoyer, D. (2010). Autonomic dysfunction and risk stratification assessed from heart rate pattern. *Open Neurology Journal, 4*, 39–49.

Hamilton, P. S., & Tompkins, W. J. (1986). Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database. *IEEE Transactions on Biomedical Engineering, BME-33*, 1157–1165.

Harro, C. C., Shoemaker, M. J., Frey, O., Gamble, A. C., Harring, K. B., Karl, K. L., et al. (2014). The effects of speed-dependent treadmill training and rhythmic auditory-cued overground walking on balance function, fall incidence, and quality of life in individuals with idiopathic Parkinson's disease: A randomized controlled trial. *NeuroRehabilitation, 34*(3), 541–556.

Hirshfield, L. M., Bobko, P., Barelka, A., Hirshfield, S. H., Farrington, M. T, Gulbronson, S., et al. (2014). Using noninvasive brain measurement to explore the psychological effects of computer malfunctions on users during human–computer interactions. *Advances in Human-Computer Interaction.* doi: 10.1155/2014/101038.

Holden, M. K. (2005). Virtual environments for motor rehabilitation: Review. *CyberPsychology & Behavior, 8*, 187–211.

Houtsma, J. A., & Van Houten, F. J. (2006). Virtual reality and a haptic master-slave set-up in post-stroke upper-limb rehabilitation. *Proceedings of the Institute of Mechanical Engineers H, 220*, 715–718.

Jelinek, H. F., August, K. G., Imam, Md. H., Khandoker, A. H., Khalaf, K., Koenig, A., et al. (2014). Influence of stroke location on heart rate variability in robot-assistive neurorehabilitation. In *Proceedings of the 2nd Middle East Conference on Biomedical Engineering* (pp. 253–256). ISBN: 978-1-4799-4799-7/14.

Jelinek, H. F., August, K., Khandoker, A., Issam, H. M., Koenig, A., & Riener, R. (2011). Heart rate asymmetry and emotional response to robot-assist task challenges in post-stroke patients. In *Proceedings of the Computers in Cardiology Conference* (Vol. 38, pp. 521–524).

Karmakar, C. K., Khandoker, A. H., Gubbi, J., & Palaniswami, M. (2011). Defining asymmetry in heart rate variability signals using a Poincaré plot. *Physiological Measurement, 30*, 1227–1240.

Karmakar, C. K., Khandoker, A. H., Jelinek, H. F., & Palaniswami, M. (2013). Risk stratification of cardiac autonomic neuropathy based on multi-scale tone-entropy. *Medical and Biological Engineering and Computing, 51*(5), 537–546. doi:10.1007/s11517-012-1022-5.

Karmakar, C. K., Khandoker, A. H., Voss, A., & Palaniswami, M. (2011). Sensitivity of temporal heart rate variability in Poincaré plot to changes in parasympathetic nervous system activity. *Biomedical Engineering Online, 10*, 17.

Khandoker, A. H., Jelinek, H. F., Moritani, T., & Palaniswami, M. (2010). Association of cardiac autonomic neuropathy with alteration of sympatho-vagal balance through heart rate variability analysis. *Medical Engineering and Physics, 32*, 61–67.

Koenig, A., Omlin, X., Zimmerli, L., Sapa, M., Krewer, C., Bolliger, M., et al. (2011). Psychological state estimation from physiological recordings during robot-assisted gait rehabilitation. *Journal of Rehabilitation Research and Development, 48*(4), 367–385.

Lakusic, N., Mahovic, D., Babic, T., & Sporis, D. (2003). Changes in autonomic control of heart rate after ischemic cerebral stroke. *Acta Medica Croatica, 57*(4), 269–273.

Lotze, M., Braun, C., Birbaumer, N., Anders, S., & Cohen, L. G. (2003). Motor learning elicited by voluntary drive. *Brain, 126*, 866–872.

Mackersie, C. L., & Cones, H. (2011). Subjective and psychophysiological indexes of listening effort in a competing-talker task. *Journal of the American Academy of Audiology, 22*(2), 113–122.

Maclean, N., Pound, P., Wolfe, C., & Rudd, A. (2002). The concept of patient motivation. *Stroke, 33*(2), 444–448.

Malik, M., & Camm, A. J. (Eds.). (1995). *Heart rate variability*. Armonk, NY: Futura.

Mancuso, D. L, Knight, K. L. (1992). Effects of prior physical activity on skin surface temperature response of the ankle during and after a 30-minute ice pack application. *Journal of Athletic Training, 27(3),* 242, 244, 246, 248-249.

McAuley, E., Duncan, T., & Tammen, V. V. (1989). Psychometric properties of the intrinsic motivation inventory in a competitive sports setting—A confirmatory factor analysis. *Research Quarterly for Exercise and Sport, 60*, 48–58.

Meyer, T., Peters, J., Zander, T., Scholkopf, B., & Grosse-Wentrup, M. (2014). Predicting motor learning performance from electroencephalographic data. *Journal of NeuroEngineering and Rehabilitation, 11*(1), 24.

Moré, J. J. (1977). The Levenberg-Marquardt algorithm: implementation and theory. In G. A. Watson (Ed.), *Numerical analysis* (Lecture notes in mathematics, Vol. 630, pp. 105–116). New York: Springer.

Morris, J. D. (1995). Observations: SAM—The Self Assessment Mannequin: An efficient cross-cultural measurement of emotional response. *Journal of Advertising Research, 35*(6), 63–70.

Nef, T., Guidali, M., & Riener, R. (2009). ARMin III—Arm therapy exoskeleton with an ergonomic shoulder actuation. *Applied Bionics and Biomechanics, 6*, 16.

Nef, T., Mihelj, M., & Riener, R. (2007). ARMin: A robot for patient-cooperative arm therapy. *Medical and Biological Engineering and Computing, 45*, 887–900.

Nguyen, D., & Widrow, B. (1990). Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In *Proceedings of the IJCNN International Joint Conference on Neural Networks* (Vol. 3, pp. 21–26).

Ohsuga, M., Shimono, F., & Genno, H. (2001). Assessment of phasic work stress using autonomic indices. *International Journal of Psychophysiology, 40*, 211–220.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6, 2*(11), 559–572.

Porta, A., Casali, K. R., Casali, A. G., Gnecchi-Ruscone, T., Tovaldini, E., Montano, N., et al. (2008). Temporal asymmetries of short-term heart period variability are linked to autonomic regulation. *American Journal of Physiology—Regulatory, Integrative and Comparative Physiology, 295*, R550557.

Rego, P., Moreira, P. M., & Reis, L. P. (2011). Serious games for rehabilitation: A survey and a classification towards a taxonomy. In: *2010 5th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1–6).

Rosenblueth, A., & Simeone, A. (1984). The interrelations of vagal and accelerator effects on the cardiac rate. *American Journal of Physiology, 110*, 42–55.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*, 1161–1178.

Saposnik, G., Teasell, R., Mamdani, M., Hall, J., McIlroy, W., Cheung, D., et al. (2010). Effectiveness of virtual reality using Wii gaming technology in stroke rehabilitation: A pilot randomized clinical trial and proof of principle. *Stroke, 41*(7), 1477–1484.

Stauffer, Y., Allemand, Y., Bouri, M., Fournier, J., Clavel, R., Metrailler, P., et al. (2009). The WalkTrainer—A new generation of walking reeducation device combining orthoses and muscle stimulation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 17*, 38–45.

Suess, W. M., Alexander, A. B., Smith, D. D., Sweeney, H. W., & Marion, R. J. (1980). The effects of psychological stress on respiration—A preliminary-study of anxiety and hyperventilation. *Psychophysiology, 17*, 535–540.

Thayer, J. F., Hansen, A. L., Saus-Rose, E., & Johnsen, B. H. (2009). Heart rate variability, prefrontal neural function, and cognitive performance: The neurovisceral integration perspective on self-regulation, adaptation and health. *Annals of Behavioral Medicine, 37*, 141–153.

Veneman, J. F., Kruidhof, R., Hekman, E. E., Ekkelenkamp, R., Van Asseldonk, E. H., & van der Kooij, H. (2007). Design and evaluation of the LOPES exoskeleton robot for interactive gait rehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 15*, 379–386.

Vichitvanichphong, S., Talaei-Khoei, A., Kerr, D., & Ghapanchi, A. H. (2014). Adoption of assistive technologies for aged care: A realist review of recent studies. In: *2014 47th Hawaii International Conference on System Sciences (HICSS)* (pp. 2706–2715). doi:10.1109/HICSS.2014.341.

Winchester, P., & Querry, R. (2006). Robotic orthoses for body weight-supported treadmill training. *Physical Medicine and Rehabilitation Clinics of North America, 17*, 159–172.

Wiemeyer, J., & Kliem, A. (2012). Serious games in prevention and rehabilitation—A new panacea for elderly people? *European Review of Aging and Physical Activity, 9*(1), 41–50. doi:10.1007/s11556-011-0093-x.

Zimmerli, L., Duschau-Wicke, A., Mayr, A., Riener, R., & Lunenburger, L. (2009). Virtual reality and gait rehabilitation: Augmented feedback for the Lokomat. In *Proceedings of the IEEE Virtual Rehabilitation International Conference* (pp. 150–153).

Chapter 11
Evaluation-Based Design Principles

Andreas Tolk, Geoffrey T. Miller, Gerald R. Gendron, and Benjamin Cawrse

Abstract The chapter describes a game-based prototype for distance teaching and independent training of medical procedures and generalizes the results for other application domains. In this project, the Microsoft Kinect system is used to observe experts conducting medical procedures. These observations are converted into a master model. Afterwards, students are observed conducting the same procedures. Their activities are compared with the master model and the evaluation is presented to the trainer. Several conceptual and computational challenges had to be overcome to make these ideas executable. The experiences in design and development of a prototype used to teach and evaluate Cardiopulmonary Resuscitation (CPR) are generalized leading towards evaluation-based design principles that focus on the use of master and student models and how to use them to evaluate processes in general. The methods are furthermore motivated by showing their foundations in Kirkpatrick's model. The focus of this chapter lies on presenting the lessons learned and to generalize the principles applied to support research in related domains to overcome similar challenges.

Keywords Master model • Process-result-dichotomy • MS Kinect • Evaluation-based design • Web-based teaching • Kirkpatrick's model

A. Tolk (✉) • G.R. Gendron • B. Cawrse
SimIS Inc., 200 High St./Ste 305, Portsmouth 23704, VA, USA
e-mail: andreas.tolk@simisinc.com; gerald.gendron@simisinc.com; benjamin.cawrse@simisinc.com

G.T. Miller
Eastern Virginia Medical School, 651 Colley Avenue/Room 215, Norfolk 23507, VA, USA
e-mail: millergt@evms.edu

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based Learning, DOI 10.1007/978-3-319-05834-4_11

1 Introduction: Why Use Kinect for Medical Procedure Evaluation?

The education system of today is changing. Today's learners, especially adult learners, access, acquire, and develop learning through technology-supported methods. More and more students are opting for information and instruction in an asynchronous and independently paced format over the Internet, supported by innovative, educational technologies. Renowned professors, previously only accessible for a few selected students, are providing their lectures over the web, and in some cases for free. While this approach is generally commendable, how can it be applied to courses that require dexterity? Is it possible to teach, train, practice, and test such skills and aptitudes over the web as well?

Even when evaluating the traditional education system in which professors and teaching assistants work with the learners in lectures and practice hours, these learners are expecting more flexibility and individualized opportunities to develop their cognitive and psychomotor skills. Is it possible that learners can practice whenever they want to? Why should they wait for an instructor? But if they practice on their own schedule, how can we avoid negative learning, i.e., ensuring that learners do not practice something the wrong way and then have to unlearn the wrong process before they can relearn the correct process?

These requests are not only driven by convenience for the learners, they have a significant impact on individuals and organizations as well. For example, for each instructional hour, a high-qualified expert, such as a surgeon, is teaching; they are removed from other productive work, such as patient care or research.

Finally, an important requirement of procedural skills assessment is that it should be objective, reliable (accurate and consistent), repeatable, and understandable. This is especially true for the development, acquisition, and assessment of medical and clinical procedures. Learners also need specific, event-level performance feedback to aid in error correction and performance improvement. But how can we ensure that we always measure and evaluate with the same accuracy, in particular when we need subject matter experts to perform it?

The underlying challenge of these observations can be captured in the following question: How can we capture the skills of these world experts and make them available to teach the next generation without having to sacrifice the productivity of this expert or the quality of the education at the same time?

To address these challenges—tele-education of manual skills, flexible, on-time, on-demand training for students, freeing up time of professionals from teaching without decreasing quality, and unbiased evaluation of test—a team of experts from the modeling and simulation industry and medical education experts developed the technical prototype for an Automated Intelligent Mentoring System (AIMS)™. Using the MS Kinect camera, this system observes experts' performance of a medical procedure to create a master model. The same system then observes students performing the same procedure and compares their psychomotor skill, processes, and results to the master model. Obviously, all four challenges are addressed:

students can learn manual procedures via the Internet, they can practice with this system whenever they have access to it, reliance on medical experts is minimized, and the system measures objectively and repeatable how a student performs.

Extending and updating the journal paper by Tolk, Miller, Cross, Maestri, and Cawrse (2013), this chapter is structured as follows:

- We will address the technical and computational challenges in our second section. This section focuses on the implemented prototype that by now has been released as a product to support Cardiopulmonary Resuscitation (CPR) education.
- The third section captures the lessons learned gained during the implementation and generalizes the findings to contribute towards evaluation-based design principles applicable to game-based education and test on the broad scale.
- The fourth section connects these ideas and findings with recognized theories of learning, in particular how game-based technology like AIMS can contribute to achieve higher levels.
- Finally, a conclusion session will summarize the main findings and recommendations.

The focus of this chapter shall be the general challenges that had to be overcome to build the successful prototype, the generalization of the evaluation procedures that are applicable to a wide range of education domains, and the justification of some approaches by tying them back to selected methods from education theory.

2 Technical and Computational Challenges of the Prototype

AIMS is a technical prototype that supports the training and education of medical students as a stand-alone system. AIMS can also be used as a distance learning/education tool that provides instructions and feedback via the Internet. It assesses users performing clinical procedural skills and provides specific, event-level feedback on their performance, both during and after the assessment. The central challenge was to create a means to identify a user's correctness while performing procedures. This is done through master models, as they are described in the introduction to this chapter. Master models are infused with the combined knowledge and experience of true medical experts and provide a means to compare the user's performance. This chapter summarizes some lessons learned from implementing the prototype. These lessons can hopefully support other researchers to overcome similar hurdles in comparable projects.

Understanding what the user must do is the first part of training. The next part is knowing what they are doing. Creating an effective training tool with gaming technology in support of these educational goals has its technical challenges. Although gaming technologies are designed for user interaction and virtual immersion at an unrestrictive cost, they have their boundaries. To achieve a satisfactory level of accuracy and the ability to identify objects, we had to extend the available features of the Microsoft Kinect, integrate computer vision techniques, and develop verification and assessment tools. The solutions of the software development kit (SDK) were good starting points, but often had to be improved.

Assessing a user's performance is worthless if it is not relayed back to them. The AIMS research objectives were gaining capability and improving proficiency and providing the proof of feasibility by implementing a technical prototype. In order to achieve this, objective feedback is provided to the user during the interaction in the form of formative, error correction feedback and after, summative feedback of overall clinical performance measures.

2.1 Creating Master Models for Medical Procedures

Meaningful learner performance assessment is best achieved when an "expert" is able to observe, measure and/or judge, and provide performance feedback regarding a learner's specific performance during each attempt. To provide the same features via a game-based tool, one needs an expert model to compare performance against a model that can be used to determine the "correctness" of physical actions, sequencing, and timing of the specific defined procedural steps.

Master models are sequential collections of actions which identify correct implementations of procedures. Actions that comprise a procedure are collected, in the order that they are performed by a Subject Matter Expert (SME). Master models are reviewed by multiple SMEs for reliability (accuracy and consistency). After the model is deemed correct, it is then dissected into parts. These parts classify portions of the procedure with varying levels of importance and the ability to identify the actions taken during those portions; how specifically we need to measure procedural performance, and what can we do to see it. In order to eliminate bias that a SME may have for the way that they perform the procedure, and differing SME body dimensions, it is good practice to collect multiple master models and compare them against each other to create a correctness threshold. This threshold creates variance from a combined master model while still allowing correctness as defined by the master model.

Master models are used to provide an accurate, objective comparison of the user performing for a specified procedural process or skill. Each action that a user performs is compared against these specific procedural master models, allowing for rich learner performance assessment and feedback, without requiring that the SMEs be present. This is done by comparing the users' real-time actions to actions, identified as correct, within the combined master model.

In the technical prototype, the master models represented text book solutions for medical procedures observed from medical teaching professionals. In order to be successful, a student had to follow the procedure as demonstrated by the experts. This is a very strict and constraining assumption. Section 3 of this chapter will generalize the evaluation criteria and add flexibility for the general case.

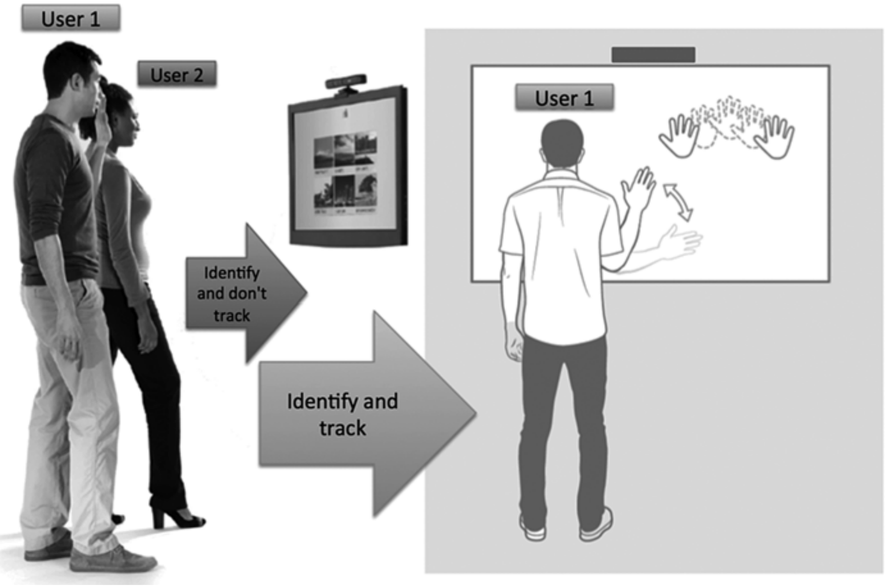


Fig. 11.1 Multiple player recognition

2.2 Computational Challenges

2.2.1 Tracking Specific Users

One issue which arises when using the Kinect is identifying who to watch if more than one person is standing in front of the camera. By default, the Kinect tracks the closest user, but it also supports more tracking schemes. Unfortunately, none of the standard schemes were robust enough for the requirements of the prototype. Default tracking of users does not function adequately in training or learning environments, which are more complex and less forgiving than a recreational environment (Microsoft Developer Network, 2014). When using depth cameras for gaming, the closest, most active user is likely the one who should be watched. This is not always the case in a training environment. The camera used in the experiment can identify six users by position, but only two of those users may be fully tracked. Fully tracked refers to the Kinect returning bone and joint information of the users as well as positions shown in Fig. 11.1. Multiple people may exist within the camera's view while training with AIMS, and we need to switch watched users as well as monitor lost users. If people other than the trainee pass by or walk near the camera, AIMS should still function as designed.

A second issue is analyzing the targeted users to be observed. Receiving user movement data is not helpful if the user's actions are obscured, and the Kinect SDK does not provide much to understand what a user is doing from raw data alone.

Identifying Who to Watch: Within the prototype, we solved the problem of tracking specific users with stances that are essentially defined by an algorithm, which compares joint positions. Before the training begins, the user is asked to mimic a given stance, one hand raised above their head for example, and with that we can identify the user exhibiting this stance as one the system would like to watch. Identifying users using stances with the Kinect for Windows can be difficult because the Kinect is only able to track joints and bones for two people at a time and only positions of six people at a time. Bone and joint information is necessary to identify the stance that a person in the scene is exhibiting, or if they are in a stance at all.

In order to check stances of more than two users within the camera's scene, the developed technique chooses the fully tracked users on each camera frame. This means on each frame it checks the stance of at most two users, and on the next frame it checks the stance of up to two other users. On each frame, all users that have not had their stance checked are put on a queue, and each user who has had their stance checked, but does not exhibit the correct stance, is put at the end of the queue. Only users who do not already exist in the queue are placed in it, which is used to decide which users should be fully tracked next. At the end of every frame, when searching for a stance, the next users to be fully tracked are taken from the queue.

This method ensures that each user in the scene that can be captured by the camera (in our prototype limited to maximal six persons) can have their stance checked until a user is found with the correct stance. After a user is found performing the correct stance, unless more than one fully tracked user is required, the system only tracks that specific user.

Switching Watched Users: Identifying and tracking of specific users, based on a stance, worked sufficiently well, unless another user exhibited the correct stance unintentionally. Because of this, a mechanism needs to switch which user is watched in case the wrong one was being selected based on a coincidental gesture that was identified as a correct stance.

When a single user is being watched, the Kinect is able to fully track a second user within the scene. Just as we identified the current user being watched, we compare the other users in the scene for the correct stance while tracking the initial "targeted" user. Since the Kinect can only fully track two users at once, the method is limited to check the stance of one additional user per frame. If a user, other than the currently watched user, is performing the stance, the switching algorithm terminates and the newly selected user is observed. If this new user holds the stance for a specified amount of frames, he becomes the watched user. To ensure this is a desired functionality and it does not inhibit AIMS during training, this technique is limited to the initial setup portion of the training phase when the variables in the scene are identified.

Watching Lost Users: Beyond identifying specific users to watch, we needed to develop a means to keep watching users, which were lost by the camera. A user may become lost (a) if they are occluded by another object, (b) if the camera fails to recognize them as a human, or (c) if they leave the camera's field of view.

The third case can be resolved by refinding the user with a stance, as described above. In order to refind users that are lost in the first two cases, AIMS keeps track of the last position for every fully tracked user. Each fully tracked user has a limited "life cycle." Once a user is initially lost, all of the users in the scene are compared against the last known position of the lost user. This comparison occurs for every frame returned by the camera. If the user is not found within the collection of visible users, then the lost user's life cycle is decremented by one. Once the life cycle reaches zero, AIMS begins searching for stances to replace the lost user.

Analyzing Watched Users: The next task is to better understand what the user is doing. Among other things, the Kinect SDK gives joint positions of users, for every frame delivered by the camera. Single instances of user positions were not enough; a history of joint positions over time is needed in order to compare user processes with the master model. The length of the history is limited for performance reasons, but having a history allows us to transform single instances of joint positions into identifiable actions and processes. This is further extended by a trending functionality that builds the history mechanism.

The trending functionality can identify a positive, neutral, or negative trend in each of the three dimensions of a joint's position across any range of that joint's collected history. This results in a means of identifying how joints move over time (Azimi, 2014). First, if a joint changes direction, the trend in one dimension of that joint changes from positive to negative, or negative to positive. Second, if a user's joint stops moving, the trend changes from positive or negative to neutral. Third, to identify more complex changes in joint movement, a sequence of trends may be used. This is very useful, for instance, if we want to track when the user has remained relatively still, we wait for a neutral trend over a large range of history. This does not force the user to remain absolutely still, but notices that they have not moved significantly. Another common use for trending is to identify change in movement. Identifying when a change occurs gives us an understanding of what the user is doing and allows us to further analyze data around that action.

While AIMS is sufficiently well equipped with assessing one user, determining how to compare the movement of two users is another challenge. When collecting the positions of joints and objects, we are able to map these points out over time, creating its path. Comparing two paths enables determining whether or not the two are moving in a similar, synchronous manor. In order to do this successfully, paths have to be recorded, rendered, and compared. This procedure allows our algorithm to produce the percent of synchronization of the observation of paths to ensure they lie within tolerable variances, etc.

2.2.2 Identifying Objects Within the Scene

For AIMS to identify objects within the scene during training, computer vision methods are applied. Specifically, color-tracking capabilities did prove useful for locating specific colors, marking instruments, or other points of special interest within the video stream given by the depth camera. The observations provided here may help other researchers to determine if the same approach is useful for them, or if alternative approaches, such as described in Han, Shao, Xu, and Shotton (2013). Examples for mathematics applicable to enhance the standard solutions have been published in Deokule and Kale (2014).

Computer Vision: AIMS identifies objects, such as manikins and medical tools, by utilizing computer vision techniques. The result of this is that we are able to know what the user is interacting with, and where they are relative to the items within the scene. Computer vision deals with analyzing digital images for identifiable traits, recognizing actions or items within an image. In our prototype, we identify colors within the scene which are simply placed on the objects we are tracking: to easily identify the chin of the manikin used for training, we apply a colored sticker to it; to identify if the right tools are used, each tool is marked with a sticker of a unique color, etc. The open source computer vision library, OpenCV©, provides this functionality (OpenCV, 2014).

When identifying objects in the scene by color, the results are more than two-dimensional pixel locations. The Kinect can map color locations to depth locations allowing the user to get three-dimensional real-world locations of objects, relative to the camera system, using a two-dimensional pixel location within the frame. Essentially, when locating one of the colors on an object, we know where that object is in relation to the coordinate space of the user.

Color Tracking: The ability to identify colors allows the system to relate the user to the rest of the objects within the scene. This added support is necessary because of the issues that arise when relying on color tracking across different environments. The principal reason different environments do not work well with color is environmental lighting conditions and variations. Different lighting conditions can substantially alter the color values returned by the camera. As this became more apparent in the research, we improved our collection and storage tools to compensate, as well as created a tool to analyze known colors.

A valuable lesson learned for all researchers utilizing this idea is the necessity to implement a calibration phase and automate this process as much as possible. The color calibration tool developed for the prototype identifies a color range from user's mouse clicks by grabbing the Hue, Saturation, and Value (HSV) values of the color in the camera's video where the user clicked. As more values are collected with clicks, the range expands and changes to include these values. To help with the lighting issue, a brightness slider exists on the tool to change the camera's brightness value of the video stream. Giving access to the brightness allows the user to get HSV values, which include different lighting conditions, making the generated color range more robust and compatible with different environments.

The prototype developed for the experiments relies on multiple color definitions per color in order to manage the issues of identifying colors across varying environments. Because of this, the color calibration tool allows storing color definitions. Every stored color is given a text descriptor, describing the color it is meant to represent: red, blue, green, etc. The experiments showed that even colors that look very similar to the human eye can vary significantly for the computer, in particular in artificial light. In the prototype, colors are set to objects by selecting the object in a dropdown, and then clicking a "Set Object Color" button. Once the button is pressed, the current color definition within the tool is set to that object, and the object can be recognized in its context by the computer. This association allows AIMS to look for the object during training with a specified color definition.

Early experiences led to the development of an automatic color calibration tool, which works as a background process during training. The auto color calibration tool attempts to find a suitable color within the environment if the currently set color, for the object it is looking for, is not being found. Because each color has a descriptor, all of the stored color ranges with the same descriptor as the one set on the object that is being looked for, may be checked on each frame from the video stream. In short, all of the color ranges, which represent the color being looked for, are checked.

In order to check a set of color ranges for compatibility in an environment, the auto color calibration tool runs them across multiple frames and retains information on how they performed. After sufficient information is collected, it is analyzed to identify color ranges with sporadic results versus those expected from a good color range. If a color range exists, after the analysis phase, which exhibits good results, it is then set to the object being searched for. If no color exists with good results after the analysis phase, then manual color calibration must be performed. It is good practice to select colors that are easily distinguishable for the computer to identify objects and to avoid colors that are neighbored in the computer spectrum.

Although manual color calibration is sometimes necessary, the collection of compatible color ranges expands and the system becomes better able to deal with varying environments. This means, whenever colors are manually calibrated and added, the new conditions may cause AIMS difficulty identifying objects, such as variances in lighting or other varying issues. The approach described in this sections mitigates these challenges. Over time, color recognition becomes more robust.

2.3 Challenges for the Evaluation

From our educational team members, we learned that half of the challenge of training and assessing users with gaming technology is identifying the actions they are taking, and the other half is relaying that information to them in an informative, concise way.

In order to effectively identify how to relay the information, robust user requirements are needed. One of our lessons learned is that users can more easily identify and formulate their needs once they have an initial prototype to work with. The iterative development within a team that comprises M&S experts, as well as medical experts, was pivotal to the success of the feasibility studies.

AIMS can provide real-time feedback during assessment so that users can better understand their running performance, and how to improve upon it. Further, summative feedback at the completion of a procedure is provided in order for learners (and instructors) to better understand and rate their performance quality, or sustainment over a period of time.

During procedural practice, real-time feedback is given by another prototype the team could integrate for the experiment: the Automated Intelligent Mentoring Instructor (AIMI). For example, during the performance of closed-chest compressions during CPR, six metrics are considered: depth of compressions, rate of compressions, recoil, hand positioning, if the users' arms are locked, and if their shoulders are directly over the midline of the manikin. Periodically while the user performs compressions, AIMI will relay how well they are performing. If they are doing well across all of the metrics, it will be a short verified, "good job," or another positive phrase. When users are not reaching an ideal or good performance rating, AIMI will identify what they are doing poorly. She may say, "Your compression depth is too shallow," or "Your arms should remain locked." This allows the user to correct their technique while still maintaining focus on the task.

Another requirement was that after performing a procedure, the user needs an overall assessment of how they did. This is provided with a "scrollable" list of measured metrics. These metrics quickly identify user performance through a relatable green, yellow, red color visual indicator schema. Based on feedback received from medical educators, each metric is also embedded within a button, which leads to more information on that metric specifically providing detail on the correct implementation.

Beyond alerting the user who performed the procedure how they did, in special cases of web-based evaluation and training, third parties must be notified as well. While the first prototype showed the feasibility of the idea that we can support teaching procedures that require dexterity using game-based technology, the next step is showing that it is also possible to teach, train, practice, and test such skills and aptitudes over the web as well. These experiments are currently ongoing.

In the current phase, our team has begun working with healthcare training companies who want to conduct remote training. In order to facilitate this, the remote company hooks their technology into a web service which the system can interact with. This allows to verify the user's credentials, identify exactly which training they should perform, and send the results of their training back to the remote company. Additional security concerns are not relevant in the context of this chapter, but are pivotal once the system reaches market maturity.

3 Generalizing the Ideas

One of the central questions to be addressed in the design of AIMS was how to evaluate a student based on the master model (derived from observing professionals conducting the task to be mastered).

In general, the master model is made up of processes and their results. In our CPR prototype, we have four processes: (1) choking, (2) Automated External Defibrillator (AED), (3) CPR, and (4) rescue breaths. These are evaluated as a series of processes that are conducted consecutively. But from our experiences with using AIMS to teach and test endotracheal intubation via direct laryngoscopy (Tolk et al., 2013), we know that there is a multitude of possibilities to evaluate the activities. Determining which option is selected is a pivotal, game-based system design decision. The selection of these metrics should be done purposefully using an evidence-based model, and not made by chance or left to implementers. As these decisions need to be made in every project similar to AIMS, i.e., using game-based technology to evaluate critical procedures, this section was written to summarize our lessons learned that can contribute to evaluation-based design principles for such systems.

3.1 Process Versus Outcome

A key consideration of assessment surrounds standard setting. Developers and SMEs must determine if the weight of the assessment is centered on the adherence to specific, sequential procedural steps, or simply achieving the desired outcome, or perhaps, some combination of these two standards (process and outcome). In the evaluation community, the dichotomy between process and outcome is as obvious as the one between process and state modeling in computer engineering. We tend to prefer measuring results and data. The reason may be that since Aristotle western philosophy of science has been heavily influenced by the idea that substantials are the main carriers of knowledge (Aristotle, 350 BC). Objects and their attributes and their relations to other objects dominate the world of knowledge representation. Processes play a subordinated role as they are merely seen as the things that create, change, or destroy objects. A recent study has shown that this view is dominant in modeling and simulation as well (Turnitsa, 2012). Only recently, the possibility to look at substantials and processes as different sides of the same coin has been discussed. Process philosophy states that processes have primacy over substantials, as a current stat is just a configuration of the underlying process, and while the processes endure, the substantials are changed by them (Rescher, 1996).

These observations are true for the evaluation of processes via their outcomes as well: one can make the argument that at the end of the process only the outcome counts, but that is not generally true. This is especially true for early learning and skill acquisition. In these cases learners are highly reliant on "rule-driven" process flows to allow them to understand, sequence, develop, and acquire a new skill.

The military practice often takes the form of a drill: soldiers practice the same procedure in exactly the same way again and again until every soldier conducts the procedure in the same way, so that when these skills are called upon during stressful or unfamiliar conditions, every soldier knows exactly what to expect from his partners or from his subordinates, ensuring the highest trust and ability to successfully conduct these procedures.

As a lesson learned, outcomes of complex procedural skills cannot be evaluated solely by looking at the final result. Subject matter experts need to define in detail what needs to be observed when. It is pivotal to understand in detail:

- What steps need to be measured?
- When—or how often—do these steps need to be measured?
- What accuracy or specificity is needed for the measurement?

3.2 Series Versus Parallel

Processes can be conducted in a series or parallel to each other. If only one person is evaluated, this challenge will not be observed so often, although it is possible that a person conducts two or three processes at the same time. The latest version of MS Kinect allows observation of several candidates at the same time, and in this case it is likely that processes will be conducted in parallel. As long as processes are executed in series, overall time constraints may be observed, and the coordination of the execution is not too challenging.

If processes are executed in parallel, the challenges become more complex. One of the best-understood alignments is temporal relations. When two processes A and B are executed in parallel, the following temporal relations for starting the processes are possible:

- A starts *before* B
- A starts *with* B
- A starts *after* B

The same observations can be made for the end of the processes:

- A ends *before* B
- A ends *with* B
- A ends *after* B

Finally, assuming that A is the process the starts earlier, we can observe

- B starts *before* A ends
- B starts *when* A ends
- B starts *after* A ends

All temporal relations associated with two processes can be expressed with these relations. In addition, a special event can be defined—such as reaching a specific state or interim result—that can be used as the synchronization point.

Whenever two processes are executed in parallel, the evaluation-based design must support capturing the underlying temporal–causal relations. In the example of AIMS, the patient transfer needs to be conducted in synchronized fashion. The learners transferring the upper part of the patient's body must be in sync with the learner transferring the lower part, or the patient will be twisted and stretched during this transfer.

As a lesson learned, for outcomes surrounding complex procedures executed in parallel, a detailed understanding of their causal and temporal relations must be thoughtfully guided and verified by the subject matter experts. In particular, the use of graphical languages, such as sequence diagrams in UML, has been proven to be very beneficial to discuss these challenges with experts (Kewley & Wood, 2012).

3.3 Enumeration Versus Collection

The discussion on series versus parallel execution can be extended by the discussion on enumerated or collected processes. Enumerated processes imply an order, either series or coordinated parallel. It is important that some processes are finished before others can start, etc.

This is not always the case. In our experiments we observed that for some processes, which we refer to as collection of services, it is simply necessary to conduct them at some point. There is no other temporal or causal relationship that binds them besides the fact that they all have to be done before a training, practice, or test run is over. For example, if a patient has several minor, not life-threatening wounds, like cuts from a broken window during an accident, it is unimportant in which order they are treated.

This observation contributed to the lessons learned that for evaluation-based design, such differences are important, as we may train and test too much if we assume specific orders that are not required. Again, the work with experts is pivotal to understand if a specific order of processes is needed, or if it is simply necessary to observe all processes at some time. Technically this creates the challenge that we need observable cues that clearly mark the start of the process enabling the system to switch into the correct observation mode for the detected process.

3.4 Variation and Deviation

It is very challenging to understand the difference between permissible variations and non-permissible deviations related to a specific procedural performance.

In our prototype, we followed the recommended approach that we will not allow variations that are not captured in the educational material. If the expert explained that a certain tool had to be used with the right hand while the left hand secured another tool, we did not allow for alternatives.

A master model needs to represent all accepted variations, which includes observable cues that allow the system to choose the correct branch for the observation. Every deviation from these accepted master models were protocolled as a student mistake.

The main lesson learned in this case was the insight that the master model must allow for acceptable variations in clinical procedural practice. If the system is consistently used for training (teaching what needs to be done), practicing (repetitive exercises of what needs to be done), and testing (comparing the actions with the master model), this is not a problem; but if students or even experts are certified with the system—and they learned the procedure to be tested following another tutorial or scheme—this may result in acceptance problems.

3.5 Deterministic Versus Stochastic

The challenge of variation and deviation becomes more complex when the element of chance is introduced. Our system so far is deterministic, no variations occur or are assumed. But many real-world problems are not deterministic at all. Also, even the same human expert will not always conduct the same procedure exactly the same. We have to work with mean values and tolerances within the statistical variation.

In particular, for practice purposes, it is recognized as useful to add random elements to avoid the student "playing the system," instead of recognizing the problem and applying the correct procedure. Video games that follow the same deterministic path are quickly losing their appeal and are viewed as boring. The same is true for serious games, which can lead to negative learning effects.

3.6 Summarizing the Lessons Learned

All these challenges do not occur exclusively, but they can happen at the same time. They can enhance each other's effects, or they may cancel each other out. A student may deviate from the right procedure, but a stochastic variation in his procedure may create a measurement that leads the system to evaluate his action as correct. Only a number of repetitions will uncover this problem.

Subject matter experts know their domain, but they may not know about the nature of processes and their evaluation. To support evaluation-based design, the engineer has to elicit the relevant information regarding serial or parallel processes, enumerated or collected processes, and what can be considered a variation and what a deviation. Taking the stochastic nature of the learning environment and learning objectives into account leads to specifications of required accuracy and consistency (AKA: reliability) of measurements, including what and when they have to be taken.

Evaluation-based design is as important as meeting the computational challenges, as they ensure that we do not measure the wrong things with high accuracy and detail and use highly sophisticated technologies to teach the wrong procedural process.

4 Alignment with Learning Theory

As compelling as technology may be, it is most relevant when it allows people to accomplish processes they need to or wish to perform more effectively. Specifically, performance matters when accomplishing activities and training evaluation is most powerful when it focuses on performance as an end to learning. As technologies have emerged throughout the early days of the Internet, a challenge has been how to best adapt learning to leverage what technology offers.

4.1 Kirkpatrick Four Level Evaluation Model

While alternative frameworks of learning and evaluation do exist, our team had experience with the Four Level Evaluation Model by Kirpatrick and Kirkpatrick (2006). Donald Kirkpatrick first published his model in the late 1950s. The book *Evaluating Training Programs* model was popularized in 1994 and has been a foundational aspect of learning theory ever since. The work of Galloway (2005) supported our selection of this approach for our domain. The four levels, listed in descending order of impact, are summarized as follows:

- *Results*. The upper level is paradoxical. Although it is the desired end of training, it is not necessarily measured as well as possible. The intent of the model is to measure returns on investment, increased effectiveness, and the like resulting from training.
- *Behavior*. This third level is among the most difficult to measure and is often impossible to measure at the time of training.
- *Learning*. This level is among the most used measure of learning—typically in the form of a test. Measures of learning intend to demonstrate the training audience has increased and retained their knowledge of the training material.
- *Reaction*. The lowest level of the model, it provides a measure of individual satisfaction and engagement among the training audience, the instructor, and the externalities (e.g., the classroom). It is often measured through satisfaction surveys.

It is common to observe measures based on how much training was accomplished or distributions of test scores. Less common is the connection between the upper two levels and lower two levels. Nickols (2011) captures this in his work using the Kirkpatrick Model in a design mode.

Figure 11.2 shows the same four levels—listed in the same hierarchical order. Nickols leverages the Kirkpatrick model to attain business results in the design stage. The basic intent is to determine the desired business results (level 4) and translate those into behaviors the team members might attain (level 3) to support those business results. This has a direct correlation in the bottom two levels in terms of designing the training. How might this approach be of assistance in designing new training technologies?

4.2 Applying a Nickols Design View to AIMS

From its inception, the AIMS approach sought to not only provide traditional measures associated with training (such as test results) but also robust measures having the higher level influences as noted by Kirkpatrick. Inspired by Nickols (2011), the various measures designed into AIMS intend to amplify the learning attained by providing tangible, business results.

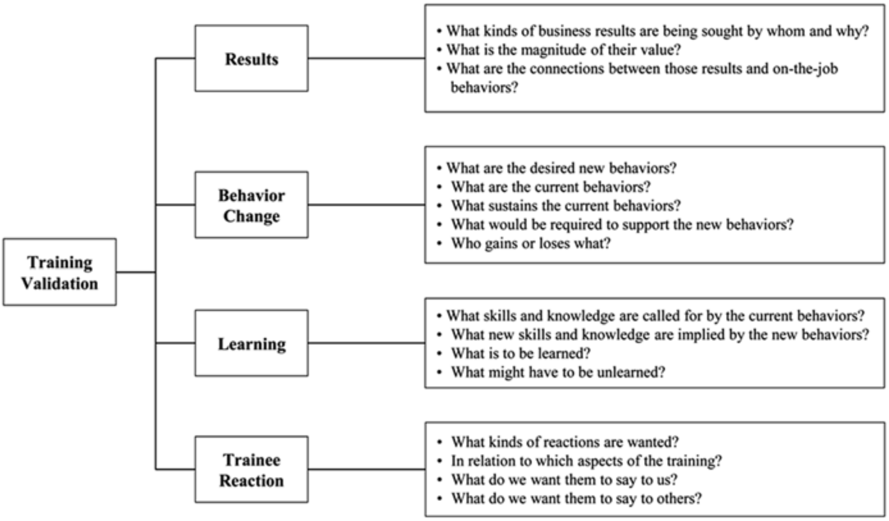

Training Validation
Results
• What kinds of business results are being sought by whom and why?
• What is the magnitude of their value?
• What are the connections between those results and on-the-job behaviors?
Behavior Change
• What are the desired new behaviors?
• What are the current behaviors?
• What sustains the current behaviors?
• What would be required to support the new behaviors?
• Who gains or loses what?
Learning
• What skills and knowledge are called for by the current behaviors?
• What new skills and knowledge are implied by the new behaviors?
• What is to be learned?
• What might have to be unlearned?
Trainee Reaction
• What kinds of reactions are wanted?
• In relation to which aspects of the training?
• What do we want them to say to us?
• What do we want them to say to others?

Fig. 11.2 Nickols' model. A design view of the Kirkpatrick model

- *Results*. The enduring philosophy of the AIMS approach has been and continues to be greater opportunity to automated expert, objective evaluation with individualized, user-specific feedback for performance improvement. Simply stated, the AIMS master model aids trainees' practice in ways similar to being under the watchful eye of a master instructor. The trainee is presented with subtle variations in technique captured as compared to the master model. This is executed using design principles that encourage skill precision and mastery level skill acquisition of a specified procedure. As depicted in the Nickols (2011) model, there is a very deliberate connection between the results and on-the-job behavior.
- *Behavior*. As noted earlier, these measures are among the most difficult to capture. In the case of AIMS, the behaviors range from willingness to perform CPR, to demonstrating confidence in the use of an AED. Results of studies surrounding AIMS training yield some interesting results. Some data captured showed some trainees felt more confident in performing CPR prior to training with AIMS and lower levels of confidence after training. When asked why, respondents indicated that they had not been assessed so thoroughly and they now understood that their technique could be improved upon. Fortunately, this class of respondents also indicated a desire to practice and master the CPR technique in order to be ready for on-the-job needs.
- *Learning*. AIMS CPR training contains seven measures to determine if the skills and knowledge necessary to enhance behavior are demonstrated in training. There are six independent measures and one measure that aggregates the first six into a single metric. Perfected performance in CPR results from good body

mechanics and proper application of force. Three measures assess the trainee's body mechanics:

1. Shoulder placement over the patient
2. "Locking" of the elbows during CPR
3. Hand position on the torso. All of these are captured and reported

   Three other measures evaluate how they are delivering CPR:

4. Depth of compression
5. Recoil from compression to allow the heart to fill
6. Rate of compression to ensure adequate blood flow

Each of these six must match the master model in order for the trainee to pass. The seventh measure aggregates the six independent measures using a weighting scheme to provide a single measure of training performance. This is not only psychologically appealing to trainees but also provides an easy benchmark for them to recall as they continue to perfect the skill.

- *Reaction.* The benefit and challenge of implementing game-based design revolves about user reaction. The success of many game-based approaches is due in part to the aesthetics of the user interface. The development team placed great emphasis on understanding market needs as well as training requirements to blend a powerful training tool with an engaging experience.

4.3 Outlook on Next Steps

The master model created for the CPR prototype has provided both the development and evaluation team with very interesting results to consider as they develop future sets of medical and human factors training technologies. Most interesting is the positive response among users, using AIMS. This is due in large part to the well-balanced use of the four levels of feedback intrinsic in the design principles of AIMS. The AIMS evaluation team will conduct continued studies on methods to enhance data-centric measures associated with behavior and results. The designers hope to assess behavior in a qualitative way during training events themselves.

5 Conclusions

With educational systems becoming more available, remote training is increasingly important to its evolution. Remote training designed to instill mental process has been tested and proven, but automated, remote training requiring physical assessment is still in its early stages. Effective automated, remote training provides students with the ability to train on their time and at their own pace. Allowing such flexibility in student training encourages students to train more. Increased training has shown to improve abilities, as long as the training does not provide negative learning.

Additionally, the benefits of automated, remote training for students are extended to the experts who would originally be the teachers. These experts become more available to conduct other productive work, which is often more lucrative for the organization than teaching students. The experts are not fully removed from the training, because they have to be involved to ensure the training is accurate and correct, but the reliance on them to fully assess students may be significantly reduced. For this to be successful, the provided training must be objective, repeatable, and understandable.

SimIS has developed, in collaboration with EVMS, the Automated Intelligent Mentoring System (AIMS) in order to address challenges present in automated, physical assessments. The system's capabilities solve a substantial number of tele-education issues related to teaching and assessing manual skills, flexible training for students, freeing up time of professionals from teaching without decreasing quality, and unbiased evaluation. Designed with the MS Kinect camera, AIMS observes experts conducting a medical procedure to create a master model. The master model is then used by AIMS to compare against students performing the same procedure. The system provides remote, accessible, repeatable, and objective training to students while freeing up the experts. The experts are required for the master model, but are no longer needed for training once it is created.

References

Aristotle. (350 BC). *Physics*. Retrieved from http://classics.mit.edu/Aristotle/physics.html.

Azimi, M. (2014). *Skeletal joint smoothing white paper*. Retrieved from http://msdn.microsoft.com/en-us/library/jj131429.aspx.

Deokule, N., & Kale, G. (2014). Human action recognition using Kinect. *International Journal of Engineering and Computer Science, 3*(7), 7199–7202.

Galloway, D. L. (2005). Evaluating distance delivery and e-learning: Is kirkpatrick's model relevant. *Performance Improvement, 44*(4), 21–27.

Han, J., Shao, L., Xu, D., & Shotton, J. (2013). Enhanced computer vision with microsoft kinect sensor. *IEEE Transactions on Cybernetics, 43*(5), 1318–1334.

Kewley, R. H., & Wood, M. (2012). Engineering principles of combat modeling and distributed simulation. In A. Tolk (Ed.), *Engineering principles of combat modeling and distributed simulation*. Hoboken, NJ: Wiley.

Kirpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluation training programs* (3rd ed.). San Francisco: Berrett-Koehler.

Microsoft Developer Network. (2014). *Tracking users with kinect skeletal tracking*. Retrieved from http://msdn.microsoft.com/en-us/library/jj131025.aspx.

Nickols, F. (2011, 21 April). Leveraging the kirkpatrick model [Web log message]. Retrieved from http://www.trainingjournal.com/blog/articles-blogs-leveraging-the-kirkpatrick-model/.

OpenCV. (2014). *OpenCV*. Retrieved from http://opencv.org.

Rescher, N. (1996). *Process metaphysics: An introduction to process philosophy*. Albany: State University of New York Press.

Tolk, A., Miller, G. T., Cross, A. E., Maestri, J., & Cawrse, B. (2013). Aims: Applying game technology to advance medical education. *Computing in Science and Engineering, 15*(6), 82–91.

Turnitsa, C. (2012). *Exploring the components of dynamic modeling techniques* (Doctoral dissertation). Retrieved from OCLC WorldCat. (819638936).

Part V
Serious Games Analytics for Learning and Education

Chapter 12
Analytics-Driven Design: Impact and Implications of Team Member Psychological Perspectives on a Serious Games (SGs) Design Framework

James Eric Folkestad, Daniel H. Robinson, Brian McKernan, Rosa Mikeal Martey, Matthew G. Rhodes, Jennifer Stromer-Galley, Kate Kenski, Benjamin A. Clegg, Adrienne Shaw, and Tomek Strzalkowski

Abstract The number of educational or serious games (SGs) available to educators has increased in recent years as the cost of game development has been reduced. A benefit of SGs is that they employ not only lesson content but also knowledge

J.E. Folkestad (✉)
Colorado State University, 244 Education Building, 80523-1785 Fort Collins, CO, USA
e-mail: james.folkestad@colostate.edu

D.H. Robinson
Colorado State University, 1588 Campus Delivery, 80523-1785 Fort Collins, CO, USA
e-mail: dan.robinson@colostate.edu

B. McKernan • T. Strzalkowski
University of Albany, SUNY, 1400 Washington Ave, Albany 12222, NY, USA
e-mail: brian.mckernan@gmail.com; tomek@albany.edu

R.M. Martey
Journalism and Technical Communication, Colorado State University,
80523-1785 Fort Collins, CO, USA
e-mail: rosa.martey@colostate.edu

M.G. Rhodes • B.A. Clegg
Department of Psychology, Colorado State University, 80523-1785 Fort Collins, CO, USA
e-mail: matthew.rhodes@colostate.edu; benjamin.clegg@colostate.edu

J. Stromer-Galley
Syracuse University, 220 Hinds Hall, 13244 Syracuse, NY, USA
e-mail: jstromer@syr.edu

K. Kenski
University of Arizona, 1103 E. University Blvd, Room 211, 85721-0025 Tucson, AZ, USA
e-mail: kkenski@email.arizona.edu

A. Shaw
Temple University, 2020 N. 13th St., Annenberg Hall,
Room 203a, 19122 Philadelphia, PA, USA
e-mail: adrienne.shaw@temple.edu

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based
Learning, DOI 10.1007/978-3-319-05834-4_12

contexts where learners can connect information to its context of use with active participation and engagement. This, in turn, improves learners' ability to recall, integrate, and apply what they learn. Much of the research on game analytics has examined learner in-game trails to build predictive models that identify negative learner actions (e.g., systematic guessing after the fact). However, analytics can also be used in the game design and development phases. Drawing on evidence-centered design (ECD), the chapter outlines ways that analytics can drive the development of scenarios and activities in a game and thus allows SGs to function as contextual apprenticeships, providing robust assessment opportunities. We describe how ECD theory was applied in a project to develop and test a SG that trains people to reduce their reliance on cognitive biases. We describe instances during the design process where our team encountered obstacles due to differing psychological and learning/teaching orientations, a topic rarely explored in the SG or ECD literature. Furthermore, we describe the final analytics-based game design features. We propose an additional element (persona) and how we anticipate incorporating that ECD extension into future projects.

Keywords Evidence-centered design • Game design • Analytics • Perspectives • Learning • Cognitive biases • Serious games

1 Introduction

In recent years, scholarly interest in serious games (SGs) has risen in the fields of education, communication, psychology, and game studies. Defined as games with an educational goal, or that do not have enjoyment as their main purpose, SGs have also increased in availability and distribution (Michael & Chen, 2005; Susi, Johannesson, & Backlund, 2007).

For many, the rich contexts and scenarios that can be created in games drive interest in SGs. Game designers can create situations where players can learn through virtual experience rather than more passive reading or observation. For example, play-learners can explore urban planning in a simulation game and witness the impact of their decisions, such as watching their city thrive or enter decay. The opportunities for students to try, fail, and learn from that failure has been appealing to many different disciplines. Accordingly, SGs have been developed to address a wide range of topics, from those more focused such as arithmetic and spelling to those more complex such as urban planning and sustainability (Michael & Chen, 2005). These SGs have been designed for young children as well as adult learners.

Barab, Gresalfi, and Ingram-Goble (2010) and Gee (2007) suggested that game design should focus on establishing meaningful and illustrative situations where play-learners adopt different roles and develop complex relationships with the educational content (see also Winn, 2002). Furthermore, they suggest that engaging situations where students can apply content understanding should be the goal of design.

Barab et al. stated that video game technologies provide “methodologies for creating curriculum that is deeply immersive, highly interactive, and experientially consequential” (p. 534). Barab et al. also suggested that game-based log files have the potential to reveal player trajectories and assess understanding through in-game actions/behavior.

Game developers have begun to take these possibilities further by embedding learner assessment tools within games. Loh, Anantachai, Byun, and Lenox (2007) suggest that the data generated by player actions can allow the educator/trainer to assess and monitor performance. Given the importance of learner assessment, it is critical that serious games analytics (SEGA) incorporate considerations of assessment in their design process (Loh et al.).

In this chapter, we describe a worked example of the design process used to develop a SG and corresponding assessments as part of a 4-year project designing and experimentally testing learning outcomes. We describe the challenges in creating a game that was engaging, effective for learning, and that incorporated data tracking mechanisms for effective assessment. This process allowed us to use evidence-based approaches for improving learner understanding and knowledge acquisition. Our team discovered that designing SGs is not a trivial undertaking due to the interconnected complexities of designing games; play mechanics, assessment, and data collection for analytics.

We discuss how Mislevy and Riconscente (2006) evidence-centered design (ECD) framework guided the design process and allowed us to incorporate analysis and assessment into all phases of the project. We describe how analytical tools were built, how they informed the design of specific activities in the game, and how resulting data were iteratively used to refine and enhance learning processes. This chapter makes two contributions. First, we provide an additional worked example of the application of ECD within a SG design and development project. We explain the importance of considering in-game data when designing situations and game features; data that will ultimately drive SEGA. Second, we provide a detailed description of how disparate team member psychological perspectives and orientations toward learning created challenges within the design process. We discuss the critical nature of these challenges and suggest the need for an extension to the ECD framework. We note that the full ECD framework is complex; its terminology and design templates are not fully explained within this worked example (see Mislevy & Riconscente, 2006, for a complete description of ECD).

1.1 Complex Assessment and SGs

SG projects are inherently complex and require teams of experts that cross disciplines. For example, the interdisciplinary team for the project discussed here, called Cycles of Your Cognitive Learning, Expectations, and Schema (CYCLES), included cognitive psychologists, educational researchers, linguists, game scholars, statisticians, and game developers. Each of these experts had a particular perspective on games, education models, and learning structures.

Coordinating and integrating disparate perspectives within diverse teams is difficult work (Behrens, Mislevy, Bauer, Williamson, & Levy, 2004). Team members from different disciplines often use similar terms that have different meanings. In addition, individuals with different perspectives towards learning and education often do not agree on the most appropriate way to teach specific content. For example, this project originally used two broad-based learning theories, the Observe-Orient-Decide-Act Loop evolutionary learning and decision-making model (Boyd, 1995) and the transformational learning theory (Mezirow, 1997). Although these learning theories provided a solid foundation for conceptualizing our overall learning elements, they were ambiguous regarding the translation of desired learning into game-based activities and features that would enhance or demonstrate learning.

We thus selected ECD to formulate a comprehensive approach to assessment design for our research purchases. ECD provided our team the guiding framework for orchestrating the complexities of communication, varying perspectives, learning objectives, and assessment measurements.

2 ECD Theory Overview

ECD (Mislevy, Behrens, Dicerbo, Frezzo, & West, 2012; Mislevy & Riconscente, 2006; Mislevy, Steinberg, & Almond, 2003) provides a framework for designing, producing, and delivering assessments. ECD works particularly well with complex assessments such as those involving SEGAs because it incorporates developments from cognitive science, technology, and statistical modeling. ECD provides a multiple-layered framework for designing assessments following Messick's (1994) questions:

1. What complex of knowledge, skills, or other attributes are to be assessed?
2. What behaviors or performances would reveal those constructs?
3. What tasks or situations would elicit those behaviors?

During the first stage, domain analysis, the knowledge, skills, and abilities reflected in the domain are collected (Mislevy and Riconscente), as are the contexts and ways in which people might use them, and the observable behaviors that would reveal them. When SGs use real-world simulations, domain analysis assures a sense of realism (Behrens et al., 2004).

The next stage, domain modeling, involves refining and organizing the domain analysis into assessment arguments. To gather evidence about the knowledge, skills, and abilities in the domain, coherent arguments are built around the contexts in which people behave that will reveal such evidence. These arguments delineate, frequently through narrative, the nature of the assessments needed. Similar to storyboarding for SGs, domain modeling allows interdisciplinary coordination when making complex assessments.

The third stage, conceptual assessment framework (CAF), refines developing design concepts and assessment arguments—specifically the more technical components of the structures and variables in the Task, Evidence, and Student Models. Such components in traditional assessments would be item types, item scoring rules, etc. However, in complex assessments such as a multi-player SG, simple items become inadequate and must be expanded.

The task model describes the environment in which the students behave and the observables that serve as evidence. Task model variables are environmental features that are crucial when interpreting students' actions. Some will be predetermined, such as the components, speed, and affordances in a particular level of a game. Others need to be tracked dynamically in response to the unfolding actions of the students, such as decisions made in the game. The observables are not discrete item responses, but artifacts that hold the potential for gleaning evidence about knowledge, skills, and abilities—as varied as the health of the avatar at the end of the game, the trace of actions the student has taken, rationales provided for actions, or the time and resources expended.

The student model addresses Messick's (1994) first question that asks what complex of knowledge, skills, or other attributes should be assessed. Student model variables include the knowledge, skills, or abilities determined in the domain analysis, at a level that suits the context of the assessment. These are latent variables in that they are not observed directly, yet they drive how decisions and high-level feedback are determined. It is thus necessary to use student behaviors in various contexts as evidence about student model variables.

The evidence model bridges the students' behaviors and the assessor's belief about the student. It creates an argument about why and how the observables in a given task situation constitute evidence about student model variables. There are two parts to the evidence model: evaluation rules for identifying and evaluating the salient aspects of observables, and the measurement model for synthesizing their importance in updated beliefs about student model variables. Human judgment or automated evaluation rules are possible, and they may be exercised at the end of an episode or identify salient events as action progresses. They produce values of observable variables. The measurement component describes, in terms of a probabilistic (psychometric) model, how the observable variables depend on student model variables. It is used to update belief about them by means of Bayes' theorem or other probabilistic models. Of particular importance in SGs are the dependencies among observations caused by the impact of past actions on a present situation, and the identification of multiple aspects of the same complex performance.

2.1 Previous ECD-Driven Serious Games Research

Several scholars have described how ECD provides a strong framework for constructing valid and reliable measures to assess the learning outcomes of SG (Behrens, Frezzo, Mislevy, Kroopnick, & Wise, 2007; Reese, Tabachnick, & Kosko, 2014;

Sweet & Rupp, 2012). These works describe how ECD can help transform the massive amount of potential data points generated in play sessions from SG into clear measures for learning assessment. In particular, many scholars consider ECD to be an ideal framework for assessing twenty-first century skills (e.g., systems thinking, creative problem solving, identity management, teamwork, perspective taking, and time management), many of which have proven to be extremely difficult to evaluate in traditional educational settings (DiCerbo, 2014; Rupp, Gushta, Mislevy, & Shaffer, 2010; Shaffer, Hatfield, Svarovsky, Nash, Nulty, Bagley, Frank, Rupp, & Mislevy, 2009; Shute, 2011; Shute and Ke, 2012; Shute, Masduki, Donmez, Dennen, Kim, Jeong, & Wang, 2010).

Drawing from these insights, several scholars have designed measures based on ECD principles to assess the learning or competency acquired in many different games. Scholars have used this framework to evaluate a broad array of skills, including urban planning (Shaffer et al., 2009), environmental planning (Sweet & Rupp, 2012), engineering (Shaffer et al., 2009), systems-thinking (Shute, 2011; Shute et al., 2010), creative problem solving (Shute, Ventura, Bauer, & Zapata-Rivera, 2009), persistence (DiCerbo, 2014), computer networking (Behrens et al., 2007; Shute et al., 2009), and fire safety (Al-Smadi, Wesiak, & Guetl, 2012). The wide range of topics scholars have used ECD to assess highlights the powerful potential of the framework's approach to in-game assessment.

Despite the promising work these scholars have done, ECD is not intended to serve only as a measure of assessment for finished projects. The founders of this approach constructed ECD as a framework to guide the entire design process, from the initial idea to the final assessment. To be fair, a few scholars have provided worked examples of how they applied ECD to the entire design process (Presser, Vahey, & Zanchi, 2013; Reese et al. 2014; Rupp et al. 2010). However, these works only report the results of each step of their team's design process and provide minimal insight into alternative approaches that were considered or how ECD principles guided selection of options. Given the complexity of the ECD framework, more worked examples of the full process are needed to help scholars gain a stronger understanding of how ECD principles can be incorporated into the design of SG.

In this chapter, we provide a detailed description of how we applied insights from ECD in our design process, including instances in which the team had to choose from several different plausible options. These decisions involved not just choices over how to approach in-game assessment, but also what learning theory should guide the scenarios instantiated in our game. Previous worked examples on applying an ECD framework to designing a SG have largely ignored the subject of selecting an appropriate learning theory. When the learning theory selected is reported, it usually reflects the overarching type of SG being designed. For example, epistemic games incorporating ECD components rely on a theory of learning known as epistemic frames hypothesis (Shaffer et al. 2009; Sweet & Rupp, 2012). Although the particular learning theory may be easy to determine for some projects, the interdisciplinary nature of many SG initiatives means that some team members may endorse different approaches to learning. Consequently, we describe the choices our team faced when selecting an appropriate learning theory in our worked example. Hopefully, these descriptions will help future teams plan accordingly when designing SGs of an interdisciplinary nature.

3 Cycles: A Worked Example

This worked example highlights the development and assessment of a multi-year game design and development project called CYCLES. Using this example, we illustrate how ECD can be applied to designing SGs reinforcing the use of ECD in gaming contexts, and discuss how analytics became a central focus of design.

The CYCLES game was designed and built to teach play-learners to recognize and mitigate cognitive biases. When making judgments, humans generally use shortcuts, or heuristics, to expedite those decisions. Although these heuristics may be adaptive, they also result in systematic and predictable distortions (biases) in decision-making (Tversky & Kahneman, 1974). Bias mitigation strategies encourage pattern recognition and behavioral modification of cognitive processes to improve judgments. However, traditional models of teaching and learning have shown little success in training people to reduce biased behavior (Fischhoff, 1982; Kahneman, 2003); in fact, training can actually exacerbate rather than reduce the bias (e.g., Sanna, Schwarz, & Stocker, 2002).

The phase-one game focused on three types of cognitive bias: fundamental attribution error (FAE), confirmation bias (CB), and bias blind spot (BBS). These biases can lead to efficient and sometimes accurate decisions but may also lead to systematic error (Harvey, Town, & Yarkin, 1981; Nickerson, 1998; Pronin, Lin, & Ross, 2002). In fact, the negative impact of these cognitive heuristics has been documented in a variety of different contexts, including medical, legal, and military settings (e.g., Nickerson, 1998).

Overall, instruction designed using behavioral and information-processing models has been largely ineffective as a training approach for helping individuals mitigate these biases. Thus, our team was interested in building SG that train play-learners in context. According to the psychological literature on cognitive biases, bias mitigation involves setting an intention to mitigate bias, identifying biased thinking patterns, and intentionally applying mitigation strategies (Kahneman, 2003). Learning to mitigate cognitive bias is heavily dependent on a perceiving-acting cycle, where agents efficiently apply mitigation strategies when needed and was the basis of the teaching model in the CYCLES games.

3.1 CYCLES Game and Design Process

The CYCLES game was designed and built for a multi-year research study that investigated the impact of various game features on play-learners' ability to recognize and avoid cognitive biases. "Results indicated that the level of bias exhibited decreased significantly from a pretest (before game play) to a posttest (after game play) and were characterized by medium to large effect sizes" (Shaw et al., 2013). As part of that process, the CYCLES team developed the learning objectives, teaching activities, and game elements including the setting, artwork, narrative, colors, and audio.

The game environment. The game's environment is an austere futuristic training facility in a sparsely furnished, bland institutional building. The play-learner's avatar is dressed in a suit with a helmet that obscures the face and hair. Each room is a puzzle related to a cognitive bias and the play-learner performs tasks in each room to complete it and progress to the next challenge. Because we wanted to emphasize the inherently human nature of cognitive biases, we used robots with human brains to exhibit biased behavior. The play-learner had to recognize when cognitive biases were present and used techniques to mitigate the effect of bias within a situation. In keeping with the training facility team, the learner was guided through each room by the computerized voice of a trainer. In addition, as the play-learner progressed from room to room, informational screens or infographics were used to provide information about cognitive biases and mitigation techniques.

During the game design and development process, three primary perspectives were used: information processing, situated learning, and social constructivism. These perspectives influenced what different team members considered evidence of student proficiency during the game experience.

Generally, game design professionals drew on situated learning concepts and focused on game features that encouraged exploration within the game space. The cognitive psychologists emphasized exposure to learning concepts through repetition/spacing and quiz-based testing to enhance retention. The education professionals highlighted behavioral learning objectives and in-game experiences that allowed play-learners to construct their understanding. The disparate, often competing, perspectives on learning created challenges during the design and development process. A description of when these challenges arose and how elements of ECD helped us alleviate these problems is described below.

3.2 Domain Analysis Layer: What Complex of Knowledge, Skills, or Other Attributes Are to Be Assessed?

The design process of the CYCLES game began with analyses of cognitive bias and the research on eliciting and mitigating biased behavior. The team of professors, graduate students, and educational game designers began by reading books (e.g., Kahneman, 2011) and refereed journal articles (e.g., Jones & Harris, 1967; Tetlock, 1985). In early stages of our project the team worked to understand each cognitive bias and the related mitigation techniques. Cognitive bias experts on our team gave brief lectures on the different cognitive biases and mitigation strategies.

In this stage, we gathered information about concepts and terminology surrounding each cognitive bias and developed rich descriptions of potential in-game tasks. We identified knowledge, skills, and abilities that each play-learner would need to master. For example, we gathered information about FAE, which is the tendency to overemphasize personal characteristics when explaining another individual's behavior. We identified and shared scenarios in which this cognitive bias presents itself. These scenarios often included rich description of tasks and practice where FAE was evident and potential strategies for avoiding FAE. For example, we rehashed the following scenario many times as we attempted to understand FAE.

> *You pull into a gas station that has only two free pumps. However, the car in front of you stops at the first pump, preventing you from getting to the second pump. You conclude the driver is inconsiderate, blaming this behavior on the individual's personality. However, after you attempt to drive around the inconsiderate driver you realize that the pump in front is covered in yellow tape and out of order. The environment was to blame, not the individual.*

We revisited this scenario often to understand how FAE affects decisions and to understand how to train learners to overcome such errors.

3.2.1 Advancing Analytics Within the Domain Analysis Layer

In the early stages of the project, the team gave little consideration to analytics. The primary focus was on understanding each bias and the behaviors and knowledge each play-learner needed to learn. However, through each design iteration and subsequent game updates and development, our team began to consider how data generated by in-game actions such as clicks could be used, including in the domain analysis layer. We identified the behaviors of those who were proficient in mitigating cognitive bias (experts) and tracked what types of data would indicate their proficiency.

In the design layer, we had to think about each behavior in terms of its benefit for analytics. However, much of what players needed to learn was metacognitive and required that players monitor their thinking, which is difficult to measure during gameplay. Therefore, we had to identify behaviors and learning objectives that could be tracked via in-game actions and resulting data log transactions.

In the gas station example above, it would be difficult to monitor metacognitive activity within the play-learner; however, we realized that the actual behavior of slowing down, or slowing down to execute decisions, could provide evidence of the play-learners' understanding and thinking process. We concluded that a critical behavior in mitigating FAE was that the decision-makers take more time to identify which bias is "at play" and consider environmental factors before making a decision. In this way, considering the analytics we needed drove us to identify a trackable behavior that would help us achieve the overarching learning objectives.

3.2.2 Challenges Within the Domain Analysis Layer

There were several challenges in the domain analysis layer of the CYCLES development process. Overall, without a comprehensive design framework such as ECD, we had few guidelines for making key decisions at this stage. For example, the CYCLES game needed to be played within about 30 min, a considerable design constraint. Thus, incorporating a design goal that slowed down players' progress generated debate among team members. This mechanism appealed to team members who drew on social constructivist perspectives because it allowed play-learners to consider their decisions and experiences carefully. Other team members worried that the slowdown mechanism would take too much time or that the mechanism was too mechanical, moving game-based fun and thus reducing engagement.

These differences significantly influenced our design and development process as decisions in the domain analysis layer impacted subsequent decisions and caused many late phase design changes. In our post hoc design analysis, we realized that identifying these perspectives on learning could have streamlined the iterative design process. As our team began to design for analytics and the collection of evidentiary data trails (assisted by ECD), we found our team conducting an interdisciplinary exchange, a "dance" that included attempting to fully understand desired learning outcomes and design game-based features that would collect evidence—all filtered through the lenses of team members' psychological perspectives.

The ECD framework helped us prioritize domain skills and knowledge based on analytics in this layer. Using ECD, our team oriented to critical domain skills that were behavior based and thus could be logged and analyzed. For our team, it was this perspective shift from designing a game based on knowledge acquisition and retention to a game designed to collect evidentiary data of performance that helped us orient our domain analysis to behavioral domain activities.

3.3 Domain-Modeling Layer

In the domain-modeling layer, we worked to organize the outcome from the domain analysis into SG assessment arguments. In an effort to highlight the design considerations that our team accommodated, we use the design pattern template to emphasize each attribute considered for FAE (see Table 12.1).

3.3.1 Advancing Analytics Within the Domain-Modeling Layer

In designing for SG, it is important to incorporate game environments and features that provide the most useful data to be collected for analytics. We needed to design game scenarios and situations that aligned with domain behavior and that would translate into data trails that would identify ideal and problematic behaviors, such as avoiding or exhibiting the cognitive biases of interest. In developing the rationale

Table 12.1 Fundamental attribution error (FAE)—design pattern

Attribute	Value(s)	SEGA assessment arguments
Summary	Identifying, recognizing, and mitigating FAE	Interact in an environment and recognize when FAE may be influencing their decisions
		Employ the correct mitigation strategy
Rationale	Individuals monitor their thinking realizing when such heuristics are in play, and mitigate effects by altering behavior	Surveys the situation and then identifies and discriminates which cognitive bias is at play
FAE skills, knowledge, behaviors	Monitor environment for situations where decisions are based on individual's attributes or behavior	Non-player characters (NPCs) pushed play-learners to make decisions when they entered the environment
	Slow down (behavior) and consider alternative factors including environmental considerations	Play-learners had the freedom to make decisions without identifying the bias
	Define FAE	Basic definitions of all three cognitive biases were introduced early and students were tested
	Identify individual characteristics versus environmental	Cognitive bias tool belt (described below)
Potential observations	Identify definition of FAE	Identify which cognitive bias is influencing their decisions
	Determine when an individual is being blamed in a situation	Identify environmental factors to mitigate impact of FAE on decisions
	Identify additional environmental factors	
Variable features	The play-learner should have the opportunity to explore the situation to gather more information	Play-learner exploration will provide evidence to make a case that the player is considering alternatives such as the environment

(see Table 12.1), we determined that the play-learner would need to enter an environment and evaluate the situation to determine which cognitive bias was at play. For FAE, we determined that play-learners had to monitor the environment and identify when they were being asked to make a decision based on an individual's characteristics and when environmental factors could also be to blame.

3.3.2 Challenges Within the Domain-Modeling Layer

A major challenge within the domain-modeling layer was reconciling team member orientations to learning and teaching. According to Mislevy et al. (2012), it is critical to make these perspectives explicit and to establish common perspectives among team members. Bringing clarity to these different perspectives is difficult, however. Here, we identify each perspective and discuss how their focus influenced the design process. We note that these are the broader patterns associated with different perspectives, but do not reflect every team member's opinions or beliefs for every aspect of design.

Game developers (*discovery-based learning perspective*). The game developers on our team brought an important and valuable discovery-based learning perspective. Through experience with designing and play-testing numerous games they understood game features and game mechanics and how to design such features. Furthermore, game developers often rely on discovery-based learning to maintain and hone their own knowledge and skills related to game development. Part of that discovery-based learning is based on fun and games. These team members carried with them this unique orientation to learning and teaching. They focused on how to design a game that allowed for exploration play, which would then be translated into understanding and learning.

Educational researchers (*social constructivist perspective*). The educational researchers on our team brought a social constructivist perspective to learning and teaching. From their perspective individuals construct new knowledge through experience and assimilation of new ideas and behaviors. These team members had a unique perspective on teaching and learning, believing that by developing sound behavioral objectives and translating those into game-based experiences the play-learners would ultimately learn the required knowledge and skills.

Cognitive psychologists (*information-processing perspective*). Cognitive psychologists viewed learning and teaching through an information-processing perspective. These team members also had a unique orientation to learning within game and analytics. They focused on ways to use the spacing effect and the testing effect, for example, to help play-learners process the information being presented on cognitive biases so as to effectively learn that information. Spacing (see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006) refers to the memorial advantage that accrues when information is studied multiple times, but not consecutively (i.e., presentations of the same item are separated by at least one other item), compared with massing information (i.e., consecutive presentations of the same item). The testing effect (see

Roediger & Butler, 2011, for a review) refers to the finding that retention is superior when learners engage in some form of retrieval of previously studied information relative to simply restudying (re-reading) this information.

Overall, these different perspectives had a significant impact on our SG design and the subsequent outcomes. Different perspectives often resulted in disputes over the appropriate design strategy or inconsistency in new designs, depending on the dominant perspective present or "the loudest" during a design decision. We believe that these different perspectives benefited our team and the CYCLES game design. However, our team is working to streamline this process. In the discussion section of this chapter, we propose an additional ECD component (persona) to assist the ECD framework. As Mislevy and Riconscente (2006) stated about the ECD framework, "the psychological perspective from which the designer views the task informs this component, since it determines the criteria for exactly which aspects of work are important and how they should be evaluated" (p. 18).

3.4 Conceptual Assessment Framework

In the third layer of ECD, we move the narrative from the domain-modeling layer to the elements and processes needed to implement gameplay and an assessment. The CAF provides a series of six models that represent areas of assessment: what, how, where, how much assessments are measured, as well as how it is presented and delivered. The CAF starts with operationalizing the tasks and work that the students will do and contributes to inferences about their proficiencies based on this evidence (Mislevy & Riconscente, 2006). We made many SG design decisions including game environment, statistical models, student work, and how analytics would be used to establish assessment arguments, described as follows.

3.4.1 Student and Evidence Models: What Behaviors or Performances Would Reveal Play-Learner Proficiency?

We used behavioral learning objectives to operationalize tasks of interest. These objectives drew on a social constructivist perspective and correspond to basic understanding tasks, understanding basic definitions, and more advanced application objectives such as considering alternative hypothesis or considering environmental factors. The student model quantifies the weight and direction of each variable as it is built to reveal something about the claim of what a student knows (Mislevy & Riconscente, 2006). It is the measurement model that begins to define the SG data based on observable variables and how those reveal something about student performance. Mislevy and Riconscente note that, "each piece of data correctly characterizes some aspect of a particular performance, but it also conveys some information about the targeted claim regarding what the student knows or can do" (p. 19). Our team quantified each student model variable (learning objective) and evaluation component as follows for our FAE example.

Table 12.2 Learning objectives and evaluation for FAE game elements

Learning objective	Evaluation
LO 1: identify the definition of FAE	Accuracy of bias definitions (correct responses)
LO 2: evaluate information to determine if FAE is at play within a given environment	Precision of identification of cognitive bias based on the given situational factors
LO 3: identify and select FAE based on situational factors	
LO 4: monitor your cognitive processes and delay making decisions (slowdown behavior) until you have considered alternatives—consider environmental factors	Consistency (measured by in-game time intervals) with which the play-learner delays decisions (slows down) and considers alternatives
LO 5: apply the appropriate mitigation strategy—consider environmental or situational factors before making a decision	Accuracy of mitigation strategy application based on the in-game situation

To identify how these factors would be measured, we identified key outcomes of the learning objectives, providing an evidence model for assessment. Table 12.2 provides the core learning objectives and their evaluation as used in the FAE elements of the CYCLES game. These ultimately drove the evidence model and task model design decisions.

3.4.2 Task Model: What Tasks or Situations Would Elicit Those Behaviors?

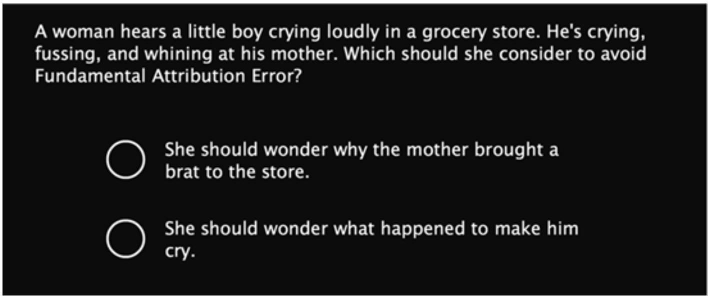
The task model of the CAF identifies specific actions or places in a learning experience where assessment can take place and involves identifying the environments in which the play-learner would perform activities. Through these performances data trails and evidence would accumulate which would then use to build claims about their proficiency. In the FAE challenges, the learning objectives were oriented around recognizing and avoiding this bias through a series of player judgments, described here in detail.

LO 1: Accuracy with which the play-learner identifies definitions (correct responses).

SG environment: Traditional multiple-choice quizzes were used after each room to collect evidence that play-learners could identify definitions of each bias (see Fig. 12.1). Play-learners were provided feedback on their answer to reinforce learning.

Data collected for analytics. Data were collected on the response selected and the length of time players took to select an answer.

Development perspectives. Quizzes were used to gauge basic understanding and to help students process the information more efficiently. Their design drew mainly on information-processing perspectives in order to leverage familiar and well-established testing techniques. There was some resistance to using quizzes from team members drawing on constructivist or game-based learning perspectives because of concerns



A woman hears a little boy crying loudly in a grocery store. He's crying, fussing, and whining at his mother. Which should she consider to avoid Fundamental Attribution Error?

○ She should wonder why the mother brought a brat to the store.

○ She should wonder what happened to make him cry.

Fig. 12.1 Quiz-based identification of FAE

that they were not sufficiently "game like" or that they diverged too much from the game setting.

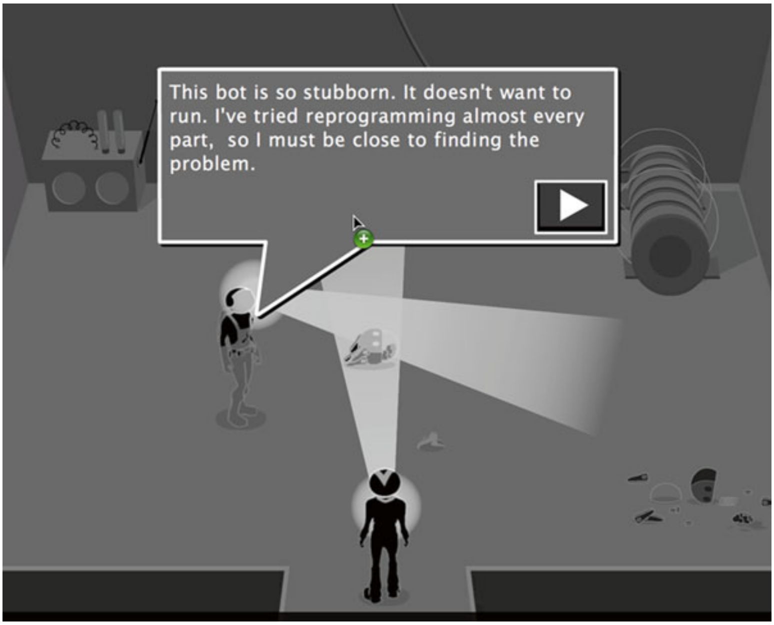
LO 2 and 3. Evaluate information to determine if FAE is at play within a given environment and identify and select FAE based on situational factors.

SG environment. The game environment would include situations where the blame could be placed on either the individual or environmental factors. We envisioned that the game would allow the play-learner to explore options that would indicate blame of the individual or explore options that indicated blame on the environment.

In the case of FAE, the play-learner entered a room where a technician is blaming a brain-driven robot for its apparent failure or malfunction. We use brain-driven robots in our scenario to make play-learners aware that cognitive biases are inherent in all human beings, the product of heuristics necessary to deal with cognitive load. Figure 12.2 provides a screen capture of and early implementation of this scenario within our game. In this screen capture you see the play-learner's avatar in all black. She is presented with a situation where a technician (dressed in overalls) is attempting to solve a problem with a brain-bot. In this situation the technician is blaming the brain-bot (i.e., the individual). It is the play-learner's responsibility to realize this and to react accordingly.

Data collected for analytics. Data is collected on the duration between mouse clicks (in Fig. 12.2 clicking on the next arrow) as the information is presented. Clicking through the content quickly provided evidence of not attending to the situation, whereas clicking through the content at an appropriate or average rate indicated that the play-learner was evaluating the situational information in an effort to determine which cognitive bias was at play.

Development perspectives. There was considerable debate about what was being learned within this situation. A primary concern was that play-learners were overloaded (extraneous cognitive load) with information about broken brain-bots, compass functions, and technicians. There was concern that these situational factors would overload the play-learner's memory and diminish learning. As was shown previously, quiz questions were added to the game to enhance information processing, whereas



This bot is so stubborn. It doesn't want to run. I've tried reprogramming almost every part, so I must be close to finding the problem.

Fig. 12.2 FAE room where technician is blaming the individual (brain-driven robot)

situation-based learning was added to the game to enhance social construction and explorative game-based play. Although these tensions between perspectives created difficulties at times, they were an important part of our design-build process. However, we feel that the ECD framework lacked a support structure for negotiating the complexities of managing the necessary explorations and negotiations among our team members.

LO 4. Monitor your cognitive processes and delay making decisions (slowdown behavior) until you have considered alternatives—in the case of FAE consider environmental factors.

SG environment. As was described above and presented in Fig. 12.2, the game environment would include situations where the blame could be placed on either the individual or environmental factors. We envision that in-game characters would behave in ways that were indicative of the cognitive bias being taught.

In this example, we designed a tool belt feature that rewarded play-learners for the desirable behavior of slowing down. In this scenario, we could have simply collected the amount of time between when the play-learner entered the room and when she made the next move or clicked on the next object. However the tool belt mechanism provided us with richer data for building our assessment argument in the situation. We designed the tool belt feature by considering a behavioral objective (slowing down) while simultaneously considering how the feature would collect data for analytics.

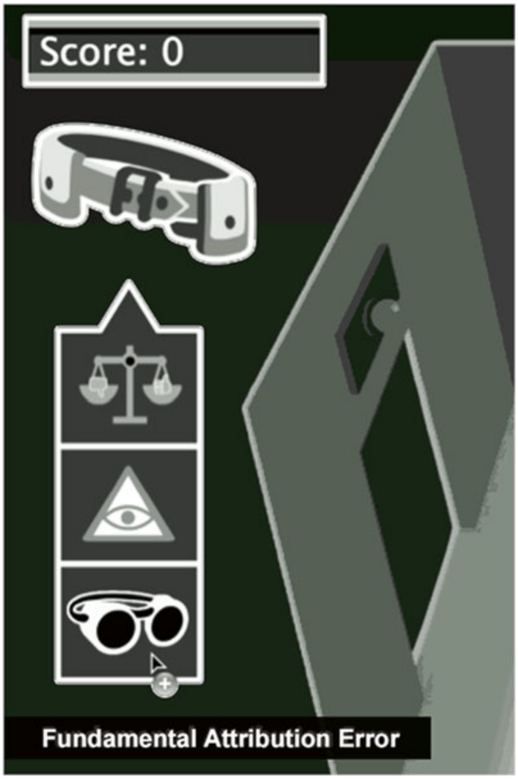


Fig. 12.3 The cognitive bias tool belt used to monitor (collect data) and reward players "slowdown" behavior

In the example above (see Fig. 12.2), the technician is focused on the problems with the brain-bot (the individual) and is ignoring environmental factors. The play-learner has the option to explore the brain-bot further following the lead of the technician or they could recognize that FAE may be at play and behave in ways to mitigate the bias.

The game explicitly teaches the play-learner to first identify the cognitive bias using the tool belt. This behavior is a metacognitive activity that is instantiated in a game-based activity. Although the play-learner is taught to slow down and evaluate each situation, they are not required to slow down and use the tool belt. Figure 12.3 provides a screenshot of the cognitive bias tool belt. This cognitive bias tool belt was designed to collect data, through play-learner behavior, on their understanding of the importance of first determining which cognitive bias is "at play" and then based on this determination, behave in ways that mitigate that cognitive bias (in this case examine or explore environmental factors).

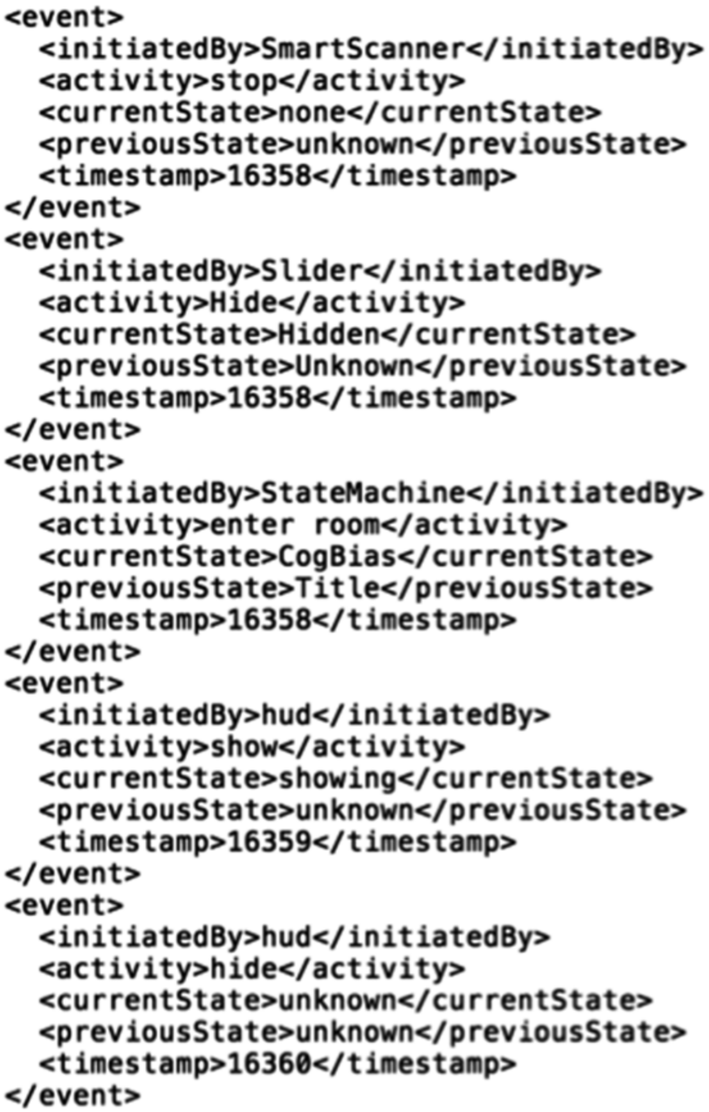
Although the slowdown behavior could be attributed to the newness or novelty of the tool belt itself, this game feature was a central game feature that the play-learner was required to use over-and-over again. Again, play-learners were encouraged to make this explicit slowdown and bias recognition behavior, as it was central to learning how to mitigate cognitive bias. Subsequently, play-learners were assessed on repeated performances of this slowdown behavior, their willingness to first identify the cognitive bias that may be "at play," and their efficiency in these performances.

Data collected for analytics. Computer log-file data was collected on which objects the play-learner clicked. The log-files were text files that included in-game actions and corresponding computer time-stamps. An example of a log-file is provided in Fig. 12.4. An important data collection feature was the cognitive tool belt. Data was collected on whether the play-learner slowed down and identified the cognitive bias before attempting to play through the room-based puzzle. Data was also collected on accuracy in response in identifying the correct cognitive bias through tool selection.

Development perspectives. During the design phase there was debate as to how many options within the environment should be available for the play-learners to explore. Some team members wanted to expand the individual and environmental options whereas others wanted to limit them. This debate was driven by two factors. First, we were required to limit the gameplay to 30 min. There was concern that too many options would extend gameplay. Second, there was concern that inundating play-learners with extraneous information would detract from learning. Again, this debate was driven by team members' orientation toward learning. For instance, the game developers wanted to add playful objects like bananas into the environment that could be explored or that were clickable. However these items were ultimately removed due to concerns of cognitive overload and their negative impact on learning.

More importantly there was considerable debate among team members as to whether the play-learner should be required to use the tool belt before progressing into a room. The debate around this game feature highlights the significance of team members' perspectives throughout design and development. Social constructivists wanted to allow for more freedom for the play-learner. Accordingly, they suggested the use of the tool belt to be completely optional, allowing some players to exhibit their thinking by first slowing down to identify the cognitive bias present within the game space. It would also allow for disparate behavior such as ignoring the tool belt and advancing to solve the in-game puzzle. In contrast, cognitive psychologists in the group advocated for a more linear and controlled approach requiring that the players utilize the tool belt before entering each puzzle. From their perspective this would reinforce the learning through spacing and testing effects by forcing them to use the tool belt (not allowing them to skip it). Game designers wanted the game to be more informative and instructive to the player, highlighting the need to use the tool belt and teaching the player within the game space how to proceed within the game. It was suggested that this be accomplished by prompting them with flashing arrows when it was time to use the tool belt.

Based on the FAE design pattern (see Table 12.1) and the behavioral learning objectives (see Table 12.2), we had extensive discussions about the tool belt game feature and its role within the game. Subsequently, during this debate we went through several design interactions the tool belt feature added and then removed, added again with reduced degrees of freedom, removed again, and then ultimately added back into the game with more degrees of freedom. These cyclical discussions

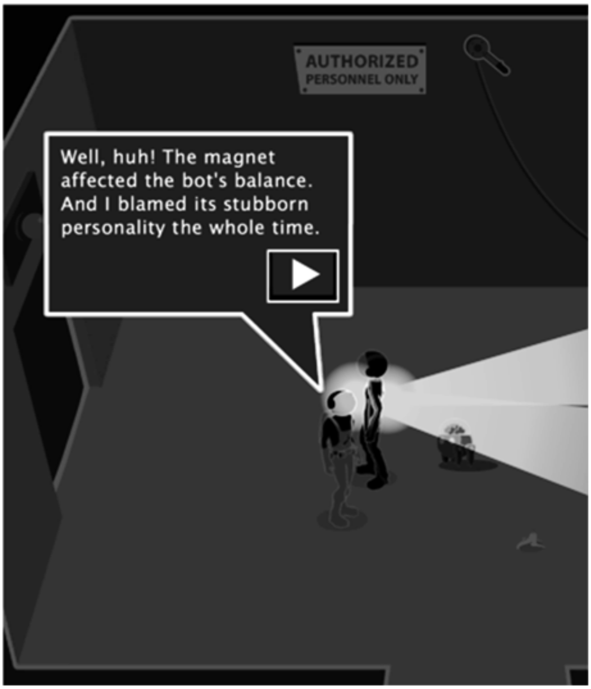


```
<event>
  <initiatedBy>SmartScanner</initiatedBy>
  <activity>stop</activity>
  <currentState>none</currentState>
  <previousState>unknown</previousState>
  <timestamp>16358</timestamp>
</event>
<event>
  <initiatedBy>Slider</initiatedBy>
  <activity>Hide</activity>
  <currentState>Hidden</currentState>
  <previousState>Unknown</previousState>
  <timestamp>16358</timestamp>
</event>
<event>
  <initiatedBy>StateMachine</initiatedBy>
  <activity>enter room</activity>
  <currentState>CogBias</currentState>
  <previousState>Title</previousState>
  <timestamp>16358</timestamp>
</event>
<event>
  <initiatedBy>hud</initiatedBy>
  <activity>show</activity>
  <currentState>showing</currentState>
  <previousState>unknown</previousState>
  <timestamp>16359</timestamp>
</event>
<event>
  <initiatedBy>hud</initiatedBy>
  <activity>hide</activity>
  <currentState>unknown</currentState>
  <previousState>unknown</previousState>
  <timestamp>16360</timestamp>
</event>
```

Fig. 12.4 CYCLES game log-file example

and subsequent design changes imposed significant additional costs (e.g., time, money, energy) to the project.

Although the challenges presented by competing psychological perspectives had been mentioned within the ECD literature, we not only underestimated the challenges that they would present to our team, but we were unsure how to deal with them using the ECD framework. Furthermore, we encountered these difficulties frequently and believe that a more formal process is necessary to enhance and



Well, huh! The magnet affected the bot's balance. And I blamed its stubborn personality the whole time.

Fig. 12.5 Mitigating FAE by considering other environmental actors

improve our ECD process. This process should include frequent discussions that identify psychological perspectives early and revisit them frequently as evidentiary data is designed into game-based features (a process is suggested below in the discussion section).

LO 5. Apply the appropriate mitigation strategy—in this case consider environmental or situational factors before making a decision.

SG environment. The game environment would include environmental factors that could be to blame for the problem. Considering these environmental factors is key to the FAE mitigation strategy.

Figure 12.5 provides a screen capture of the expanded environment that includes environmental factors. Play-learners were given the option to explore these environmental factors. For example, the play-learner can manipulate his/her avatar over to the wall and attempt to click on the switch. By exploring these environmental factors, before exploring the brain-bot (the individual), the play-learner was exhibiting mitigation behavior.

Data collected for analytics. Data was collected on which objects the play-learner explored and the sequences of those clicks and exploration.

Development perspectives. Again, driven by each team members' psychological perspective, there was considerable debate about these environmental objects. This included a debate about how many objects should be presented and what types of objects would be most beneficial to or simply distract from learning.

3.4.3 Assembly Model

The assembly model describes how multiple pieces of evidence come together to make a claim about student knowledge (Mislevy & Riconscente, 2006). In our FAE example, we assembled evidence from each task including correct responses on quiz-based questions, identification of biases based on situational factors, slow-down behaviors within each environment, and behaviors that related to mitigation. Within this assembly model the target tasks, accuracy and evidentiary data was identified. Table 12.3 provides a listing of these model elements for the FAE gameplay.

3.4.4 Sample Knowledge Representations

The sample knowledge representation is the general blueprint of assessment tasks and their details that are used for implementation. There are number of systems that have been developed, such as Principled Assessment Designs for Inquiry, that formalizes these representations (see Mislevy & Riconscente, 2006, for a review). As Table 12.2 illustrates, we chose to represent our assessment tasks by operationalizing behavioral learning objectives. Our team frequently revisited these learning objectives as a way to refocus our design efforts and design decisions. Revisiting these learning objectives frequently was effective in returning our team's focus to the principles outlined out by ECD and detailed in the definitions of student model variables, work products, evaluation procedures, and the task model variables.

3.5 Assessment Implementation

After finalizing the CAF, we moved to SG development and implementation. In this stage, we created assessment game features and simulations. Several of these game features have been described above (see Figs. 12.1, 12.2, 12.3, and 12.5) to illustrate the SG environment within the CAF. These FAE game features provide examples of how we translated CAF student model and corresponding behavioral learning objectives into a SG assessment implementation.

As noted previously, our CYCLES design team began developing our SG without a formal design framework. It was only over time, and out of necessity, that we adopted elements of the ECD framework to improve our game development process. We believe strongly that some of the difficulties we encountered in our early design iterations could have been minimized if we used a more formal design process such as ECD. In our early designs, we moved too quickly into the assessment implementation phase programming game features and simulations prior to establishing a solid CAF. A significant amount of the difficulties surrounded the differences between team members' perspectives on learning and cognition. We have attempted to provide clear examples of how these differing perspectives impacted

Table 12.3 FAE assembly model including task, accuracy and evidentiary data

Task	Accuracy	Evidentiary data
Correctly identifies definitions for each bias	100 %	Quiz question response within click-log data
Recognize relevant environmental fact and select the proper tool within the cognitive–bias tool belt	100 %	Tool selection based on click-log data
Exhibit the slowdown behavior, within game, by delaying their actions within the game environment	Determined based on historical averages of gameplay data	If a play-learner enters the room and very quickly clicks on the cognitive bias tool belt, we will assume that they did not consider what was going on in the room before doing so, or they failed to slow down. And, if a play-learner very quickly enters a room, does not click on the tool belt, but rushes to click on other objects with in the room, this will be evidence that they are not slowing down to consider which cognitive bias could be impacting their decisions
This will be measured by calculating the duration of time between when they enter a room (within game) and the time that they start clicking on objects within that room or the time that they click on the cognitive bias tool belt		
Apply an appropriate mitigation strategy. This will be collected and measured through click log sequences or chains of events within gameplay	Sequences of events logged that will provide evidence that the play-learner is behaving in ways to mitigate cognitive bias	Sequence of events:
		First: slow down
		Second: selection of correct cognitive bias (tool belt selection)
		Third: which objects the play-learner decides to investigate. In our FAE example the play learner could choose to investigate the brain-bot further (investigating the characteristics of the individual) or could choose to investigate and consider environmental factors, the correct behavior for mitigating FAE

SG design decisions. Although we continue to encounter tensions, it is through the exposure and acceptance of these differences that we have improved the design process and improved our SG design.

4 Conclusions: ECD and Personas

The number of SG research projects has risen rapidly since 2002. Many of these research projects have investigated existing games and their impact on learning. However as funding agencies begin to fund design and development work, researchers begin to cross the line into the complexities of game design, serious assessment design, and analytics. These complexities include managing and orchestrating large diverse teams of experts. These experts possess great domain knowledge that is steeped in an orientation to learning or psychological perspective towards knowledge acquisition in teaching. It is imperative that researchers who are attempting to cross this line and become effective game designers recognize the difficulties associated with these large-scale efforts. In designing for SGs it is important to adopt a design framework that is holistic and considers evidence-based assessment and evaluation throughout each design phase.

As was discussed previously, we began our multi-phase SG design and development project without such a design framework. We quickly realized that we needed a more formalized process and adopted elements of ECD. Although we did not adopt all elements of the ECD framework, the framework was extremely beneficial allowing us to more tightly orchestrate the activities of our team. It helped us focus our domain analysis, behavioral learning objectives, and subsequent game features on knowledge and skills that would collect the evidentiary data needed for our assessments.

ECD is a framework that provided a holistic approach from initial considerations of expert performance to design features that collect evidence of such performance. In this chapter, we presented a worked example of how we used the ECD framework, moving from identifying critical mitigation behaviors (mitigation of cognitive bias) moving those toward game-based features that would provide us with evidentiary data that indicate such proficiency. We feel that the descriptions presented in this chapter will significantly help others understand the importance of such a framework for the development of a SG that will support work and analysis in the emerging field of SEGA. Specifically, we hope that the detailed description of how we translated the desirable behaviors (based on expert behaviors) into behavioral learning objectives and then into game base features and subsequent evidentiary data will provide a concrete example for those researchers who find themselves attempting to design and develop SG for SEGA.

A critical perspective on this worked example may be that designing game features (e.g., the cognitive tool belt) and heavily rewarding behavior (slowing down) may change play-learners' in-game behaviors that don't translate to understanding. Although learning transfer and retention are outside the scope of this chapter, we have found medium to large training effect sizes in our initial analysis (Shaw et al., 2013). Our initial results indicate that the largest effect sizes appear when play-learners respond to questions that require them to mitigate each cognitive bias (situation-based questions) as opposed to simply recognizing a definition or discriminating between each bias. Although our initial findings are intriguing, the ability of play-learners to transfer learned behaviors to out-of-game situations warrants additional research.

Working with and attempting to orchestrate different perspectives among team members continues to be a significant challenge for our team. Although we recognize that our game design benefited from these differences, these differences continue to create situations that are less than optimal for our design and development process. We hypothesize that a more formal process for describing and discussing the differences between team members' orientation to learning and teaching would be beneficial. Additional research should focus on incorporating structures such as personas (Pruitt & Adlin, 2010) into the design process, which would help surface the differing perspectives of team members. Personas are fictional characters that are created to represent different designer types (designers working on an SG project) that might view an orientation to learning in similar ways. We propose that these personas be used to depersonalize, and thus reduce the tensions, these perspectives create. We theorize that these personas would be presented and discussed as an integral part of each ECD layer, as we found and have illustrated, these perspectives impacted decision in each layer. We theorize that by depersonalizing these discussions and by explicitly discussing each persona at each ECD layer, a stronger SG design, that incorporates the strengths of each perspective, will emerge.

Acknowledgements This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory contract number FA8650-11-C-7176. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

References

Al-Smadi, M., Wesiak, G., Guetl, C., & Holzinger, A. (2012, July). Assessment for/as learning: Integrated automatic assessment in complex learning resources for self-directed learning. In *2012 Sixth International Conference on Complex, Intelligent and Software Intensive Systems (CISIS)* (pp. 929–934). Palermo, Italy: IEEE.

Barab, S. A., Gresalfi, M., & Ingram-Goble, A. (2010). Transformational play using games to position person, content, and context. *Educational Researcher, 39*(7), 525–536.

Behrens, J. T., Mislevy, R. J., Bauer, M., Williamson, D. M., & Levy, R. (2004). Introduction to evidence centered design and lessons learned from its application in a global e-learning program. *International Journal of Testing, 4*(4), 295–301.

Behrens, J. T., Frezzo, D., Mislevy, R., Kroopnick, M., & Wise, D. (2007). Structural, functional and semiotic symmetries in simulation-based games and assessments. Assessment of problem solving using simulations. *Simulation-Based Games and Assessments, 4*, 59–80.

Boyd, J. R. (1995). The essence of winning and losing. A five slide set.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*, 354–338.

DiCerbo, K. E. (2014). Game-based assessment of persistence. *Educational Technology & Society, 17*(1), 17–28.

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422–444). Cambridge, MA: Cambridge University Press.

Gee, J. P. (2007). *What video games have to teach us about learning and literacy*. Revised and updated edition. New York: Macmillan.

Harvey, J. H., Town, J. P., & Yarkin, K. L. (1981). How fundamental is "the fundamental attribution error"? *Journal of Personality and Social Psychology, 40*, 346–357.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology, 3*(1), 1–24.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*, 697–720.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Macmillan.

Loh, C. S., Anantachai, A., Byun, J., & Lenox, J. (2007, July). Assessing what players learned in serious games: *in situ* data collection, information trails, and quantitative analysis. In *Proceedings of the computer games: AI, animation, mobile, educational & serious games conference*. Wolverhampton: University of Wolverhampton.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Mezirow, J. (1997). Transformative learning: Theory to practice. *New directions for adult and continuing education, 74*, 5–12.

Michael, D. R., & Chen, S. L. (2005). *Serious games: Games that educate, train, and inform*. Boston: Thomson Course Technology.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., Frezzo, D. C., & West, P. (2012). Three things game designers need to know about assessment. In *Assessment in game-based learning* (pp. 59–81). New York: Springer.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Mahwah, NJ: Erlbaum.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220.

Presser, A. L., Vahey, P., & Zanchi, C. (2013, June). Designing early childhood math games: A research-driven approach. In *Proceedings of the 12th International Conference on Interaction Design and Children* (pp. 376–379). New York: ACM.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin, 28*, 369–381.

Pruitt, J., & Adlin, T. (2010). *The persona lifecycle: Keeping people in mind throughout product design*. Amsterdam: Elsevier.

Reese, D. D., Tabachnick, B. G., & Kosko, R. E. (2014). Video game learning dynamics: Actionable measures of multidimensional learning trajectories. *British Journal of Educational Technology, 46*(1), 98–122.

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *The Journal of Technology, Learning and Assessment, 8*(4), 1–48.

Sanna, L. J., Schwarz, N., & Stocker, S. L. (2002). When debiasing backfires: Accessible content and accessibility experiences in debiasing hindsight. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(3), 497–501.

Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., & Bagley, E. (2009). Epistemic network analysis: A prototype for 21st-century assessment of learning. *International Journal of Learning and Media, 1*(2), 33–53.

Shaw, A., Kenski, K., Stromer-Galley, J., Martey, R., Clegg, B., Lewis, J., Folkestad, J. E., Strzalkowski, T. (2013). *Serious Efforts at Bias Reduction: The Effects of Digital Games and Avatar Customization on Three Cognitive Biases*. Washington, D.C.: National Communication Association.

Shute, V. J. (2011a). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction, 55*(2), 503–524.

Shute, V. J., & Ke, F. (2012). Games, learning, and assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning* (pp. 43–58). New York: Springer.

Shute, V. J., Ventura, M., Bauer, M., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning. In *Serious games*: *Mechanisms and effects* (Vol. 2, pp. 295–321). Philadelphia, PA: Routledge/LEA.

Shute, V. J., Masduki, I., & Donmez, O. (2010). Conceptual framework for modeling, assessing, and supporting competencies within game environments. *Technology, Instruction, Cognition, and Learning, 8*(2), 137–161.

Susi, T., Johannesson, M., & Backlund, P. (2007). Serious games: An overview. *Technical report HS-IKI-TR-07-001*. Sweden: University of Skövde.

Sweet, S. J., & Rupp, A. A. (2012). Using the ECD framework to support evidentiary reasoning in the context of a simulation study for detecting learner differences in epistemic games. *Journal of Educational Data Mining, 4*(1), 183–223.

Tetlock, P. E. (1985). Accountability: A social check on the fundamental attribution error. *Social Psychology Quarterly, 48*(3), 227–236.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

Winn, W. (2002). Current trends in educational technology research: The study of learning environments. *Educational Psychology Review, 14*, 331–351.

Chapter 13
Design of Game-Based Stealth Assessment and Learning Support

Fengfeng Ke and Valerie Shute

Abstract In this chapter, we describe the processes of designing and validating game-based learning assessment and/or support in two different games—Portal 2 (by Valve Corporation) and Earthquake Rebuild. The games represent cases of possible game-based learning (i.e., domain-generic and domain-specific) and provide good vehicles for testing the design decisions underlying stealth assessment and learning support. The chapter starts with a critical review of prior research on game-based assessment of competencies and learning support mechanisms in games, and then focuses on the particular design processes and findings from our two game cases. The review and findings suggested that process-oriented data mining and learning analytics methods help to capture the complex and open-ended learning trajectories in a game setting. They also illustrated how the evidence-centered assessment design and the learning context/task design should and can be interwoven in the early phase of game development. We conclude with a discussion relevant to developing and integrating the assessment and support of learning into other learning-game platforms.

Keywords Game-based learning assessment • Learning support • Stealth assessment

1 Introduction

In this chapter, we review and examine two important issues related to the next generation of learning games: (a) the real-time capture and analysis of gameplay performance data (i.e., game-based stealth assessment) and (b) the provision of adaptive learning supports based on the assessment information.

Historically, learning in games has been assessed indirectly and/or in a post hoc manner (Shute, 2011). What's needed instead is real time and valid assessment of learning based on the dynamic performance of players, which should be seamlessly

F. Ke (✉) • V. Shute
Florida State University, 3205-C Stone Building, Tallahassee, FL 32306-4453, USA
e-mail: fke@fsu.edu; vshute@fsu.edu

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based Learning, DOI 10.1007/978-3-319-05834-4_13

woven into the game to capture play-based competency development. This assessment information would then provide the basis for targeted and dynamic learner support. Incidental learning is often a consequence of playing well-designed games (MacCallum-Stewart, 2011; Prensky, 2001). However, creating substantive, improvisational learning experiences in games is difficult because knowledge and skill acquisition usually involves conscious elements (e.g., processing information, constructing mental models) in addition to subconscious processes, such as insight. A relevant game design hypothesis is that learning within gameplay will proceed from being improvisational (i.e., acting spontaneously in the environment without pre-planning) to meta-reflective (i.e., considering various points of view), or moving from a tacit experience to an aware, strategic, and reflective application of the target knowledge/skills. The underlying challenge of this design hypothesis is to integrate the learning-analytics-based support (or scaffolding) of meta-reflective learning into the game world and mechanics while not disrupting what is enjoyable about games.

In this chapter, we will describe the processes of designing and validating game-based assessment and/or learning support in two games. The first game, Portal 2 (by Valve Corporation), is an existing, commercial off-the-shelf (COTS) game that we hypothesized would foster spatial skills. The second game, Earthquake Rebuild, is currently under development. It is an architectural game that aims to promote mathematical understanding and math-related problem-solving skills. The two games represent typical cases of possible game-based learning (i.e., domain-general and domain-specific) and provide good vehicles for testing the design decisions underlying game-based stealth assessment and learning support. The chapter starts with a critical review of prior research on game-based assessment of domain-relevant competencies and learning support mechanisms in games, and then focuses on the particular design processes and findings from our two game cases. We conclude with a discussion relevant to developing and integrating the assessment and support of learning into other learning-game platforms.

2 Literature Review

There is rapidly growing interest in data mining and analytics in education, learning sciences, and other academic fields. Research on the automated collection or monitoring of user-generated data has been conducted in multiple fields, such as that on telemetry in computer science (Yairi, Kawahara, Fujimaki, Sato, & Machida, 2006) and geospatial data mining in Geographic Information System (Miller & Han, 2009). Educational data mining (EDM), highlighted in this chapter, is the process of exploring and extracting descriptive patterns from large amounts of data—“big data”—in educational settings (e.g., logs of student–computer interaction) to provide insights into instructional practices and student learning (Baker & Yacef, 2009; Romero, Ventura, Pechenizkiy, & Baker, 2011; Witten & Frank, 2005). In recent years, EDM has been used to infer students’ computer-supported learning engagement and behaviors and hence the development of effective and dynamic learning support (e.g., Baker, 2007; Baker, Corbett, & Koedinger, 2004; Beck & Mostow, 2008; Shute, Ventura, & Kim, 2013).

Closely related to EDM is the learning analytics research that refers to collecting, measuring, analyzing, and reporting data about learners and contexts to understand and optimize learning and the environments in which it occurs (SoLAR, 2011). Similar to EDM, learning analytics (LA) focuses on data-intensive approaches to education, although EDM often uses automated discovery with models while LA leverages more human judgment (Siemens & de Baker, 2012).

Prior research has suggested that both EDM and LA can and should be used together to exploit game-based performance data to inform on students' attributes, their on-task or off-task behaviors, competency development related to the targeted subject matter, and hence the effectiveness and design of learning supports. Four recent projects (by Dede, 2012; Levy, 2014; Shaffer et al., 2009; Shute et al., 2013) can exemplify the current state of game-based learning assessment via EDM and/or LA.

2.1 Game-Based Learning Assessment Through Data Mining and Analytics

Evaluation of Save Patch: In a recent study, Levy (2014) employed the approach of evidence-centered assessment design (Mislevy, Steinberg, Almond, & Lukas, 2006) and the method of Bayesian Networks to evaluate student performance in Save Patch, an educational game targeting rational numbers in math.

The process started by a cluster analysis that classifies the gameplay log to extract a list of solution strategies (behaviors) for successful gameplay actions, and that of misconceptions associated with unsuccessful actions. The cluster analysis results then served as categories of values of observable variables. For each targeted math competency in the game, Levy (2014) specified a dichotomous latent variable with its categorical values as mastery (coded as 1) and nonmastery (coded as 0). A dichotomous latent variable was also specified for each of the misconceptions, with its categorical values coded as 1 or 0 based on whether the student possessed that misconception or not. A Bayesian network model was created and calibrated to investigate conditional probabilities of observable categories (values) of each latent variable for individual students (or specific student groups) at different points in time and for each game level. The constructed psychometric model also encompassed transitions from nonmastery to mastery in certain latent variables by specifying the probability that a student is a master at time $t+1$, given they were a nonmaster at time t and had a particular value for the observable at time t.

Game-based stealth assessment: Similar to Levy (2014), Shute, Masduki, and Donmez (2010) and Shute, Ventura, and Kim (2013) adopted the approach of *evidence-centered assessment design* (*ECD*) to design and validate the framework of educational assessments in terms of user-generated in-game (gameplay) data, named stealth assessment, in various game evaluation studies. For example, Shute and her colleagues described the development of competency, evidence, and task models for the assessment of systems thinking in the game Taiga Park (Shute, Masduki, & Donmez, 2010). The ECD-based, stealth assessment framework has also driven the design and validation of Physics Playground (formerly called Newton's Playground), a learning game intended to help secondary school students understand qualitative physics (Shute & Ventura, 2013). The central evidentiary component of stealth assessment for Physics Playground is the game log file that captures multiple gameplay variables (e.g., time spent on a level, number of trials, types of objects created, the trajectory of objects, number of gold trophies obtained). Analyses revealed a significant correlation between in-game assessment indicators (e.g., gold trophies earned) and the external learning measure (qualitative physics test score). Furthermore, students (167 middle school students) significantly improved on the external physics test (administered before and after gameplay) despite no formal instruction in the game. Students also enjoyed playing the game (reporting a mean of 4 on a 5-point scale in where 1 = strongly dislike and 5 = strongly like), and boys and girls equally enjoyed the game.

Epistemic network analysis for epistemic games: Different from the aforementioned studies that emphasized data mining with a quantitative, psychometric modeling approach, Shaffer and his colleagues (2009) adopted a learning analytic approach by collecting and analyzing qualitative data from the game-extended records and interactions in gameplay. Specifically, they performed a systematic coding and aggregation with the qualitative data to identify salient elements of an epistemic frame (i.e., competency), then quantified the coded results by calculating the co-occurrence frequency of each pair of epistemic frame elements. They then created a cumulative network graph that is similar to a social network, "where frame elements (nodes) that are linked more often in the data are closer to each other than those that are linked less often in the data" (p. 7). In the structural network analysis, the unit of analysis is a strip (or segment) of activity "into which ongoing activities are divided for the purpose of analysis" (p. 8). By summing the strips of activities up to a particular time, the trajectory of development of an epistemic frame can be mapped as a dynamic network graph, or series of slices (or phases) over time, with each slice showing the state of the players' epistemic frame at certain time. A descriptive, visual comparative analysis can then be used to examine the trajectories of frame development of a subgroup of players (e.g., novices versus experts), or the knowledge structure of the targeted competency (e.g., by computing the relative weight or centrality of each node in the epistemic network).

EcoMUVE assessment: With an emphasis of data visualization, Dede (2012) described multiple analytical methods relating to learning trajectories in virtual-reality-based, complex inquiry tasks. These methods include:

- Event path analysis and visualization via the heat map. This method involves using the server-side log data to generate event paths and then providing a visual and diagnostic analysis on players' scientific inquiry skills. The path analysis comprises a series of visual slides depicting the relative frequencies of learning events performed by subpopulations of students, aggregated by prespecified virtual-world location and time unit, for comparative analyses (e.g., high-performing vs. low-performing students). The heat map shows which hotspots the players prefer—where hotspots are highlighted and can be used diagnostically to inform various misconceptions.

- Behavior analysis with the usage of guidance tools and pedagogical agents. This process uses the prediction analysis (e.g., regression and correlation analyses) to examine the effects of various learning support mechanisms and how they relate to student performance. The guidance tool uses individual players' interaction histories to generate real-time, customized support.
- Structured benchmarking task assessments. The last method entails a series of mini-modules (or inquiry tasks) in the virtual reality environment. The tasks are created as benchmarking assessments to provide information on skill mastery and promote transfer of learning.

In summary, the aforementioned game-based learning projects illustrate the multifaceted nature of assessment through data mining and learning analytics. All projects adopt a data-intensive, evidence-based approach, but differ in terms of: (a) the assessment objectives (i.e., to model or predict students' competency development, or to analyze the structure of domain-specific competency or epistemic frame, or to examine the association between the learning trajectory and the learning support and context design), (b) the resources of data (e.g., in-game log data, or game-extended behaviors), (c) analysis methods (e.g., quantitative psychometric modeling, network or structural analysis, and path analysis), and (d) type of visualization (e.g., algorithms, models, network graphs, or spatial and chronical maps).

2.2 In-Game Learning Support

In a recent meta-analysis that synthesized 29 studies on instructional/learning support in game-based learning, Wouters and Van Oostendorp (2013) classified learning support features into two major categories—ones that support the selection of relevant information, and ones that facilitate information organization and integration via reflection and explication. Of the articles reviewed in the study, more than half explicitly studied the in-game learning support features. These in-game support features are based on their associations with game-design elements (i.e., game world, game actions, and rules) and can be categorized as: (a) cues and feedback, (b) explicit training or instruction, (c) probes or prompts for self-explanation and reflection, (d) in-game learning tools, (e) incentive structures, and (f) level sequencing or progression.

Adaptive instructional or learning support is emerging as a prominent feature of serious and learning games (Kickmeier-Rust & Albert, 2010; Leemkuil & de Jong, 2012; O'Rourke, Haimovitz, Ballweber, Dweck, & Popović, 2014). Adaptive feedback, intelligent pedagogical agents, and adaptive level progression, in particular, feature prominently in such games. For example, O'Rourke et al. (2014) designed four metrics (named "brain points") to capture and reward players' novel and incremental content-related game performance. They found that the "brain points" version of the game, in comparison with a control version of the game, increased overall time played, strategy used, and perseverance after challenge. Hwang, Sung, Hung, Huang, and Tsai (2012) examined the role of game level sequencing or navigation in a role-playing science game. They reported that students who learned with the personalized game level sequencing (by matching their learning styles with the game level navigation style—linear or nonlinear) showed significantly greater learning achievement, motivation, and acceptance towards game-based learning than those who learned with the game without personalized sequencing. These support tools or mechanisms are typically based on the nonintrusive, stealth assessment of in-game performance via the creation and tracking of evaluation indices and threshold values (Shute et al., 2013; Zapata-Rivera, VanWinkle, Doyle, Buteux, & Bauer, 2009).

We now present two more detailed examples of game-based assessment design. In Example 1, we report a completed, controlled evaluation of domain-generic skills development in a COTS game. In Example 2, we describe how the development of a domain-specific, stealth assessment mechanism is aligned and associated with the design of the game world, game mechanics, and in-game learning support in an underdeveloped math learning game.

3 Game-Based Learning Assessment Design for Portal 2

3.1 Portal 2

Portal 2 is the name of a popular linear first-person puzzle-platform video game developed and published by Valve Corporation. Players take a first-person role of Chell in the game and explore and interact with the environment. The goal of Portal 2 is to get to an exit door by using a series of tools. The primary game mechanic in Portal 2 is the portal gun, which can create two portals. These portals are connected in space, thus entering one portal will exit the player through the other portal. Any forces acting on the player while going through a portal will be applied upon exiting the portal. This allows players to use, for example, gravity and momentum to "fling" themselves far distances through the air. This simple game mechanic is the core basis of Portal 2.

Other tools that may be used to solve puzzles in Portal 2 include Thermal Discouragement Beams (lasers), Excursion Funnels (tractor beams), Hard Light Bridges, and Redirection Cubes (which have prismatic lenses that redirect laser beams). The player must also disable turrets (which shoot deadly lasers) or avoid their line of sight. All of these game elements can help in the player's quest to open locked doors, and generally help (or hinder) the character from reaching the exit.

The initial tutorial levels in Portal 2 guide the player through the general movement controls and illustrate how to interact with the environment. Characters can withstand limited damage but will die after sustained injury. There is no penalty for falling onto a solid surface, but falling into bottomless pits or toxic pools kills the player character immediately.

There are several plausible ways for a person to acquire and hone spatial skills as a function of gameplay in Portal 2.

3.2 Spatial Skills

Of particular importance in understanding the role of video gameplay relative to spatial cognition is the distinctions among: (1) figural, (2) vista, and (3) environmental spatial skills (Montello, 1993; Montello & Golledge, 1999). Figural spatial skill is small in scale relative to the body and external to the individual. Accordingly, it can be apprehended from a single viewpoint. It includes both flat pictorial space and 3D space (e.g., small, manipulable objects). It is most commonly associated with tests such as mental rotation and paper-folding tasks. Vista spatial skill requires one to imagine an object or oneself in different locations—small spaces without locomotion. Vista spatial skill is useful when trying to image how the arrangement of objects will look from various perspectives (Hegarty & Waller, 2004). Environmental spatial skill is large in scale relative to the body and is useful in navigating around large spaces such as buildings, neighborhoods, and cities, and typically requires locomotion (see Montello, 1993, for a discussion of other scales of space). It usually requires a person to mentally construct a cognitive map, or internal representation of the environment (Montello & Golledge, 1999). Environmental spatial skill depends on an individual's configurational knowledge of specific locations in space and is acquired by learning specific routes. Configurational knowledge depends on the quality of an individual's cognitive map, or internal representation of an environment. In this map-like representation, all encountered landmarks and their relative positions are accurately represented.

A game-like Portal 2 has the potential to improve spatial skills due to its unique 3D environment that requires players to navigate through problems in often complex ways. Over the past 20 years, a growing body of research has shown that playing action video games can improve performance on tests of spatial cognition and selective attention (e.g., Dorval & Pepin, 1986; Feng, Spence, & Pratt, 2007; Green & Bavelier, 2003, Spence, Yu, Feng, & Marshman, 2009; Uttal et al., 2012). Recently, Ventura, Shute, Wright, and Zhao (2013) showed that self-reported ratings of video game use were significantly related to all three facets of spatial cognition, and most highly related to environmental spatial skill. Feng et al. (2007) found that playing an action video game improved performance on a mental rotation task (i.e., small-scale or figural spatial cognition). After only 10 h of training with an action video game, subjects showed gains in both spatial attention and mental rotation, with women benefiting more than men. Control subjects who played a nonaction game showed no improvement.

Recently, Uttal et al. (2012) conducted a meta-analysis of 206 studies investigating the effects of training on spatial cognition. Of these 206 studies, 24 used video games to improve spatial skills. The effect size for video game training was .54 (SE = .12). Findings like these have been explained due to the visual-spatial requirements of 3D action games which may enhance spatial skills (e.g., Feng et al., 2007; Green & Bavelier, 2003, 2007).

3.3 External Measures of Spatial Skills

To both validate the in-game (stealth) measures of spatial skills (e.g., number of portals shot on average, per level) and test for any learning of them from 8 h of gameplay, Shute et al. used three existing, validated assessments for figural, vista, and environmental spatial skills. To measure figural (or small-scale) spatial skill, they used the Mental Rotation test (Vandenberg & Kuse, 1978). To assess vista spatial skill, they administered the Spatial Orientation Test (Hegarty & Waller, 2004). And to measure environmental spatial skill, they developed and validated an assessment called the Virtual Spatial Navigation Assessment. Each is now described.

- *Mental Rotation Test* (MRT). The MRT was adapted from Vandenberg and Kuse (1978). In this test, participants view a three-dimensional target figure and four test figures. Their task is to determine which of the test figure options represent a correct rotation of the target figure. The total score is based on the total number of items where both correct objects are found.
- *Spatial Orientation Test* (SOT). The SOT requires the participant to estimate locations of objects from different perspectives in one picture (Hegarty & Waller, 2004). In each item the participant is told to imagine looking at one object from a particular location in the picture and then point to a second location. Each response is scored as a difference between the participant's angle and the correct angle (scores range from 0 to 180°). Larger differences between a participant's drawn angle and the correct angle indicate lower vista spatial skill.
- *Virtual Spatial Navigation Assessment* (VSNA). The VSNA (Ventura et al., 2013) was created in Unity. In the VSNA, a person explores a virtual 3D environment using a first person avatar on a computer. Participants are instructed that the goal is to collect all the gems in an environment and return to the starting position. Participants first complete a short familiarization task that requires them to collect colorful gems in a small room. The VSNA consists of an indoor environment consisting of halls in a building (i.e., a maze), and an outdoor environment consisting of trees and hills. In each environment the participant must collect the gems twice—training and testing phases. The VSNA collects data on the time taken to collect all gems and return to the starting position, as well as the distance traveled in the training and testing phase of an environment. The main measure used in the current study consists of the time to collect all gems and return home. Less time suggests greater navigational skill.

3.4 Results from a Controlled Evaluation of Portal 2

A recent study reported by Shute, Ventura, and Ke (2015) tested 77 undergraduates who were randomly assigned to play either Portal 2 or a control game condition (i.e., the popular brain training game suite called Lumosity) for 8 h. Before and after gameplay, participants completed a set of online tests related to their spatial skills. Results revealed that participants who were assigned to play Portal 2 showed a statistically significant advantage over Lumosity on the composite measure of spatial skill. Portal 2 players also showed significant increases from pretest to posttest on specific small-scale (MRT) and large-scale (VSNA) spatial tests while those in the Lumosity condition did not show any pretest to posttest differences on any measure. Finally, Portal 2 *in-game performance data* (e.g., number of portals shot on average, per level) significantly correlated to MRT and VSNA after controlling for the respective pretest scores. These findings suggest that performance in Portal 2 predicts outcomes on different (small- and large-scale) spatial measures beyond that predicted by their respective pretest scores.

The improvement of subjects on their spatial skills as a function of playing Portal 2 is likely due to the repeated requirement in Portal 2 to apply and practice their spatial skills to solve problems. This result supports other work investigating video game use and spatial skill (e.g., Feng et al., 2007; Uttal et al., 2012; Ventura et al., 2013). There were no improvements for the Lumosity group on any of the three spatial tests. Overall, the findings of between-group differences on the MRT and VSNA measures, combined with the significant Portal 2 pretest–posttest gains in MRT and VSNA, give strong evidence that playing Portal 2 causes improvements in small- and large-scale spatial skills. Moreover, the fact that a conservative control group was used gives even greater credence to the finding that playing Portal 2 can improve spatial skills over other game-related activities that claim to improve cognitive skills (i.e., Lumosity games). Finally, while video gameplay has been previously shown to improve MRT performance (e.g., Uttal et al., 2012), this is the first research study to provide experimental evidence that video game play can improve performance in large-scale spatial skill.

4 Game-Based Learning Assessment and Support Design for Earthquake Rebuild

Earthquake Rebuild (E-Rebuild) is a 3D architecture game that intends to promote versatile representation and epistemic practice of mathematics in design and building quests (Ke, Shute, Erlebacher, Clark, & Ventura, 2014) and is on the development and user testing phase at the time of writing. The overall goals of E-Rebuild are to plan, design, and rebuild an earthquake-damaged space to fulfill diverse design parameters and needs. The intermediate game goals involve completing each level of the design quest to gain new tools, construction materials, and credits (e.g., game scores in terms of architectural design efficiency, structural soundness, and complexity in structures—which comprise an overall credit that enables a player to perform subsequent game levels).

A learner in E-Rebuild performs multiple types of gameplay (or architectural design) actions: *collection*, *construction*, space and energy *allocation*, and materials *trading*. All four gameplay actions act as both the source and the application of math understanding. The target math topics of E-Rebuild, aligned with the Common Core State Standards (CCSS) for mathematics Grade 6–8 (CCSSI, 2011) are: (a) ratio and proportional relationships; (b) angle measure, area, surface area, and volume; and (c) numeric and algebraic expressions.

Different from most projects in which learning assessment design (via EDM or LA) is a post hoc practice conducted after game development, E-Rebuild is integrating stealth assessment design directly into the game design process. This section introduces the process of interweaving assessment of learning and game design in this *ongoing*, design-based research project.

4.1 Interweaving the Design of Game World and that of the Game Log File

A major component of the game world design in E-Rebuild is to design various game objects, such as constructional materials (e.g., planks, pillars, bricks, prefabricated container houses) and game characters (e.g., victims or residents to be accommodated). The design of the relationship structures and the properties of these objects are aligned with the design of the game log file in terms of the variables and events logged. For example, the key properties of each construction element include its mass (solid vs. hollow, primitive vs. composite), texture, geometric form, size, volume, location, and position or angle. With each object and its element there will be a list of potential actions to be performed, such as clicking, moving, joining, cutting, and scaling. The original state of the objects' properties, the specific actions performed, and hence the state or characteristic change (e.g., increased happiness) following the actions performed (along with the time stamp and occurrence frequency) will all be captured in the game log file for a future sequential analysis.

4.2 Aligning Game Mechanics with Competency-Based Learning Actions

To enable an authentic, performance-based assessment, we align the E-Rebuild game mechanics (i.e., gameplay actions and rules) with math learning actions. Specifically, the integration of two gameplay modes (i.e., the adventure and

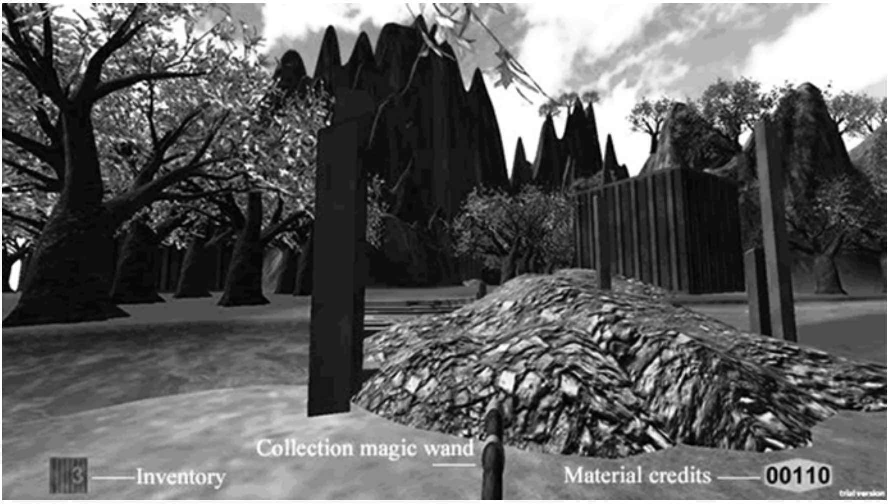


Fig. 13.1 Collection action in the adventure mode

construction modes) aims to extract integral, multistranded math learning actions. That is, in the adventure mode (see Fig. 13.1), players are requested to engage in exploration- and collection-based math concept representation (e.g., identifying a construction item in a specific prism and size) and experience-based reflection (e.g., evaluating their math-specific design performance by seeing how a designed structure collapsed in the earthquake or failed to address the needs). In the construction mode (see Fig. 13.2), players are mainly involved in construction-oriented math calculation and problem-solving (e.g., cutting/scaling an item to a desirable size, measuring/rotating the construction site based on a landmark, managing materials).

4.3 Designing Game Tasks Based on the Competency and Evidence Models

A *game task library* is being developed based on the target math competencies and the corresponding specifications of the competency and evidence models for the game-based stealth assessment. The competency and evidence models are being explicitly aligned with the Common Core State Standards (CCSS). They follow the structure of a Bayesian network and have guided the design of specific game tasks and the arrangement of these tasks within and across game levels.

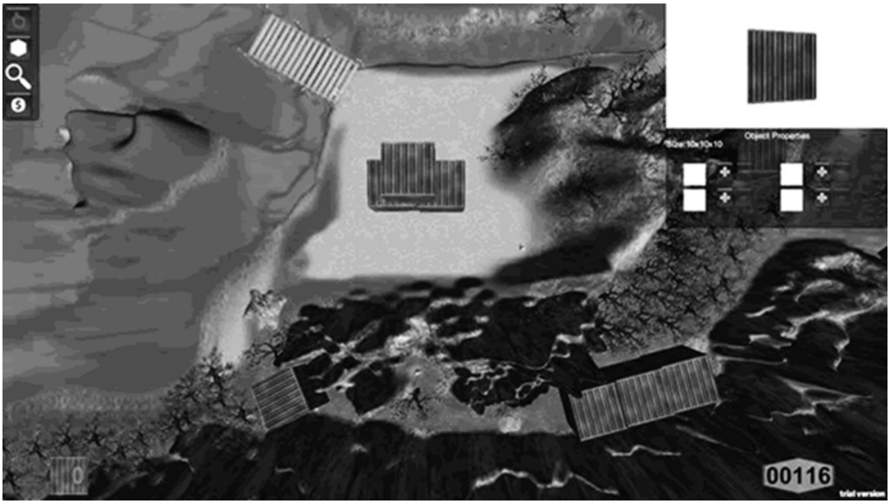


Fig. 13.2 Construction action in the construction mode

4.4 Representing Learning in the Game-Scoring Mechanism

The major variables used to evaluate successful knowledge and skill acquisition in E-Rebuild include: (1) *time* taken to complete the current task (e.g., tasks are speeded with a risk/progress bar related to an earthquake-hit) and (2) successful handling of multiple *design constraints* imposed by the needs of the area's residents, the landscape, and the limited construction materials. The first criterion measures the fluency while the second criterion measures the accuracy of math-related architectural problem-solving performance. A composite game score, along with sub-scores embedded in the game reward mechanism (e.g., time credit, material credit, happiness of residents), is then calculated based on the evaluation of the aforementioned evaluation criteria and presented to portray a player's learning profile.

4.5 Learning Support Design as Both the Source and the Application of Data Mining

Intuitive interfaces are important to successful human–computer interactions. In E-Rebuild, we design in-game learning supports as an intermediary interface between the player and the game (Fig. 13.3). This interface will support content engagement during gameplay while capturing the processes related to solving a complex math task. For example, a user-testing, comparative analysis with the control-meter interface and the current text-entry box (for feeding numerical values

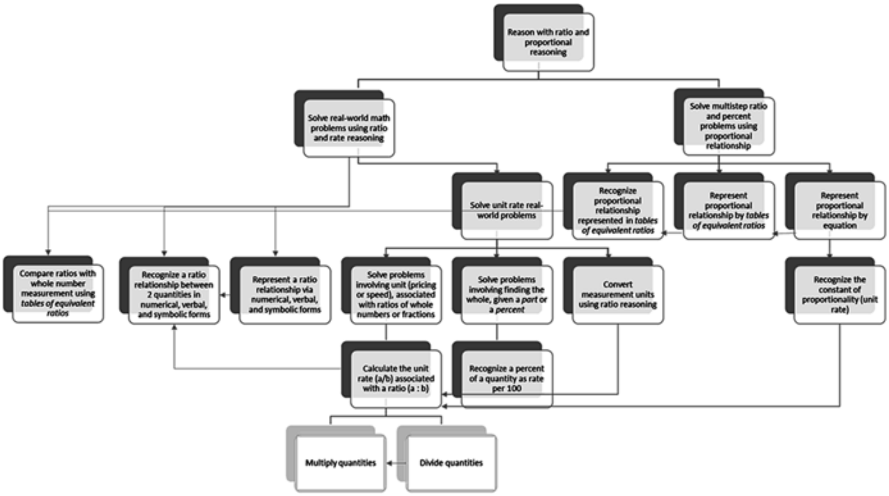

Reason with ratio and proportional reasoning
Solve real-world math problems using ratio and rate reasoning
Solve multistep ratio and percent problems using proportional relationship
Solve unit rate real-world problems
Recognize proportional relationship represented in *tables of equivalent ratios*
Represent proportional relationship by *tables of equivalent ratios*
Represent proportional relationship by equation
Compare ratios with whole number measurement using *tables of equivalent ratios*
Recognize a ratio relationship between 2 quantities in numerical, verbal, and symbolic forms
Represent a ratio relationship via numerical, verbal, and symbolic forms
Solve problems involving unit (pricing or speed), associated with ratios of whole numbers or fractions
Solve problems involving finding the whole, given a *part* or a *percent*
Convert measurement units using ratio reasoning
Recognize the constant of proportionality (unit rate)
Calculate the unit rate (a/b) associated with a ratio (a : b)
Recognize a percent of a quantity as rate per 100
Multiply quantities
Divide quantities

Fig. 13.3 Competency model of the ratio and proportional reasoning defined by CCSS for grades 6–8

of the x, y, and z coordinates in a scaling tool, see Fig. 13.4) indicates that the text-entry box is obviously associated with less wild guessing or trial-and-error play and more mindful math calculations. Every attempt of using this specific scaling tool, along with the values entered, is captured in the game log file to enable a diagnostic analysis. The results of the diagnosis will then be presented as dynamic feedback in a Scratch Pad screen (see Fig. 13.5). This scratch pad also includes an internal calculator and enables the typing of calculation steps, thus working as the record of mathematical processing performed by the player for the future data mining.

5 Conclusions, Discussion, and Future Research

5.1 Heuristics of Game-Based Learning Assessment Design

A salient feature of the aforementioned game-based assessment projects is the diagnostic and formative measurement of multiple domain-relevant steps or cognitive processes underlying each task solution or performance. Process-oriented data mining and learning analytics methods, such as Bayesian networks, social networks or structural analysis, visual or graphical analysis of event paths, and sequential analysis of time series, will capture the complex and open-ended learning trajectories in a game setting.

Game-based assessment should leverage and integrate both quantitative, model-based automatic discovery and qualitative interpretation with human judgment. Frequently, the interpretation and extraction of meaningful patterns from the game

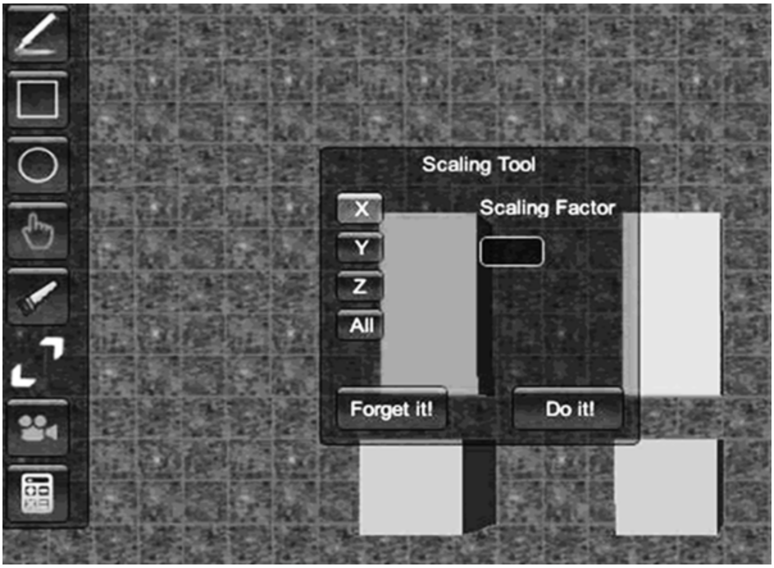
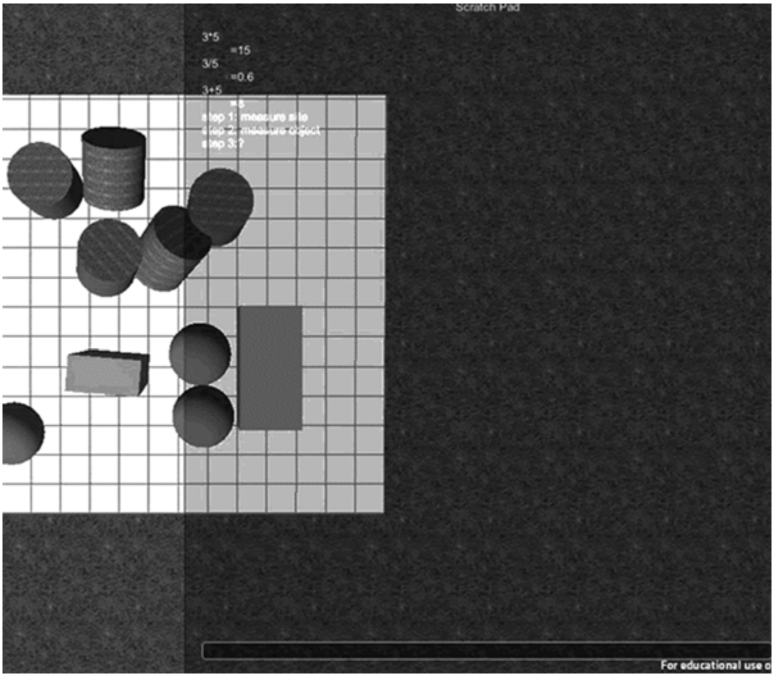


Fig. 13.4 Scaling tool



```
3*5
    =15
3/5
    =0.6
3+5
    =8
step 1: measure side
step 2: measure object
step 3: ?
```

Fig. 13.5 Scratch pad

log or extended performance data are in need of the perspectives and expertise of stakeholders (e.g., content experts, game designers, student users). The rules for evidence identification and the categorization of observable values also emerge based on the integration of expert decision and data-driven calibration. The sources and products of analytical methods in game-based learning assessment, as the examples in this chapter illustrated, comprise not only numerical values and algorithms but also discourses, descriptive frames, and graphical models.

Notably, prior work in this area, as well as our own research has suggested that performing diagnostic and stealth assessment with game-based learning is especially challenging when the assessment strategies are a post hoc design decision enforced on an existing game. Both Dede (2012) and Levy (2014) have reported on particular challenges of using data mining or learning analytics techniques to evaluate learning in an existing game or simulation. These challenges include but are not limited to the difficulty of inferring knowledge-mastery transition due to the insufficient and unbalanced task-specific data across game levels, and the difficulty of mapping the event path when a game-based learning task does not involve location exploration. Similarly, it is difficult to collect and analyze all action-based evidences of spatial skills in the Portal 2 study since only part of gameplay actions or object attributes were recorded in the game log file. Moreover, the tagging of variables and events in the current game log file of Portal 2, like that of many commercial games, changes across game levels and makes it extremely difficult to interpret and clean the log data for an automatic pattern discovery. In other words, the strategy and scope of data recording are not well aligned with the method and objective of stealth assessment when the assessment design occurs *after* game development.

A promising solution to the above challenge, as argued by Dede (2012) and Shute and Ventura (2013), is to interweave the evidence-centered assessment design and the learning context/task design in the early phase of game development. The E-Rebuild project illustrated that the development of domain-relevant competency, evidence, and task models should underlie the design and sequencing of tasks within and across game levels. The design of the game log file, in terms of the log's content, structure, and tagging, should be aligned with the game world design and evidence identification rules to enable automatic data cleaning and processing.

5.2 Implications for Future Game Design and Evaluation Efforts

This chapter has focused on methods for achieving two interrelated goals that we believe can have a significant impact on both formal and informal learning. The first goal is to get more children, particularly females and certain underrepresented minorities (e.g., Black and Hispanic children), excited about and interested in developing STEM-related skills and knowledge—such as spatial skills and understanding ratios and proportional reasoning (which serve to undergird many higher math areas).

Recognizing that interest alone is not enough, our second goal is to identify ways to facilitate and deepen learning in immersive, rich, and authentic environments. Well-designed digital games represent a promising vehicle for meeting both goals: capturing children's interest in STEM fields in general, and supporting their learning. More research is needed about the optimal design to be used for valid assessments and real-time learning support. We agree with the conclusion presented by Clark et al. (2011) that more research is needed that provides "supports for students to help them articulate their intuitive understandings from game play with the explicit formal concepts and representations of the discipline" (p. 2192). Our future research will focus on iterative design processes to refine the integration of stealth assessment and learning support in E-Rebuild. Data will be collected via both qualitative and quantitative methods over time to build up a body of evidence on the design generalizations and effectiveness of the learning game and its assessment/support mechanism.

Acknowledgments The research was supported by the National Science Foundation, Award Number #1318784 and the James T. and Catherine D. MacArthur Foundation #11-99517. We also would like to thank Matthew Ventura for his work on the Portal 2 project—particularly relative to spatial skill assessment.

References

Baker, R. S. (2007, April). Modeling and understanding students' off-task behavior in intelligent tutoring systems. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 1059–1068). San Jose, CA: ACM.

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004, January). Detecting student misuse of intelligent tutoring systems. In *Intelligent tutoring systems* (pp. 531–540). Berlin: Springer.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3–17.

Beck, J. E., & Mostow, J. (2008, January). How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. In *Intelligent Tutoring Systems* (pp. 353–362). Berlin: Springer.

Clark, D. B., Nelson, B., Chang, H., D'Angelo, C. M., Slack, K., & Martinez-Garza, M. (2011). Exploring Newtonian mechanics in a conceptually-integrated digital game: Comparison of learning and affective outcomes for students in Taiwan and the United States. *Computers and Education, 57*(3), 2178–2195.

Common Core State Standards Initiative. (2011). *Common core state standards for mathematics*. Retrieved May, 2011, from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf.

Dede, C. (2012, May). Interweaving assessments into immersive authentic simulations: Design strategies for diagnostic and instructional insights. In *Invitational Research Symposium on Technology Enhanced Assessments.* Retrieved from http://www.k12center.org/rsc/pdf/session4-dede-paper-tea2012.pdf.

Dorval, M., & Pepin, M. (1986). Effect of playing a video game on a measure of spatial visualization. *Perceptual and Motor Skills, 62*, 159–162.

Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science, 18*, 850–855.

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature, 423*(6939), 534–537.

Green, C. S., & Bavlier, D. (2007). Action-video-game experience alters the spatial resolution of vision. *Psychological Science, 18*(1), 88–94.

Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective- taking spatial abilities. *Intelligence, 32*, 175–191.

Hwang, G., Sung, H., Hung, C., Huang, I., & Tsai, C. (2012). Development of a personalized educational computer game based on students' learning styles. *Educational Technology Research and Development, 60*(4, Special Issue on Personalized Learning), 623–638.

Ke, F., Shute, V., Erlebacher, G., Clark, K., & Ventura, M. (2014, June 9–10). *Earthquake rebuild: Math learning through modeling and design.* Poster presented at Cyberlearning Summit 2014, Madison, WI.

Kickmeier-Rust, M. D., & Albert, D. (2010). Micro-adaptivity: Protecting immersion in didactically adaptive digital educational games. *Journal of Computer Assisted Learning, 26*(2), 95–105.

Leemkuil, H., & de Jong, T. (2012). Adaptive advice in learning with a computer-based knowledge management simulation game. *Academy of Management Learning & Education, 11*(4), 653–665.

Levy, R. (2014). *Dynamic Bayesian network modeling of game based diagnostic assessments* (CRESST Report 837). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

MacCallum-Stewart, E. (2011). Stealth learning in online games. In S. de Freitas & P. Maharg (Eds.), *Digital games and learning* (pp. 107–128). London: Continuum.

Miller, H. J., & Han, J. (Eds.). (2009). *Geographic data mining and knowledge discovery*. London: CRC Press.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., & Lukas, J. F. (2006). Concepts, terminology, and basic models of evidence-centered design. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 15–47). Hillsdale, MI: Lawrence Elbaum.

Montello, D. R. (1993). Scale and multiple psychologies of space. In A. U. Frank & I. Campari (Eds.), *Spatial information theory: A theoretical basis for GIS* (Proceedings of COSIT'93. Lecture notes in Computer Science, Vol. 716, pp. 312–321). Berlin: Springer-Verlag.

Montello, D. R., & Golledge, R. G. (1999). Scale and detail in the cognition of geographic information. Report of the Specialist Meeting of Project Varenius, Santa Barbara, CA, May 14–16, 1998. Santa Barbara: University of California.

O'Rourke, E., Haimovitz, K., Ballweber, C., Dweck, C., & Popović, Z. (2014, April). Brain points: A growth mindset incentive structure boosts persistence in an educational game. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (pp. 3339–3348). ACM.

Prensky, M. (2001). *Digital game-based learning*. New York: McGraw-Hill.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. (Eds.). (2011). *Handbook of educational data mining*. London: CRC Press.

Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., et al. (2009). Epistemic network analysis: A prototype for 21st-century assessment of learning. *International Journal of Learning and Media, 1*(2), 33–53.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–524). Charlotte, NC: Information Age Publishers.

Shute, V. J., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. Cambridge, MA: MIT Press.

Shute, V. J., Masduki, I., & Donmez, O. (2010). Conceptual framework for modeling, assessing, and supporting competencies within game environments. *Technology, Instruction, Cognition, and Learning, 8*(2), 137–161.

Shute, V. J., Ventura, M., & Ke, F. (2015). The power of play: The effects of Portal 2 and Lumosity on cognitive and noncognitive skills. *Computers & Education, 80*, 58–67.

Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in Newton's playground. *Journal of Educational Research, 106*(6), 423–430.

Siemens, G., & de Baker, R. S. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252–254). ACM.

SoLAR, (February 27–March 1, 2011), 1st International Conference on Learning Analytics and Knowledge, Banff, Alberta; as cited in George Siemens and Phil Long, "Penetrating the Fog: Analytics in Learning and Education," *EDUCAUSE Review,* vol. 46, no. 5 (September/October 2011).

Spence, I., Yu, J. J. J., Feng, J., & Marshman, J. (2009). Women match men when learning a spatial skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1097–1103.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., & Warren, C., et al. (2012). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin.* Advance online publication. doi:10.1037/a0028446

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotation, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills, 69*, 915–921.

Ventura, M., Shute, V. J., Wright, T., & Zhao, W. (2013). An investigation of the validity of the virtual spatial navigation assessment. *Frontiers in Psychology, 4*, 1–7.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. San Francisco, CA: Morgan Kaufmann.

Wouters, P., & Van Oostendorp, H. (2013). A meta-analytic review of the role of instructional support in game-based learning. *Computers & Education, 60*(1), 412–425.

Yairi, T., Kawahara, Y., Fujimaki, R., Sato, Y., & Machida, K. (2006, July). Telemetry-mining: a machine learning approach to anomaly detection and fault diagnosis for space systems. In *Space Mission Challenges for Information Technology, 2006. SMC-IT 2006. Second IEEE International Conference on* (8-pp). Pasadena, CA: IEEE.

Zapata-Rivera, D., VanWinkle, W., Doyle, B., Buteux, A., & Bauer, M. (2009). Combining learning and assessment in assessment-based gaming environments: A case study from a New York City school. *Interactive Technology and Smart Education, 6*(3), 173–188.

Chapter 14
An Application of Exploratory Data Analysis in the Development of Game-Based Assessments

Kristen E. DiCerbo, Maria Bertling, Shonté Stephenson, Yue Jia, Robert J. Mislevy, Malcolm Bauer, and G. Tanner Jackson

Abstract While the richness of data from games holds promise for making inferences about players' knowledge, skills, and attributes (KSAs), standard methods for scoring and analysis do not exist. A key to serious game analytics that measure player KSAs is the identification of player actions that can serve as evidence in scoring models. While game-based assessments may be designed with hypotheses about this evidence, the open nature of game play requires exploration of records of player actions to understand the data obtained and to generate new hypotheses. This chapter demonstrates the use of the 4R's of Exploratory Data Analysis (EDA): revelation, resistance, re-expression, and residuals to gain close familiarity with data, avoid being fooled, and uncover unexpected patterns. The interactive and iterative nature of EDA allows for the generation of hypotheses about the processes that generated the

K.E. DiCerbo (✉)
Pearson, 400 Center Ridge Dr, Austin, TX 78753, USA
e-mail: kristen.dicerbo@pearson.com

M. Bertling • Y. Jia
Educational Testing Service, MS-T-03, 660 Rosedale Rd, Princeton, NJ 08541, USA
e-mail: mbertling@ets.org; yjia@ets.org

S. Stephenson
GlassLab Games, 209 Redwood Shores Pkwy, Redwood City, CA 94065, USA
e-mail: shonte.berkeley@gmail.com

R.J. Mislevy
Educational Testing Service, Center for Advanced Psychometrics,
MS 12-T, 660 Rosedale Rd, Princeton, NJ 08541, USA
e-mail: rmislevy@ets.org

M. Bauer
Educational Testing Service, 16-R, Turnbull Hall, Rosedale Rd, Princeton, NJ 08540, USA
e-mail: mbauer@ets.org

G.T. Jackson
Educational Testing Service, 660 Rosedale Rd, MS 16-R, Princeton, NJ 08541, USA
e-mail: gtjackson@ets.org

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based
Learning, DOI 10.1007/978-3-319-05834-4_14

observed data. Through this framework, possible evidence pieces emerge and the chapter concludes with an explanation of how these can be combined in a measurement model using Bayesian Networks.

Keywords Exploratory data analysis • Game-based assessment • Evidence model • Data visualization • Re-expression • Residuals

1 Introduction

The past decade has seen a growing push for games in learning spaces (Gee, 2003). A new generation of promising educational games has emerged allowing for deep exploration of broad concepts (Klopfer, Osterweil, & Salen, 2009). Games support sociocultural and situative approaches to learning in which players interact with peers and their environment to develop knowledge and understanding of the world (Steinkuehler, 2004). In addition, data from games provide information about the process a player used to arrive at a final product, suggesting great potential for generating new insights regarding student actions as they relate to complex knowledge, skills, and attributes (Mislevy, Behrens, DiCerbo, Frezzo, & West, 2012). Game-based assessments (GBAs) have the potential to combine the rich problems, engagement, and motivation from games with the evidentiary arguments of assessment.

However, the potential of games as assessment tools can be met only if replicable methods for aligning game play with learning standards and formative assessment objectives can be developed. New interactive digital games elevate both the availability of student micro-patterns (small, repeatable segments of play actions) and the importance of understanding them as they reflect variation in strategy or evolving psychological states. While the richness of the data holds promise for making important inferences, standard methods for scoring and analysis do not exist. In addition, the open nature of many games means students often engage in unexpected actions in the game. This requires multiple cycles of data exploration, hypothesis generation, and confirmation on the part of the analyst to fully understand the relationships of game play actions to inferences about players.

Assessment is fundamentally about designing situations which elicit evidence about aspects of what learners know and can do. Evidence-Centered Design (ECD; Mislevy, Steinberg, & Almond, 2002) provides a framework for specifying these arguments. It defines the following models:

- Student model—What we want to know about the learner
- Task model—What activities the learner will undertake
- Evidence model—How we link the work produced in the task to the constructs in the student model. The evidence model contains two pieces:
- Scoring model—How we will identify evidence in the learners' work product
- Measurement model—The statistical techniques we use to link the evidence to the elements in the student model

This chapter will focus largely on the scoring model, or the identification of the important elements in the record of player actions to extract and pass to our measurement models. For multiple choice items, the scoring model is simple. The work product is a list of selected options. The scoring rule for each item is, “if selection matches correct response, then mark correct, otherwise mark incorrect.” However, when the work product is a log file of actions a student has taken in a game, it is less clear how to identify the scoring rules, much less apply them. What are the actions in the game that will tell us about the knowledge, skills, and attributes of interest? Our usual assessment routines and psychometric processes cannot be easily lifted from our traditional assessments and applied to GBAs.

In designing GBAs, the specification of the scoring model is an iterative process. Design begins with hypotheses about what player actions will be important for making inferences. However, most games are complex systems. Before diving directly into confirming these hypotheses, it is important we understand the data obtained from the game and also seek to uncover unexpected patterns in the data that may generate new hypotheses. Exploratory Data Analysis (EDA; Tukey, 1977) provides a helpful framework by which to consider the processes of hypothesis generation and exposition of patterns in data. While EDA techniques are not new, the application of these older (but often overlooked) methods in this new context provides a way to facilitate new ways of identifying evidence for inferences about player knowledge, skills, and attributes. This chapter will focus on the use of EDA to gain close familiarity with game-based assessment data, avoid being fooled, and uncover unexpected patterns while developing an understanding of what features of player game play provide evidence about our constructs of interest. The final section of the chapter will demonstrate how these uncovered evidence fragments can then be inserted into a measurement model to estimate proficiency of game players. The scoring model and measurement model in combination allow the translation of game play into inferences about knowledge, skills, and attributes. The chapter will use analysis of data from SimCityEDU to demonstrate the concepts of the EDA framework.

1.1 Exploratory Data Analysis

EDA is a conceptual framework with a core set of ideas and values aimed at providing insight into data, and to encourage understanding probabilistic and nonprobabilistic models in a way that guards against erroneous conclusions (Behrens, DiCerbo, Yel, & Levy, 2012). EDA also provides a set of tools that allow researchers to become intimately familiar with their data. It encourages the development of mental models of the data and processes that created them.

EDA holds several complementary goals: to find the unexpected, avoid being fooled, and develop rich descriptions. The primary analogy used by Tukey (1977) to communicate these goals is that of the data analyst as detective. The work is essentially exploratory and interactive, involving an iterative process of generating hypotheses and looking for fit between facts and the tentative theory or theories. Detective work also provides a solid analogy for EDA because both are essentially bottom-up processes of hypothesis formulation and data collection.

Tukey (e.g., 1986) did not consider methodology as a bifurcation between exploratory and confirmatory, but considered quantitative methods to be applied in stages of exploratory, rough confirmatory, and confirmatory data analyses. In this view, EDA is aimed at the initial goals of hypothesis generation and pattern detection following the detective analogy. It is therefore differentiated from the (correctly) maligned practice of snooping through data to find the data and model that will most likely lead to significant results. Rather, EDA generates hypotheses that are later confirmed with separate data. Rough confirmatory data analysis is sometimes equated with null-hypothesis significance testing that is often what is taught in statistics courses. Strict confirmatory analyses involve the more sophisticated testing of specific relationships and contrasts that is less common in research practice. As a researcher moves through these stages, she moves from hypothesis generation to hypothesis testing and from pattern identification to pattern confirmation.

In the context of EDA, the data analyst performs an iterative series of interactions with the data, all the while generating various observations and hypotheses about the forms of the data and the likely underlying processes that generated them. Therefore, to return to the original problem, EDA allows us to iteratively generate hypotheses about the patterns in the game data and their relationships to levels of knowledge, skills, and attributes. EDA provides a set of tools by which to accomplish this. We can think of them in relation to four R's (Hoaglin, Mosteller, & Tukey, 1983): revelation, re-expression, resistance, and residuals. Revelation refers to uncovering the unexpected, largely through visualization. Re-expression involves careful understanding of the distributions of variables. Resistance implies using methods that are not overly influenced by extreme or unusual data. Finally, residuals provide a means by which to evaluate and iterate with models. Each of these will be discussed further with examples in the remainder of the chapter.

1.2 Context

For illustrative purposes, references will be made throughout the chapter to SimCityEDU (www.simcityedu.org), developed by GlassLab. SimCityEDU, based on the popular SimCity commercial game, offers players various challenges that ask players to solve problems facing a city, generally requiring them to balance elements of environmental impact, infrastructure needs, and employment. The game scenarios are designed to assess systems thinking. Often named on lists of twenty-first century skills, systems thinking is also a cross-cutting concept in the Next Generation Science Standards (NGSS; NGSS Lead States, 2013). Essentially, it is the understanding of how various components of a system influence each other.

Table 14.1 Systems thinking learning progression from SimCityEDU

Level 1—Acausal
The player is not reasoning systematically about causes and effects
Level 2—Univariate
The player tends to focus on a single causal relationship in the system
Level 3a—Early multivariate
The player has considered multiple effects resulting from a single cause
Level 3b—Multivariate
The player has considered multiple causes in relation to their multiple effects
Level 4—Emergent patterns
The player attends to and intervenes on emergent patterns of causality that arise over time

Starting with a strong research-based theory or cognitive model is preferable (but not required) in the development of GBAs because it can provide clear hypotheses to design, categorize, and evaluate evidence that can be further explored through EDA. The aim is to jumpstart the design of GBAs using an initial psychological theory of students' likely changes in competency toward the learning goals during game play. This approach leverages existing models of learning and how peoples' understanding of concepts potentially progress through qualitative changes in a particular developmental sequence (e.g., learning progressions of how their thinking develops from simpler more univariate concepts to more complex interactive systems; c.f., Heritage, 2008). These learning progressions, or cognitive models, help to inform design and development of GBAs, but the models themselves are also subject to iterative refinement as data are collected during playtesting, mini-tryouts, pilots, and larger-scale studies. Following a review of existing conceptualizations of systems thinking, a learning progression for the construct was developed as part of the student model for the game. Table 14.1 presents a summary of the systems thinking learning progression used in SimCityEDU.

The examples described in this chapter relate to efforts to uncover evidence in players' game actions related to systems thinking. While SimCityEDU consists of four scenarios, discussion here focuses on the third, which requires players to balance maintaining enough power in the city with reducing air pollution. Players explore the city and find that coal plants are primary producers of pollution, while other industrial areas also contribute to the problem. Players can reduce pollution by bulldozing coal plants, but that will reduce power in the city. They can dezone industrial areas, but that alone will not result in large enough changes to please the city inhabitants (and get the player to a full three-star solution).

The process of analyzing the various actions players take in the game relies on the telemetry system of the game, or the remote collection of player actions and game states. Log files of telemetry data are collected for every game session and detail actions the player has taken in chronological order. The following sections seek to identify elements of game play that may provide insight into players' systems thinking using the principles of revelation, resistance, re-expression, and residuals with SimCityEDU data from 751 US middle school players who participated in beta testing of the game.

2 Revelation

Revelation refers to Tukey's (1977) statement that "The greatest value of a picture is when it forces us to notice what we never expected to see" (p. vi). Graphics are the primary tool for the exploratory data analyst. Graphical representations can display large amounts of information using relatively little space and expose relationships among pieces of information better than other representations. Here we are talking not about visualization for public display, but for finding patterns in relationships. Tools for this include things like boxplots and scatterplot matrices (a grid of scatterplots similar to a correlation matrix except with graphs) in addition to interactive graphics that allow the analyst to explore relationships with a few clicks. For example, a scatterplot may reveal a cluster of outliers. Interactive graphics allow the analyst to highlight them on the screen and examine their values on other variables to further understand what differentiates this group.

The initial goal of data analysis should be to become very familiar with the data. Instead of beginning an analysis by producing tables of descriptive statistics, followed by a big correlation matrix, EDA suggests beginning by looking at histograms, followed by scatterplots, scatterplot matrices, and boxplots. Let's take an example from the third scenario of SimCityEDU. The successful player will find out that coal power plants are the biggest pollution generators as well as the major energy producers and, therefore, both important and destructive for the city. Further, students engaged with this scenario need to discover that there are other energy sources available for them, such as solar or wind plants that are environmentally friendly. They have to figure out how replacing of coal power plants with green energy sources will allow them to reduce pollution while maintaining power in the city. We began with a rough hypothesis that just bulldozing coal plants without placing green energy would indicate a lower level of systems thinking because it indicated players were only considering a single effect of coal plants (namely pollution) rather than the multiple effects (power and pollution).

One of the first types of analyses is simply to examine the different actions and outcomes of game play. A common next step is to run the means and standard deviations, resulting in a table like the one in Table 14.2. Pollution is the final amount of

Table 14.2 Means and standard deviations of select outcomes and actions from SimCityEDU

Outcomes and actions	Mean	SD
Pollution	15,956,941	18,258,294
Bulldoze coal	3.17	1.952
Place new coal	0.20	0.745
Turned off coal	0.41	0.970
Turned on coal	0.17	0.678
Place wind/solar	2.58	2.366
Bulldoze wind/solar	0.26	1.008
Turned on wind/solar	0.07	0.370
Turned off wind/solar	0.09	0.430

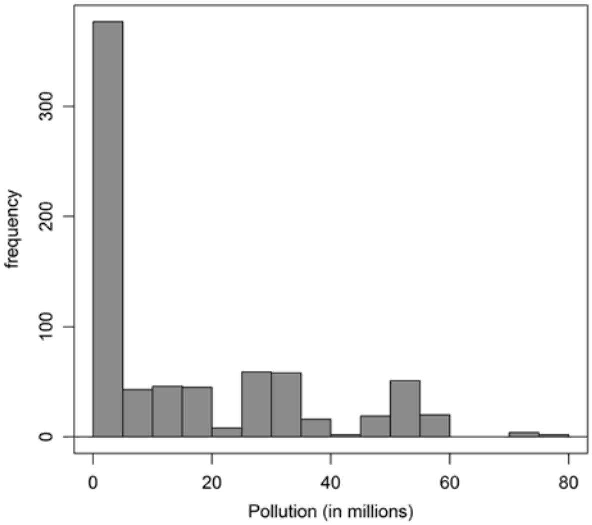


Fig. 14.1 Histogram of final pollution values

pollution in the city. Bulldozing refers to how players can eliminate buildings in the city (they use a bulldozing tool to knock them down). Placing refers to putting a new building in the city. Turning off and on are options to allow the energy plants to be active or not. Coal refers to coal plants while solar/wind refer to the alternative energy power plants available. So, on average, players knocked down 3.17 coal plants during their play, for example.

This representation does not tell us about the distribution or the outliers. However, a histogram like that in Fig. 14.1 for pollution does a better job showing these. If researchers start with visualizations first, they will better be able to interpret what numbers like those in Table 14.2 are indicating (or not indicating).

Here we see that pollution is quite skewed towards low values and actually appears to be trimodal. These three apparent groups in the outcome variable were not initially expected. The game was designed such that lower levels of pollution should be indicative higher levels of systems thinking, as players need to understand the system in order to successfully lower pollution without driving the city into a power failure. The identification of three groupings of pollution scores, however, was not intentional and raises questions about how game actions relate to these outcomes, and to systems thinking. While the groupings do not mean that the intended relationship of lower pollution to higher levels of systems thinking do not hold, it does mean that we must determine whether these groupings are artifacts of game design or whether they map to the levels of systems thinking. The latter would be a beneficial, but unexpected, result.

In Fig. 14.2, we can see the distributions of some of the other game actions. Note that bulldozing 4–6 coal plants is common. There are only six possible coal plant/generators in the original city, so anyone who bulldozed more than that must have placed new ones down. Understanding both the skew of the distributions and the location of outliers will lead into the re-expression and resistance work to follow.

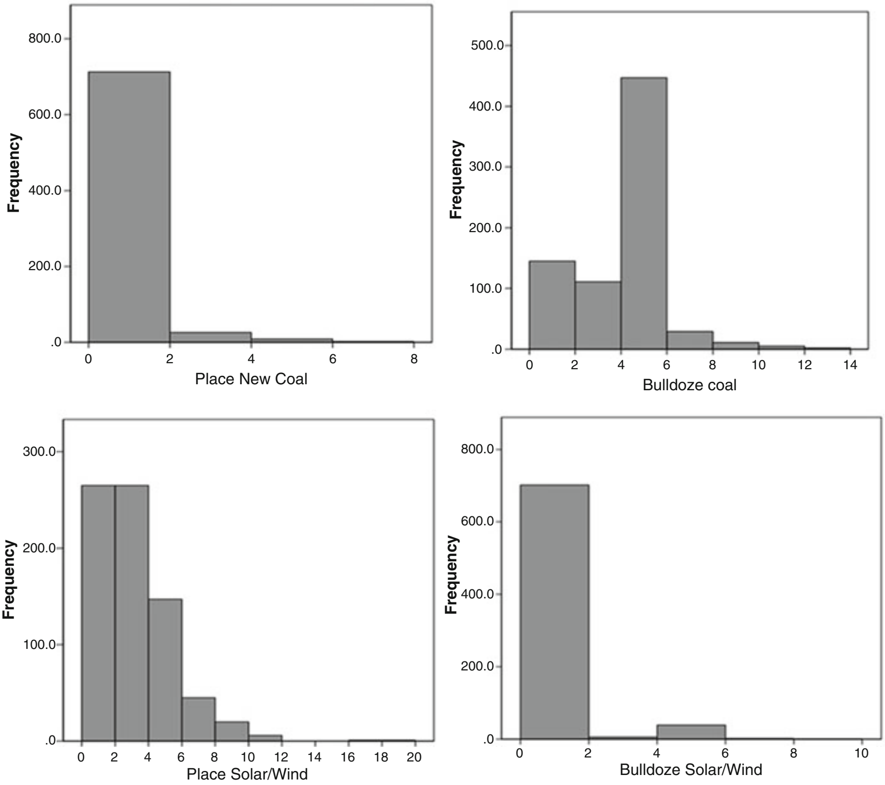


Fig. 14.2 Histograms of placing and bulldozing energy sources

Table 14.3 Correlation matrix among coal events, alternative energy events, and end state pollution

	Pollution	Bulldoze coal	Place new coal	Turned off coal	Turned on coal
Pollution	1.00				
Bulldoze coal	−.54	1.00			
Place new coal	.07	.35	1.00		
Turned off coal	−.03	−.31	−.04	1.00	
Turned on coal	.06	−.19	−.01	.82	1.00

Once we looked at this univariate information, we started looking at relationships between variables. A common technique to examine bivariate relationships is the creation of a correlation matrix like that in Table 14.3.

This suggests a moderate negative correlation between pollution and bulldozing coal, but not with other variables related to coal levels. However, the numbers themselves do not provide information about the patterns of relationship (for example, linearity and nonlinearity). To see those, scatterplot matrices such as the ones in Fig. 14.3 are helpful.

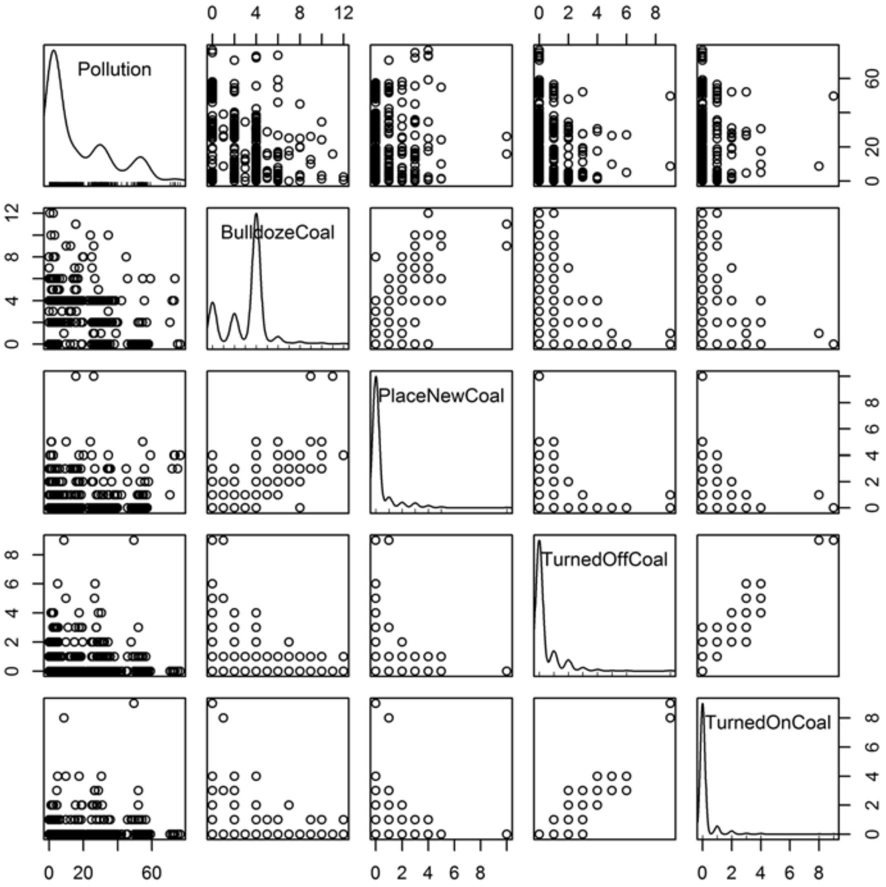


Fig. 14.3 Scatterplot matrix of relationship between coal activities and pollution

Figure 14.3 shows all of the actions that can increase or decrease the amount of coal production in the city. This matrix works like a correlation matrix such that each square is the scatterplot of the row and column variable with distributions of each variable on the diagonal. So the second box on the top row shows us the relationship between bulldozing coal and pollution. One thing that is apparent looking at this box is that there are some players that do not bulldoze any coal plants, but still end up with low pollution. These will require more investigation. Looking at the far right column, the fourth box down shows the relationship between turning coal plants off and on. When a player enters the game, all of the coal plants are on. This graph suggests that many of the players that turn a plant off proceed to turn it back on again. This behavior coincides with observations made during play testing that the turning off behavior is often a "testing" behavior in which the player can test the effect of turning a coal plant off without the permanency of bulldozing it. However, it is clear that this action is often reversed by turning it back on. Therefore, when we are looking

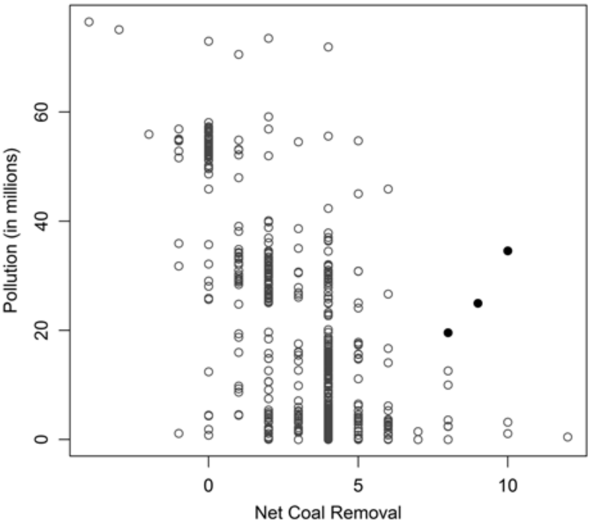


Fig. 14.4 First attempt to visually analyze relationship between net coal removal and pollution

to determine the total amount of coal removal, what we actually need is a measure of net removal that takes into account the reversal of removal actions.

We therefore created a variable adding the number of bulldozing coal and turning off coal actions and then subtracting the placing new coal and turning on coal actions to get a measure of net coal removal. The first time we created this variable and plotted the net coal removal against pollution, the result was that shown in Fig. 14.4.

In analyzing this, we were drawn to the three filled in black data points in the lower right of the figure. These were apparently individuals who had high net coal removal but continued to have relatively high pollution values. In order to search for other explanations for their pollution values, we returned to their log files. Rather than finding some other variable to explain the high pollution, we found that they had in fact placed additional coal plants that had not been properly coded in the automated scripts that clean the data. Here our visualizations helped us identify an error in our own data cleaning processes, and avoiding being fooled by incorrect data. Going back and fixing the coding of these values yielded the graph in Fig. 14.5.

3 Resistance

Because a primary goal of EDA is to avoid being fooled, resistance is an important aspect of using EDA tools. Resistant methods are methods that are less sensitive to large disruptions in small parts of the data (Mallows, 1983). Thus, they help us reduce the effects of extreme or unusual data. Note that this is different than robustness in that robustness deals with the ability of a statistic to give adequate estimates when assumptions are violated. Resistant methods are those that generally do not have these assumptions. In general, there are three primary strategies for improving

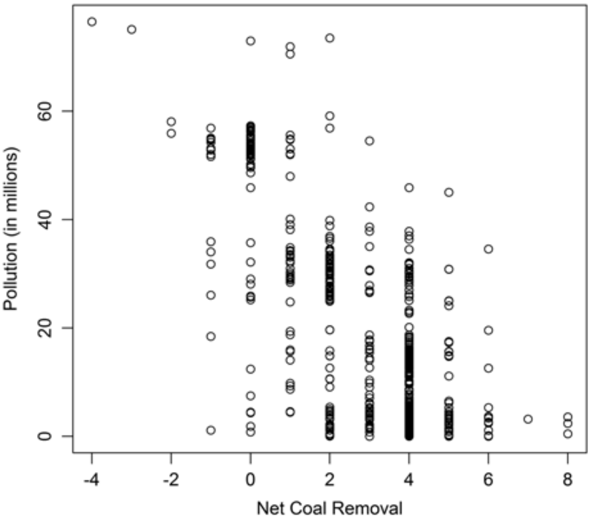


Fig. 14.5 Corrected scatterplot of relationship between net coal removal and pollution

resistance. The first is to use rank-based measures (e.g., the median) and absolute values, rather than measures based on sums (e.g., the mean) or sums-of-squares (such as the variance). While the mean has a smaller standard error than the median, and so may be an appropriate estimator for many confirmatory tests, the median is less affected by extreme scores or other types of perturbations that may be unexpected or unknown in the exploratory stages of research. For measures of spread, the interquartile range is the most common resistant method. The second general resistance building strategy is to use a procedure that emphasizes more centrally located scores, and uses less weight for more extreme values. This category includes trimmed statistics in which values past a certain point are weighted to zero, and thereby dropped from any estimation procedures. A third approach is to reduce the scope of the data one chooses to model on the basis of knowledge about extreme scores and the processes they represent. Depending on the application and the intended use of results, different methods will be appropriate in different situations.

3.1 Dealing with Outliers

Because an important goal of EDA is to develop understandings and descriptions of data, it is important to recognize that the data arise in specific contexts and contain background assumptions, even when these assumptions are unrecognized. This context and background can help us determine how to deal with outliers. Do we keep them or pull them out?

The fundamental question to ask is: Do we know something about these observations that suggests they come from a different process than the process we are seeking to understand? In games, numerous unintended processes could lead to outlying

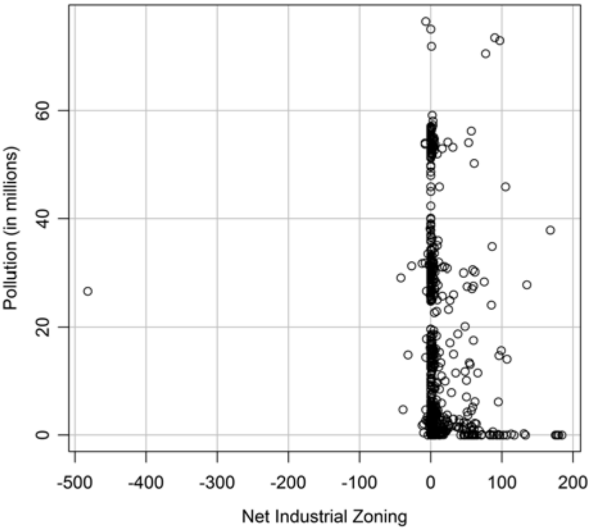


Fig. 14.6 Scatterplot of pollution and net industrial zoning

values: failure to understand instructions, exploring the environment, following their own goals, failure to pay attention to the task, or equipment or data failures. Games often encourage exploration and player agency, which means that players can often be observed doing things unrelated to the processes we wish to observe.

As an example, when we create a scatterplot of the net industrial zoning (another factor that should reduce pollution) versus pollution, we get the plot in Fig. 14.6. There is clearly one outlier who removed more than 500 industrial zones from the city. Further examination of this individual's log file revealed this individual also bulldozed 349 residential structures (median for the sample = 4) and 64 commercial structures (median = 3) while also dezoning 556 residential areas (median = 3) and 171 commercial areas (median = 0). This is an individual who appears to be seeking to destroy or eliminate most of the pre-built city. This is clearly a different goal than that intended and means we really cannot make any inferences about this individual's level of systems thinking. As a result, this is a case where it is justifiable to remove an outlier.

Alternately, when we look back at the scatterplot in Fig. 14.5, we could call the two values in the upper left outliers. They have higher pollution than any other players and lower net coal removal. However, these players' frequencies on other bulldozing and zoning variables are consistent with other players. There is no evidence that these players are not attempting to reduce pollution, they just are not doing it very well. Therefore, they were left in the sample, but the inclusion of their more extreme values point to the need for reporting of medians and interquartile ranges when reporting descriptive statistics. The most important aspect in either case is that a careful and detailed description of the full data, the reduced data, and the impact of the outlying data be reported. Unfortunately, the extremely terse descriptions seen in a lot of research reporting is inconsistent with this highly descriptive approach.

4 Re-expression

Data often come to the exploratory data analyst in messy, nonstandard, or simply not-useful ways. This may be overlooked if one assumes the data distributions are always well behaved, or that statistical techniques are sufficiently robust that we can ignore any deviations that might arise, and therefore skip detailed examination. In fact, it is quite often the case that insufficient attention has been paid to scaling issues either in advance, or during the modeling phase, and it is not until the failure of confirmatory methods that a careful examination of scaling is undertaken. Addressing appropriate scaling in advance of modeling is called re-expression and is a fundamental activity of EDA. Recently, advances in modeling have resulted in the ability to model distributions and nonlinearity, but still require careful consideration of underlying distributions in order to specify the appropriate model. Re-expression here refers solely to attempts to address the scaling of the data, as opposed to smoothing, for example, which aims at reducing the variability of the data.

The distribution most commonly "assumed" by statistical tests is the "normal" distribution. In EDA, the term "normal distribution" is avoided in favor of "Gaussian distribution" to avoid the connotation of prototypicality or social desirability. A Gaussian shape is sought because this will generally move the data toward more equal-interval measurement through symmetry, will often stabilize variance, and can quite often yield forms of the data that lend themselves to other modeling approaches (Behrens, 1997).

4.1 Re-expression Prior to Modeling

Although mathematically equivalent to what is called transformation in other traditions, re-expression is so named to reflect the idea that the numerical changes are aimed at appropriate distributions rather than radical change. An appropriate re-expression can often be found by moving up or down the ladder of re-expression (Tukey, 1977). The ladder of re-expression is a series of exponents one may apply to original data that show considerable skew. Recognizing the raw data exists in the form of X1, moving up the ladder would consist of raising the data to X2 or X3. Moving down the ladder suggests changing the data to the scale of X1/2, −X-1/2, −X-1, −X-2, and so on. The position on the ladder occupied by X0 is generally replaced with the re-expression of log(X), where the log is usually either taken to be the base 10 logarithm or the natural logarithm; the choice between them is arbitrary but may be made for interpretation. Gelman and Hill (2007) for example, suggest that the base 10 logarithm yields easier interpretation of data while the natural logarithm yields easier interpretation of coefficients in models. Note that the Box-Cox power transformation is one more formal method by which to search for and apply the best means of re-expression.

To choose an appropriate transformation, one moves up or down the ladder (i.e., takes each data point and applies the appropriate exponent) toward the bulk

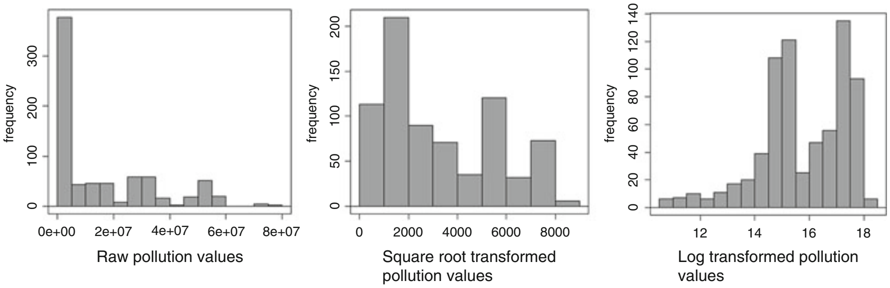


Fig. 14.7 Re-expression of end state pollution variables

of the data. This means moving down the ladder for distributions with positive skew and up the ladder for distributions with negative skew. To demonstrate this process, we can examine the distribution of the end state pollution values. These are initially highly skewed and were re-expressed with both a square root and log transformations (Fig. 14.7). The square root transformation shifted the distribution somewhat to the right but still leaves some skew (skew: 0.48). However, the log of pollution shifted the data too far, resulting in a negatively skewed distribution (skew: −2.91).

A common objection to re-expression is that the results of analyses involving re-expressed variables are difficult to interpret. This is true in some cases, however, we wish to provide an example of interpretation of log re-expression in regression to demonstrate that this should not be a barrier for some of our most common analyses. In the situation where the dependent variable is re-expressed as a log of the original variable while the independent variables are not, we say that a one unit change in the independent variable yields a 100*coefficient percent change in the dependent variable. In the case where the independent variable is re-expressed as a log but the dependent variable is unchanged, we interpret the result as a 1 % change in the independent variable results in a coefficient/100 change in the dependent variable. When both the independent and dependent variables are re-expressed as logs, we can interpret the regression result to mean that a 1 % increase in the independent variable leads to a coefficient percent increase in the dependent variable. It should be noted that re-expression alters the relative distance between data points. So, although the points all remain in the same order, there is a loss of information that may be undesirable when those distances are meant to be interpretable, such as might be the case with variables such as age or GPA (Osborne, 2002).

Although some researchers may reject the notion of re-expression as "tinkering" with the data, our experience has been that this view is primarily a result of lack of experience with the new scales. In fact, in many instances individuals use scale re-expressions with little thought. For example, the familiar practice of using a proportion is seldom questioned, nor is the more common re-expression to z-scores. Many common measurements, such as the Richter scale and decibel are transformations.

4.2 Modeling Distributions

The re-expression discussed up to this point has involved re-expression of individual variables prior to model fitting. This work is important in that it builds familiarity with the data, helps to understand different possible strategies, and suggests possible approaches for picking a computational model for an analysis. Rodgers (2010) discusses a "quiet methodological revolution" (p. 1) in which the traditional null hypothesis–testing paradigm is replaced with one of building, evaluating, and comparing models. The focus of the new paradigm is on developing models that best fit the data, rather than manipulating the data to fit the assumptions of a test of a null hypothesis and may involve re-expression to better bring out relationships.

After completing the scale-motivated methods discussed earlier, exploratory analyses often take advantage of the strengths of generalized linear models. For example, while a binary variable may be transformed to a series of logits for early data exploration, the development of a predictive model is most likely to be accomplished using a logistic regression form with all the availability of predictive values, residuals, and so forth available in common generalized linear models. In other words, data may be re-expressed for some analyses, but also left in its raw form and models incorporating the non-Gaussian distributions used. For example, count data are commonly analyzed using Poisson (log-linear) models without initial re-expression of the data. Weighted least squares can be used when variability is not constant across groups (heteroscedasticity). Gelman and Hill (2007) provide excellent examples on the application of generalized linear models following approaches largely or altogether consistent with the views expressed here. Finally, nonparametric methods can be explored, although often at the expense of power and loss of information from interval level scales.

5 Residuals

George Box (1976) succinctly summarized the importance of aligning model choice with the purpose of the analysis writing: "All models are wrong, some are useful" (p. 3). Residuals allow us to understand how our models are wrong. This emphasis on residuals leads to an emphasis on an iterative process of model building: A tentative model is tried based on a best guess (or cursory summary statistics), residuals are examined, the model is modified, and residuals are reexamined over again.

It is worth a pause here to describe how these models are built. In a traditional research view, models are developed from the hypotheses of experts with domain knowledge and the existing research base. However, in GBAs we often have very weak or nonexistent hypotheses about the relationships among variables of interest. As a result, it is prudent to examine recent advances in methods of model building. For example, researchers can submit data to Kaggle and set up a competition among data scientists to find the best models of the data, essentially crowdsourcing model building. Alternately, statistical techniques such as symbolic regression can be used to discover the relationships among variables in a model. In using any of these traditional or new techniques, a key is understanding not just the fit of the model, but where misfit is occurring.

In statistics, we use the term residual to mean what is left unexplained by the predictor(s). If you are trying to predict someone's test score by how much they studied, you are going to be wrong, for some people by a little and for some people by a lot. That amount you are off is what is "left over" of the test score after the effect of study time is accounted for. It is the residual. Different models will lead to different patterns of residuals. It is not just a case that some are big and some are small, but that when they are graphed, we can see patterns. In many models, there are assumptions about residuals. For example, in linear regression, a well-fit model will have residuals with a mean of 0 and variance should be constant. However, even without specific assumptions, examining the pattern of error terms, can yield information about how models fit the collected data.

In the EDA tradition, residual is not simply a mathematical definition, but a foundational philosophy about the nature of data analysis. The primary focus of EDA is on the development of compact descriptions of the world. However, these descriptions will never be perfect so there will always be some misfit between our model and the data, which really means a misfit between our model and the world.

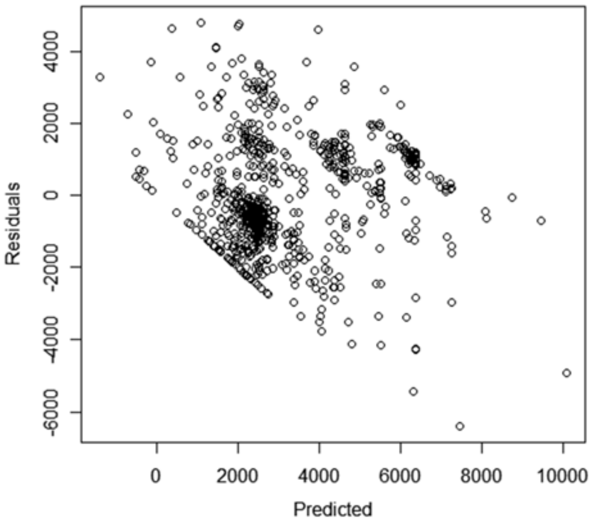
In the third scenario of SimCityEDU, we wanted to examine the factors that led to decreased pollution outcomes in the game (hypothesizing that players who ended up with lower pollution while maintaining power had a better understanding of the system). In order to test the variables explored above, we ran a linear regression model predicting the square root of pollution from the net coal removal, net industry removal, and net alternative energy placement. The model was significant, $F(3, 745) = 303.5$, $p < .001$, $\mathrm{R}^2 = .55$, Cohen's $f^2 = 1.22$. Importantly, all three predictors were significant, indicating they all contribute to pollution values above and beyond the other predictors, and engaging in those activities is likely related to understanding of the system. These three variables plus the end state are the beginnings of evidence we will include in our measurement model.

While the model is statistically significant, we should not stop there. We can graph the predicted pollution outcomes for each person versus the residuals (see Fig. 14.8).

Looking at the graph, we see that there is a clear pattern in which lower predicted values of pollution have higher residuals and higher values of pollution have smaller residuals. A biased homoscedastic pattern such as this suggests there is likely an unmodeled predictor variable.

Based on this information, we will want to adjust our model. We can do this in a number of ways. We might try to statistically model the pattern. In this case, we tried a general linear model using raw pollution values and a Poisson distribution. This yielded an even more extreme linear pattern. Our next path will be to find another predictor to add to the equation. It may be that the group that is under-predicted did something else to decrease pollution in the city. Going back to exploratory mode and/or using some other data mining techniques might uncover this.

Fig. 14.8 Scatterplot of actual pollution versus residuals



There are two cautions with this process. First, there is a point of diminishing returns where the improvements made to the model no longer have meaningful impact on the decisions to be made. For example, the slightly greater precision in estimation of ability may not be useful in informing instructional decisions. Second, going down the iterative exploratory road fits a model to a particular data set, and confirmation on independent data would be required.

6 Psychometric Techniques

To finish the discussion of evidence models in GBAs, we will briefly review how the pieces of evidence identified in EDA are combined using a measurement model in order to estimate players' levels of systems thinking. The EDA processes we saw above may yield everything from action counts to times to final scores as evidence fragments. While these individually may be interesting, we must also find a way to combine these disparate pieces of information to estimate the values of the latent traits we are ultimately interested in assessing. This is the work of the measurement model.

The simplest psychometric models are classical test theory (CTT) models or observed score models, in which scores based on observable variables are added. CTT works well when the multiple measures at issue are similar pieces of evidence about the same thing—in familiar assessments, for example, correctness across many similar test items; in GBAs, this would correspond to independent attempts at similar problems, as long as learning is negligible across those attempts. With familiar tests, CTT models also prove serviceable for collections of unlike items—as long as the collection doesn't change. Since CTT addresses the overall score, changing game scenarios or player actions changes the meaning of the scores; it does not lend itself to the rapid versioning of games or their mix-and-match character. In general, CTT does not work as well for situations that are more complicated in any of several ways: for example, where the evidence comes in different forms, has dependencies among some of its pieces, pieces depend on different mixes of skills in different combinations, proficiencies are changing across the course of observation, or different players contribute different amounts or different types of evidence. In the example above, we have information such as net coal removal events and total end state pollution. It would not make sense to simply add these values up. Latent variable models were invented to deal with assessments with these features.

Commonly used latent variable models used in educational measurement include item response theory (IRT; Yen & Fitzpatrick, 2006) and diagnostic classification models (von Davier, 2005). More detail about latent variable models can be found in Mislevy et al. (2014). Developed in traditional assessment environments, these models often have constraints on independence of observations and single dimensionality of observations that are routinely violated in GBA. While modifications of the models, such as multidimensional IRT have been developed, the multidimensional, dependent evidence with polytomous or continuous observations continue to challenge these items. Bayesian inference networks offer another option and have shown to be useful in complex assessment systems with nontraditional evidence (Almond & Mislevy, 1999; Mislevy & Gitomer, 1996; VanLehn, 2008). There is no need to pick “a” model from among them to use in GBA, because different kinds of observable variables (counts, strategy usage, features of an system diagram) can all be modeled as depending on the same latent variables by using appropriate conditional probability distributions (link functions). Furthermore, it is sometimes useful to have multiple models running in parallel, or to have them running at different levels of the hierarchical organization of GBA interactions.

We focus here on Bayesian inference networks and provide a numerical example to give some insight into how the model works in GBA. Bayesian inference networks, or Bayes nets for short, are a broad class of models for interrelationships among categorical variables. They can express or approximate the various latent-variable models mentioned above, and are particularly well suited to flexible combination of modules that express recurring relationships among kinds of evidence or between evidence and proficiencies (a characteristic that serves well in domains such as jurisprudence, intelligence analysis, and medical diagnosis; Schum, 1994). The model enables us to take advantage jointly of information from theories about a learning domain, from design strategies, and accumulating data from players. At the beginning, we posit models that reflect our initial beliefs about the targeted aspects of proficiency and the features of situations (tasks) that will evoke them. We build these hypotheses into the forms and the parameterizations of the models. By modeling conditional probabilities in terms of parameters, we can express our initial expectations as prior probability distributions for the parameters. As data arrive, Bayesian machinery allows us to get increasingly improved estimates of the model parameters and to examine where and how well the data fit the model. This information helps us fine-tune models to better manage evidence, or to modify game situations to provide better evidence. This is a particular advantage of Bayesian networks; as long as the student model variables (SMVs) remain the same, it is straightforward to incorporate additional forms of evidence, such as new evidence fragments discovered from educational data mining (EDM) or new game levels added to the game.

Koenig, Lee, Iseli, and Wainess (2010) and Shute (2011) illustrate the use of Bayes nets in GBA, with ECD as the design framework. VanLehn (2008) provides a good overview for related uses in intelligent tutoring systems. An example from the Sierra Madre challenge in SimCityEDU illustrates key ideas.

6.1 A Numerical Example

Figure 14.9 gives a numerical example of a part of a Bayes net for the third scenario. As we will see, Bayes nets generally require categorical states for the observable variables. Recall that in the above analysis the final pollution state appeared to have a trimodal distribution. Three groups were identified in the pollution result and combined with final power state to yield five levels of an End State observable variable. Similarly the net coal removal variable and net industry zoning variables identified above were combined with the net alternate energy placement variable to form a Remove Replace variable. Systems Thinking is the latent SMV and Remove Replace and End State are two observable variables, as shown in Fig. 14.9 (this is a small piece of the Bayes net for demonstration purposes). Recall that Systems Thinking has five levels. However, the design of the game targeted gathering evidence for the first four. Therefore, the top two levels are collapsed given that with the evidence

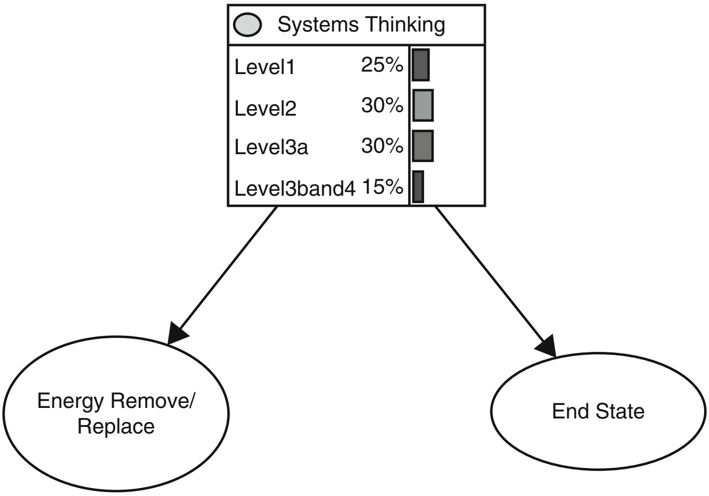
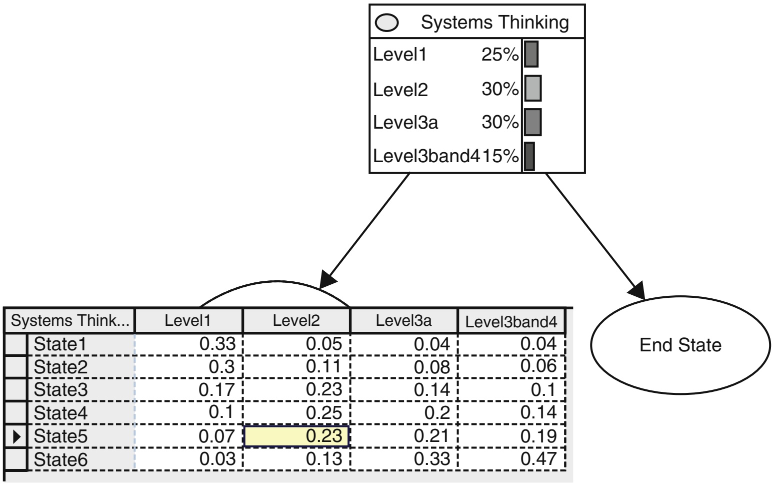


Fig. 14.9 Bayesian Network with no observed evidence



Systems Think...	Level1	Level2	Level3a	Level3band4
State1	0.33	0.05	0.04	0.04
State2	0.3	0.11	0.08	0.06
State3	0.17	0.23	0.14	0.1
State4	0.1	0.25	0.2	0.14
State5	0.07	0.23	0.21	0.19
State6	0.03	0.13	0.33	0.47

Fig. 14.10 Probability table linking evidence node to student model variable

available in the game we cannot differentiate between the two. Figure 14.9 shows the prior probabilities we assign to a student being at these levels, before observing her performance. Note that although it is traditional for diagrams to display latent variables as circles and observed variables as squares, Bayesian Network software consistently displays all variables as circles and uses squares when displaying the probability distributions, as seen in Fig. 14.9. The values shown there represent beliefs that correspond to how we expect the game to be used. That is, most players would be at level 1, 2, or 3a with respect to this context and content, and not at 3b or 4; however, without evidence, there are near equal probabilities that a player is at level 1, 2, or 3a.

We then create probability tables, like that shown in Fig. 14.10, that list the probability of observing each category or state of energy remove and replace given a level of systems thinking. So, for example, someone at level 1 of the systems thinking progression would have a .33 probability of not removing any coal or industry (State 1). The numbers for prior probability and conditional probabilities were first justified in terms of what we know about the situation—expectations based on knowing the kinds of students who would be players, research on Systems Thinking, and the numbers in the example are initial expert-opinion refined by data from a small scale try-out test. As the general release of the game brings in much large volume of data, the Bayes net allows for coherent updating of the conditional probabilities (Mislevy, Almond, Yan, & Steinberg, 1999). The model also allows for comparing the patterns in the data with the patterns the model can express, so that the model or the data-gathering situations can be improved (Levy, 2006; Williamson, Mislevy, & Almond, 2000).

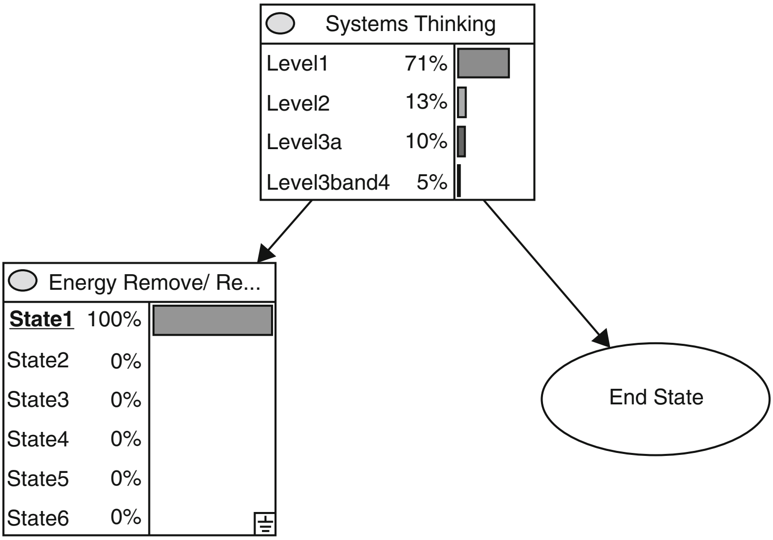


Fig. 14.11 Bayes Net after observing play

We created a similar probability table for the variable EndState. Once the probability tables are created, we can use the Bayes nets to estimate probabilities, as shown in Fig. 14.11. For example, if we see that someone has not removed any coal plants or rezoned any industrial areas, we can then update their probabilities of being at each level of the systems thinking progression. In this case, the updating results in the estimate that there is a .76 probability that the player is at level 1 (Acausal Thinking) in the progression. In this way, we are able to link in-game actions to estimates of levels of the learning progression.

7 Conclusions

The goal of the exploratory analysis of SimCityEDU was to identify potential pieces of evidence in game play related to systems thinking. The tools presented here were useful in identifying these "fragments" of evidence that could then be combined via statistical tools such as Bayesian networks. They allowed us to identify errors in our data process, suggested how actions might be used by players (e.g., turning off coal plants as a test), identify outliers and assess their inclusion in models, and judge whether our efforts to identify meaningful variables was complete. The methods of EDA are summarized in Table 14.4.

Table 14.4 Summary of 4R's of EDA (based on Hoaglin et al., 1983)

	Definition	Example
Revelation	Uncovering the unexpected, most often through visualization	Identification of a cluster of players whose actions led to unexpected game outcomes
Resistance	Using methods that are not overly influenced by extreme or unusual data	Identification of players whose actions are so different than average that they are likely pursuing a different goal in game play, suggesting we should not make inferences about their skill based on our known evidence rules
Re-expression	Ensuring match between data distributions and modeling techniques	Use of square root or log-transformed variables to better fit models
Residuals	Evaluation of where models do not fit the data, encouraging iteration	Identification of overprediction of a model at the lower end of a scale, suggesting variables are missing from the model

We believe that EDA offers a complementary approach to other analysis traditions. For example, EDM is a group of methods aimed discovering novel and useful information from large amounts of educational data. Baker and Yacef (2009) identified the following five areas of work characteristic of EDM: prediction, clustering, relationship mining, distillation of data for human judgment, and discovery with models. It is our position that EDA allows for intimacy with data prior to the use of these more complex methods. In our experience, practitioners of EDM often wait until their models fit poorly to begin investigating the issues of data familiarity and distribution discussed here. In addition, EDA serves as a theory-generating process which can inform the data mining models being built (for example, informing the list of features used in building automated detectors). We believe beginning with EDA techniques would likely result in better fitting EDM models and more thorough understanding of results. The use of residual techniques will lead to better evaluation of the resulting models as well. In the SimCityEDU project, analysis will likely move to EDM techniques in an attempt to identify other variables involved in the prediction of pollution scores.

The work of identifying evidence from GBAs ultimately requires cycles of exploration, hypothesis generation, and confirmation. While EDA is likely good practice in analysis of all kinds of data, the generally weak initial hypotheses about links between game play and evidence combined with the open nature of game play in GBAs make GBA data a prime candidate for the use of the techniques. Our psychometric techniques have progressed to the extent that we can receive the traditional correct/incorrect data, fit our established models, and review the output with known methods to examine fit. However, GBAs result in new work products and new kinds of evidence that do not easily translate into these techniques. By looking back to Tukey's Exploratory Data Analysis tools, we find a framework and powerful tools to lead us forward in analyzing our new game-based assessment data.

References

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23*, 223–237.

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*, 3–16.

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2*(2), 131–160.

Behrens, J. T., DiCerbo, K. E., Yel, N., & Levy, R. (2012). Exploratory data analysis. In I. B. Weiner, J. A. Schinka, & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology* (2nd ed., pp. 34–70). New York: Wiley.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association, 71*, 791–799.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York: Palgrave Macmillan.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/ hierarchical models*. New York: Cambridge University Press.

Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment*. Washington, DC: Council of Chief State School Officers.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis*. Hoboken, NJ: Wiley.

Klopfer, E., Osterweil, S., & Salen, K. (2009). *Moving learning games forward: Obstacles, opportunities, and openness*. Cambridge, MA: The Education Arcade.

Koenig, A. D., Lee, J. J., Iseli, M., & Wainess, R. (2010). *A conceptual framework for assessing performance in games and simulation.* (CRESST Report 771). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Levy, R. (2006). *Posterior predictive model checking for multidimensionality in item response theory and Bayesian networks*. Doctoral dissertation, University of Maryland at College Park.

Mallows, C. L. (1983). Data description. In G. E. P. Box, T. Leonard, & C.-F. Wu (Eds.), *Scientific inference, data analysis, and robustness* (pp. 135–151). New York: Academic Press.

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where the numbers come from. In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 437–446). San Francisco: Morgan Kaufmann.

Mislevy, R. J., Behrens, J. T., DiCerbo, K. E., Frezzo, D. C., & West, P. (2012). Three things game designers need to know about assessment: Evidence-centered design for game-based assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 59–84). New York: Springer.

Mislevy, R. J., Corrigan, S., Oranje, A., Dicerbo, K., John, M., Bauer, M. I., et al. (2014). *Psychometric considerations in game-based assessment*. New York: Institute of Play.

Mislevy, R. J., & Gitomer, D. H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction, 5*, 253–282.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*(4), 477–496. doi:10.1191/0265532202lt241oa.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.

Osborne, J. W. (2002). Notes on the use of data transformations. Practical Assessment, Research & Evaluation, 8 (6). Retrieved from http://pareonline.net/getvn.asp?v=8&n=6.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist, 65*, 1–12.

Schum, D. (1994). *The evidential foundations of probabilistic reasoning*. New York: Wiley.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–524). Charlotte, NC: Information Age.

Steinkuehler, C. A. (2004). Learning in massively multiplayer online games. In Y. B. Kafai, W. A. Sandoval, N. Enyedy, A. S. Nixon, & F. Herrera (Eds.), *Proceedings of the sixth international conference of the learning sciences* (pp. 521–528). Mahwah, NJ: Erlbaum.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Tukey, J. W. (1986). Data analysis, computation and mathematics. In L. V. Jones (Ed.), *The collected works of John W: Vol. IV. Philosophy and principles of data analysis: 1965–1986* (pp. 753–775). Pacific Grove, CA: Wadsworth. Original work published 1972.

VanLehn, K. (2008). Intelligent tutoring systems for continuous, embedded assessment. In C. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 113–138). Mahwah, NJ: Erlbaum.

von Davier, M. (2005). *A class of models for cognitive diagnosis*. Research Report RR-05-17. Princeton, NJ: ETS.

Williamson, D., Mislevy, R. J., & Almond, R. G. (2000). Model criticism of Bayesian networks with latent variables. In C. Boutilier & M. Goldszmidt (Eds.), *Uncertainty in artificial intelligence* (pp. 634–643). San Francisco: Morgan Kaufmann.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. Brennan (Ed.), *Educational measurement* (3rd ed., pp. 111–153). Portsmouth, NH: Praeger/Greenwood.

Chapter 15
Serious Games Analytics to Measure Implicit Science Learning

Elizabeth Rowe, Jodi Asbell-Clarke, and Ryan S. Baker

Abstract Evidence Centered Game Design (ECgD) is an increasingly popular model used for stealth game assessments employing education data mining techniques for the measurement of learning within serious (and other) games (GlassLab, Psychometric considerations in game-based assessment. Institute of Play. Retrieved July 1, 2014, from http://www.instituteofplay.org/work/projects/glasslab-research/). There is a constant tension in ECgD between how pre-defined the learning outcomes and measures need to be, and how much important, but unanticipated, learning can be detected in gameplay. The EdGE research team is employing an emergent approach to developing a game-based assessment mechanic that starts empirically from what the players do in a well-crafted game and detects patterns that may indicate implicit understanding of salient phenomena. Implicit knowledge is foundational to explicit knowledge (Polanyi, The tacit dimension. University of Chicago Press, Chicago, IL,1966), yet is largely ignored in education because of the difficulty measuring knowledge that a learner has not yet formalized. This chapter describes our approach to measuring implicit science learning in the game, *Impulse*, designed to foster an implicit understanding of Newtonian mechanics using a combination of video analysis, game log analyses, and comparisons with pre-post assessment results. This research demonstrates that it is possible to reliably detect strategies that demonstrate an implicit understanding of fundamental physics using data mining techniques on user-generated data.

Keywords Implicit learning • Science learning • Assessment • Educational data mining

E. Rowe (✉) • J. Asbell-Clarke
EdGE at TERC, 2067 Massachusetts Avenue, Cambridge, MA 02140, USA
e-mail: elizabeth_rowe@terc.edu; jodi_asbell-clarke@terc.edu

R.S. Baker
Teachers College, Columbia University, 525 W. 120th St., Box 118, New York, NY 10027, USA
e-mail: baker2@exchange.tc.columbia.edu

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based Learning, DOI 10.1007/978-3-319-05834-4_15

1 Introduction

Games have long been recognized as natural assessments (Gee, 2003, 2007). However, it was the call for games as *stealth assessments* (Shute, Ventura, Bauer, Zapata-Rivera, 2009) that encouraged game-based learning researchers to think more about switching from using formal pre-post assessments to using assessments embedded within and/or consisting solely of gameplay data. In this move to stealth assessments, most instantiations use an Evidence-Centered Game Design (ECgD) model (GlassLab, 2014; Halverson, Wills, & Owen, 2012; Plass et al., 2013; Shute et al., 2009) where explicit learning outcomes and measures are designed and developed as part of the game design process. The EdGE research team builds upon the ECgD framing with an emergent approach to detect implicit learning from complex patterns within data generated from a game whose mechanics are grounded in science. Grounded in videos of learners playing the game, EdGE studies where students' strategic game behavior is consistent with an implicit understanding of the science content and validates the use of those strategies against an external measure of implicit science learning (Asbell-Clarke, Rowe, & Sylvan, 2013; Asbell-Clarke & Rowe, 2014). Implicit science learning is expressed in brief instances of play, but unfolds and changes over course of play. This chapter outlines the theoretical lenses with which we view game-based science learning and describes the methods we use to measure that learning.

2 Implicit Science Learning in Games

Implicit knowledge (also called tacit knowledge) has a variety of forms or definitions. Polanyi (1966), a philosopher and scientist, argued that tacit knowledge is foundational to all explicit knowledge. Within tacit knowledge, Collins 2010) distinguishes between somatic tacit knowledge of primal tasks such as walking and talking; collective tacit knowledge in a community such as language and humor; and tacit relational knowledge, the tacit knowledge that with effort can become related to explicit, or formalized, knowledge. Tacit relational knowledge is likely of most direct consequence to formal education.

The ways in which implicit knowledge can impact learning and teaching is not completely new to education. Vygotsky (1978) described *preparedness for learning* as the abilities and understandings a learner brings to a learning situation that can be scaffolded by a teacher, environment, and tools. Late in the last century, much literature in US science education turned attention to implicit learning in the form of misconceptions that may get in the way of a learner's conceptual development (e.g., McCloskey, 1983; Minstrell, 1982). diSessa (1993) notes the robustness of physics misconceptions with over half of respondents agreeing with several common misconceptions about basic physics, such as Newton's Laws of Motion. diSessa also distinguishes between the intuitive knowledge that novices hold—that a book will not fall through a table or that a glowing filament is hot—from an expert understanding of these phenomena. For novices, these understandings guide behavior, but are not necessarily expressible in formalisms or questioned in a deeper sense. Experts, however, not only think about a phenomenon in a more nuanced sense, but also may seek consistency across phenomena to be able to abstract their experiences towards more general principles about the world (diSessa, 1993).

Implicit knowledge is, by definition, largely unexpressed by the learner making it particularly challenging to measure. Games may provide an innovative assessment solution as a growing body of research shows how games may engage learners in cognitive processes that are not necessarily perceived by learner or recognized in external learning assessments (Gee, 2013; GlassLab, 2014; NRC, 2011; Thomas & Brown, 2011).

The unique affordances that games offer for the measurement of implicit science learning include (a) the ability to engage learners by encouraging them to dwell in scientific phenomena over repeated trials towards success (with appropriate scaffolding and feedback) and (b) the wealth of information that can be recorded during game play to provide evidence of their implicit learning. These features open opportunities to reveal tacit learning previously invisible to educators.

3 Stealth Assessments

In the past decade, researchers have begun assessing learning occurring in interactive environments such as games (Fisch, Lesh, Motoki, Crespo, & Melfi, 2011; Halverson et al., 2012; Shute & Ventura, 2013). A common way researchers have assessed learning in games is through pre-/post-tests or tasks before and after a specified period of gameplay. In contrast, *stealth assessments* measure learning using tasks embedded within the gameplay itself to “support learning, maintain flow, and remove (or seriously reduce) test anxiety, while not sacrificing validity and reliability” (Shute, Masduki, Donmez, & Wang, 2010, p. 10). To satisfy validity and reliability requirements, researchers often use an Evidence-Centered Design (ECD) framework that seeks to establish a logically coherent, evidence-based argument between the domain being assessed and assessment task design and interpretation (Mislevy & Haertel, 2006).

GlassLab (2014) describes how their team applied the ECD framework to the assessment of learning in SimCityEDU, creating an Evidence-Centered Game Design (ECgD) approach that carefully defines how game and assessment design must work in concert to produce an evidentiary model for learning with an explicit framework for characterizing that evidence. Other researchers have developed stealth assessments guided by the ECgD framework using educational data mining techniques to discern evidence of learning from the vast amount of click data generated by online science games and virtual environments such as *Progenitor X* (Halverson et al., 2012), *EcoMUVE* (Baker & Clarke-Midura, 2013), *Newton's Playground* (Shute, Ventura, & Kim, 2013), and *Surge* (Clark et al., 2011).

Within ECgD, measures of learning must be considered and designed along with the game mechanics. Plass et al. (2013) argue that game mechanics, learning mechanics, and assessment mechanics must be designed in symbiosis with each other. For example, game mechanics to launch projectiles or maneuver objects through gravitational fields may have as their learning mechanics the development of specific understanding of forces and motion. The assessment mechanics in this case are the game behaviors (and often achievements) that correspond to the consistent use of strategies to grapple successfully with the forces and motion created by the gravitational effects.

However, there is a constant tension in ECgD to make sure gameplay is designed to support and measure meaningful learning, while also remaining open to important learning that may occur during gameplay but that designers may not have considered from the start. This is especially important in game spaces with hundreds or, in some cases, thousands of play patterns where the player can be successful. There can also be a tension between the most enjoyable game mechanics and the most effective learning and assessment mechanics.

EdGE seeks to remain as open as possible to emergent evidence of implicit learning in games while still pursuing the logical coherence of the ECgD framework. We do this through a more open-ended, bottom-up iterative design process that optimizes game design for learner engagement (i.e., would they choose to play this game in their free time?) and allows the assessment mechanisms to emerge from observations of gameplay rather than place any constraints on the game design. The remainder of this chapter describes the EdGE research team's attempt to push game assessment mechanic development towards that more emergent end of the spectrum while maintaining validity and reliability. We describe this process in the context of the game, *Impulse*, which has been played or downloaded by over 10,000 players online and through the iOS and Android app stores.

4 Impulse

EdGE designed *Impulse* to foster and measure implicit learning about Newton's First and Second Laws of motion by placing a simple game mechanic (get your particle to the goal without crashing into other particles). *Impulse* immerses players in an n-body simulation of gravitationally interacting particle in which they must predict the Newtonian motion of the particles to successfully avoid collisions and reach the goal (see Fig. 15.1). For a better understanding of this work, readers are encouraged to play *Impulse* at http://www.edgeatterc.com/edge/games/impulse/.

The motions of all particles in the game obey Newton's laws of motion and gravitation, including accurate gravitational interactions and elastic collisions among ambient particles with varying mass. Players use an impulse (triggered by their click or touch) to apply a force to particles. If the player's particle collides with any ambient particle, the level is over and they must start again. Each level of the game gets more complex, requiring players to grapple with the increasing gravitational forces of an increasing number of particles and also particles of different

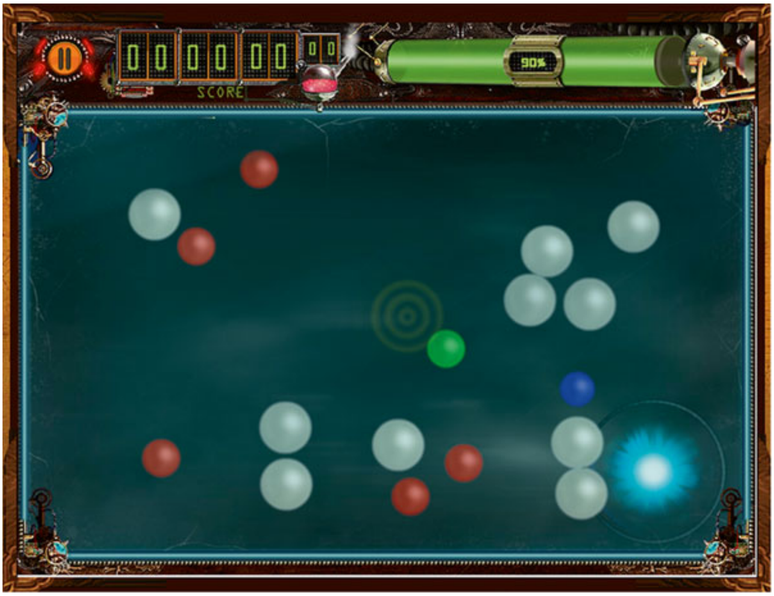


Fig. 15.1 A screenshot from *Impulse*. The player is the *green* particle and is going towards the cyan goal in the *bottom-left corner*

mass (and thus inertia). For each level, they must accomplish this goal with 20 impulses. Each impulse depleted the energy available to the player in the game (measured by the green bar in the upper right corner of Fig. 15.1). Once they exceed 20 impulses, the player no longer has energy left to apply any force to the particles.

Newton's First Law states that an object in constant motion will stay in constant motion unless acted upon by an external force. This is counterintuitive for many learners because we rarely encounter a frictionless environment in real life (McCloskey, 1983). Newton's Second Law states that the acceleration an object experiences from a force depends on the mass of the object. The *game mechanic* increases n, the number of particles, to increase the complexity and difficulty of each level and also uses particles of different mass to provide opportunities for players to grapple with phenomena governed by Newton's First and Second Laws. The *learning mechanic* is designed assuming that as players dwell in increasingly complex situations in the game, they may build tacit knowledge that is foundational for explicit learning of the behaviors governed by these laws.

The *assessment mechanic* is designed to measure players' behaviors that may indicate they are gaining an implicit understanding of Newtonian motion. We look for patterns of play in the game data logs that reflect behaviors that players demonstrate that are consistent implicit understanding. For example, players may let a ball "float" with added force, and then use an opposing force to stop the ball's motion—both consistent with an understanding of Newton's first law of motion. Even more directly, a player might consistently use more force to accelerate a heavier object than a lighter one—demonstrating an implicit understanding of Newton's second law.

5 Assessing Implicit Science Learning

The EdGE research team is taking three steps to build assessment mechanics of Newton's First and Second Law for *Impulse*. First, we coded videos in terms of specific strategic moves, noting which strategic moves are consistent with an understanding of Newton's first and second laws (i.e., the phenomena in which they are dwelling). Second, we mined the game log data for evidence consistent with an implicit understanding of those laws. Finally, we will be validating those play patterns against learner performance on a pre-post assessment of those concepts. These steps vary slightly for each of Newton's Laws. While evidence for Newton's First Law can be found in a player's single actions (clicks), evidence for Newton's Second Law relies on the relationship between sequences of actions (i.e., how many times they click on particles of different masses within a short time).

We hypothesize that advancing to higher levels in *Impulse* depends upon, fosters, and demonstrates an implicit understanding of Newton's laws. While navigating among particles that are colliding and are attracted or repelled by each other, players need to "study" the particles' behavior. They must predict the motion of the particles so that they can avoid them as they travel to the goal. Specifically, we expect players to increase their understanding that each particle will keep moving on its path without an impulse or force from another particle (Newton's First Law) and that different mass particles react differently to the same force (Newton's Second Law).

5.1 Video Coding as Ground Truth

Two researchers, one the game designer with a physics background and the other with expertise in the learning sciences and limited background in physics, began developing the coding system using video recordings from two play test sessions, one with 10 high school students from urban and suburban schools in the northeastern US, and another with six Physics graduate students from a small university in Canada. These samples represent players with novice and near-expert understandings of Newton's Laws of Motion.

Players' interactions with *Impulse* were recorded with Silverback software (Clearleft Ltd, 2013) capturing both players' onscreen game activities and video of their faces and conversations. Students were asked to "think aloud" while playing. Typically students played in groups, one student per computer, prompting conversation about gameplay and phenomena they observed. Silverback solves many synchronization problems others have experienced using multiple video cameras to record screen activity, facial expressions, and conversations.

Data from a larger number of learners were needed to build detectors based on this coding system. These data were collected over 6 hour-long workshops conducted in March–June 2013 with 69 high school students (29 female) from urban and suburban schools in the Northeastern United States. A third coder with no physics background was trained using the coding system and coded randomly selected 3-min segments from all 69 videos. Segments were randomly chosen above Level 20 whenever possible to ensure players had already mastered the game mechanic and had encountered particles of different masses. Twenty-nine of the players (42 %) did not reach level 20 and had time segments earlier in the game. Two additional coders and one of the designers of the coding system double-coded the segments from 10 videos for inter-rater reliability checking.

The final version of this coding system presented here was developed through repeated coding of hundreds of clicks with different play styles. These codes are not mutually exclusive (i.e., it is possible for one click to be both a "Float" and a "Move Toward Goal"). Each click was coded with at least one of these codes. Table 15.1 includes definitions of the codes with inter-rater (human-human) Kappas exceeding 0.70 and the implicit understanding of Newton's First Law we claim they reflect.

When coding, we distinguished between intended and actual game moves—what the player wanted to accomplish with each click versus what actually happened. Player intentions are judged based not only on their screen actions, but also audio commentary and mouse over behaviors. Often players hold their mouse over spots, ready to click if needed, providing visible clues of their intended path or strategy. While not directly visible in the clickstream data, these behaviors are observable in

Table 15.1 Video codes, definitions, and kappas for Newton's first law (NFL)

Intended strategy code label	Game-based move	Implicit understanding	Kappa
Float	The learner did not act upon the player particle for more than 1 s	Player particle will move in a straight path if no force is applied (NFL)	0.759
Move toward goal	The learner intended to apply force to direct the player particle toward the goal	Control movement of player particle by applying force	0.809
Stop/slow down	The learner intended to use opposing force on player particle in the path of the player particle to stop/slow it down	Slow particle down by using an opposing force (NFL)	0.720
Keep player path clear	The learner intended to apply force to non-player particles to keep them out of the path of the player particle	Player particle will move in a straight path if no force is applied (NFL)	0.819
Keep goal clear	The learner intended to apply a force to non-player particles to keep the goal clear by removing the non-player particle	Control movement of non-player particles by applying force	0.832
Buffer	The learner intended to apply a force between the player and other particles to avoid collision	Control movement of player and non-player particles by applying force	0.772

Source: Rowe, Baker, Asbell-Clarke, Kasman, and Hawkins (2014) video and aid interpretation. For actual moves, we coded whether or not intended and actual moves matched and, if not, which of five unanticipated outcomes occurred. These unanticipated outcomes include (1) no effect on the target particle; (2) rapid acceleration of the target particle (i.e., click was too close to the particle and made it accelerate more rapidly than expected); (3) moved the player particle closer to another particle (i.e., causing a potential collision); (4) moved the player particle away from the goal (in the absence of reason to do so); and (5) the target particle did not move as expected with no negative consequences as is the case with the other outcomes. The reliability of this code depends on the reliability of the intended codes. If they did not agree on the intended strategy, it is likely they would not agree whether the actual move was as intended or not. Therefore, it was not surprising that the coding of unanticipated outcomes (Kappa = 0.35) was much less reliable than the coding of intended moves (see Table 15.1).

Players clearing a particle from their path towards the goal may show evidence of their implicit understanding of Newton's First Law in that they are predicting that the particle will stay at constant motion in the absence of a force (and thus will collide), so they impart the force to move it away. Even more compelling evidence of an implicit understanding of Newton's First Law is when the player directly opposes straight-line motion with their impulse (Stop/Slow Down), explicitly providing the force needed to stop their particles' motion. When a player uses a Float strategy, particularly when accompanied by a mouseover trailing along with the particle, their behavior is consistent with an implicit understanding that an external force is not needed to keep the particle moving at a constant speed (Newton's First Law).

For evidence of an implicit understanding of Newton's Second Law, we coded information about the target of the click and whether or not the target of the current click was the same as the previous click (see Table 15.2). Together, these codes were used to determine if the player treated the different mass balls differently, more specifically if they consistently used more force (clicks) to move the heavier particles than the lighter ones.

There were four different colored particles besides the player with each color signifying a different mass (in order from least to most massive): blue, red, white, dark grey. The color of the target was recorded alongside the target. The blue, red, and white balls also increased in size (consistent with the same density of ball), but the grey ball was most massive and smallest in size. This was to ensure that mass was being differentiated in players' behaviors rather than size. From these codes, the number of consecutive clicks for each color target was calculated.

Table 15.2 Video codes, definitions, and Kappas used for measuring Newton's second law

Code	Definition	Kappa
Target	Type of particle (player, other, both) the learner intended to move	0.920
Same as last target	The learner intended to move the same target as the last action	0.869

Source: Rowe, Baker et al. (2014)

5.2 *Game Log Analyses*

As the learner plays *Impulse*, the game logs every game event as well as the location of every object in the game space. Recorded game events include level starts/ends, pausing and resuming the game, clicks (impulses) in the game space, collisions between particles, collisions between the particles and the walls of the game space, and collisions of the player with the goal. The game state is recorded along with the event. The final outcome of each game level is also recorded: Advance with energy remaining, Advance without energy remaining, Collision with energy remaining, Collision without energy remaining, Restart, and Quit. Players have a limited amount of energy (20 clicks) at each level of the game, so if they "Advance without energy remaining," it means they floated into the goal after they ran out of energy.

From this raw game log, we have distilled a set of 60+ features in five major categories: (1) Location/Vector Movement of Player Particle; (2) Timing and Location of Impulses; (3) Number and Location of Other Particles; (4) Overall Game Characteristics, and (5) Game Outcome. The feature distillation process explicitly selected features thought by domain experts to be semantically relevant to the strategies observed by the human coders (Sao Pedro, Baker, & Gobert, 2012). Table 15.3 gives a non-exhaustive list of examples. The distilled features were added to the original backend data. Using the synchronized timestamps, these features are then aggregated at the click level to map to the labels provided by the video coder (Sao Pedro, Baker, Gobert, Montalvo, & Nakama, 2013).

5.2.1 Building Detectors of Strategic Moves: Evidence for Newton's First Law

With the distilled data and the human-coded data, we followed a standard process for developing a model that could replicate the human judgments using the distilled log files. In other words, the goal of these analyses was to develop software that could look at the logs of student interaction with the software and come to the same judgments as a human being.

Specifically, we developed classifiers that could infer the human-coded data (1 for the presence of a specific category, 0 when it was absent), in RapidMiner 5.3. A separate classifier was developed for each human-coded construct (strategic move), six classifiers in total.

Four algorithms were tried for the first three classifiers developed:

- W-J48—a "decision tree" algorithm which makes a set of yes/no decisions based on the data to make an eventual decision with a known confidence; based on the first decision, the second decision will be different (Quinlan, 1993).
- W-JRip—a "decision rules" algorithm which makes a set of yes/no decisions based on the data to make an eventual decision with a known confidence; the order of decisions is always the same regardless of previous decisions.

Table 15.3 Distilled feature categories, examples, and rationale

Category	Distilled feature examples	Rationale
Player particle		
1	Distance between player and goal	Players use different strategic moves when close to the goal than when farther away
2	Current speed of player particle	When the player is moving faster they need to use different strategic moves than when slow
3	Distance travelled since last event	This provides an indication of how much the game state has changed
4	Change in angle between player's path and a straight-line path to goal	Strategic moves vary depending on whether or not player has a straight-line clear path to the goal
Impulses		
1	Proximity of impulse to player particle	Identifies the likely intended target (player particle or other) of the impulse
2	Time since last impulse	Very quick actions may indicate panicking or intentional increased force; very slow actions may indicate floating strategies
3	Distance from impulse to three closest other particles and their color	Identifies the likely intended target (player particle or other) of the impulse and identifies if players click more near certain color particles
Other particles		
1	Number of other particles in play space	Describes the potential complexity of the play space
2	Number of particles in path between player and goal	Describes difficulty of immediate task of getting to goal
3	Number of particles in current path of player particle	Describes immediate danger of collision
Overall game characteristics		
1	Total time spent playing this level across multiple rounds	Describes difficulty of the level
2	Total number of times playing this level	Describes players' experience with the level

Source: Asbell-Clarke, Rowe, Sylvan, and Baker (2013)

- Logistic regression—regression conducted using a logistic function in order to predict a binary variable rather than the quantitative variable predicted in linear regression.
- Step regression—regression conducted using a step function rather than a logistic function or a linear function using the standard software RapidMiner 5.3 with the Weka Extension Package. Step regression is not to be confused with stepwise regression.

These algorithms were selected based on their success in past problems where researchers attempted to classify student behavior within online learning environments for science inquiry (cf. Baker & Clarke-Midura, 2013; Baker, Ocumpaugh, Gowda, Kamarainen, & Metcalf, 2014; Sao Pedro et al., 2012, 2013), as well as in other domains. W-J48 worked best for the first three constructs, and so W-J48 was the only algorithm attempted for the remaining three. W-J48 is a decision tree algorithm with several virtues: it produces relatively interpretable models, is fast to create and use (facilitating both validation and use in a running system), and tends to be conservative (reducing the risk of over-fitting, where a model is fit to the noise in the data as well as the signal).

The models were validated in the following fashion. For each construct, the algorithm was validated using fourfold student-level cross-validation. The students were randomly distributed into four groups. The algorithm was run, training a model on data from three of the groups. Then the model was applied to the data from the students in the fourth group and tested to see how well the model functioned on this unseen group. It is important to use student-level cross-validation to avoid training and testing a model on the same student; if a student's behavior is idiosyncratic, then the model may become over-fit to that student and less able to function effectively for other students. Student-level cross-validation penalizes models that over-fit to the specific student. Within student-level cross-validation, the number of folds may lie between 2 and the number of students. This type of cross-validation is thought to be asymptotically equivalent to the Bayesian Information Criterion (Moore, 2003); while the choice of number of folds remains arbitrary, four is a common number of folds that leads to models repeatedly being built on 75 % of students and tested on the remaining 25 %.

In this study, two goodness (performance) metrics were used to determine how effective each detector was: Cohen's Kappa (Cohen, 1960) and A′ (Hanley & McNeil, 1982). Each of these metrics was applied at the level of the 3-min segments coded from the video data.

Cohen's Kappa assesses the degree to which the detector is better than chance at identifying which segments involve a specific code. For example, a Kappa of 0.865 would indicate that a detector is 86.5 % better than chance for a specific code. A Kappa of 0 indicates that the detector performs at chance, and a Kappa of 1 indicates that the detector performs perfectly.

A′ is the probability that the detector will correctly identify whether a specific code is present or absent in a specific clip, taking model confidence into account when comparing clips to each other. A′ is equivalent to W, the Wilcoxon statistic, and closely approximates the area under the Receiver-Operating Curve (Hanley & McNeil, 1982). A model with an A′ of 0.5 performs at chance, and a model with an A′ of 1.0 performs perfectly. For example, an A′ of 0.967 indicates that a detector of "keep player path clear" can distinguish a student demonstrating that strategy within a 3-min segment from a player not demonstrating that strategy, 96.7 % of the time.

These two metrics have different virtues. Cohen's Kappa assesses the quality of a model's final decisions (and is therefore a better assessment of how well the model will perform when used to drive interventions in the most common fashion, assigning interventions when confidence is over 50 %), while A′ assesses a model's confidence in its decisions (and is therefore a better assessment of how well the model will perform when used in discovery with models analyses, which typically take percent confidence into account).

Table 15.4 Kappas and A′ for each intended strategic move

Intended strategic move	Kappa	A′
Float	0.738	0.901
Move toward goal	0.757	0.907
Stop/slow down	0.512	0.779
Keep player path clear	0.865	0.967
Keep goal clear	0.772	0.943
Buffer	0.759	0.928

Source: Rowe, Baker et al. (2014)

There are no specific cut-off values for the use of these metrics in educational data mining, as acceptable performance tends to depend on the usage and the expectations in the current domain; medical tests are published and used with A′ values of 0.75–0.80 or higher; affect detectors are published as of this writing with Kappa values as low as 0.15 and A′ values as low as 0.65 (Pardos, Baker, San Pedro, & Gowda, 2013; Sabourin, Mott, & Lester, 2011). Kappa values above 0.5 and A′ above 0.8 tend to represent state-of-the-art performance in most educational domains as of this writing. Table 15.4 shows the performance of the specific models created in this chapter.

Hence, we have developed models that can judge a learner's strategic moves relevant to Newton's First Law, successfully drawing many of the same conclusions a human being can (for six codes). These models were assessed based on their ability to agree with a human rater on entirely new, unseen data and achieve comparable reliability. They met this test, achieving reliability similar to the human coders (and much better than most automated detectors of this type in the published literature).

The ability to detect these strategic moves reliably in the game data logs means we can now compare the learning of those players who use these moves consistently to those who don't. We hypothesize that players who use these moves consistently will be better prepared to learn Newton's first law of motion in class having developed this implicit foundational knowledge.

5.2.2 Mining Sequences of Clicks: Evidence of Newton's Second Law

To seek evidence of implicit knowledge of Newton's Second Law of motion ($F=ma$), we analyzed sequences of fast clicks. In specific, we looked at the length of sequences where players clicked near each color particle to move it. Each color of particle has different mass and size, represented by the different colors. By looking at how frequently the players click near the same particle in a short amount of time, we can see if they recognize that more massive particles require a greater degree of force to be moved the same distance—or if they confuse mass and size.

We examined this for a range of operationalizations of a "short time", e.g., fast clicking, treating the cut-off as being 1 s, 2 s, up to 10 s. The overall pattern of results was very similar across time lengths; within this chapter, we will just show values for 4 s, a time threshold long enough to include all students repeatedly clicking to move the same particle, but brief enough for students to avoid cases where the student is clicking on the same particle for different reasons. So, for each particle color, we looked for cases where a student clicked to move the same particle (as coded by the human coder) in under 4 s after the previous action. Then we look for how many times this happened in sequence (which would be 1 if the player clicked to move a particle once in under 4 s after the previous action and then did something else; 2 if the player clicked to move the same particle twice in under 4 s after the previous action and then did something else, and so on).

Within this analysis, we compared the sequence length for different particle colors, across all sequences. A between-subjects comparison was used, as different students played different levels and therefore received different particles (and some students did not click near all the particles they saw). This discards some within-subjects information leading to a conservative assumption (leading to less statistical power to find significant results). We compared each color particle to each other color particle, using a two-sample *t*-test. Then we applied the Benjamini and Hochberg (1995) post-hoc correction to control for having run six statistical tests. Benjamini and Hochberg is a "false discovery rate" post-hoc method that controls for the number of tests run while avoiding the over-conservatism that characterizes family-wise error rate methods such as the Bonferroni correction.

The Benjamini and Hochberg correction requires a smaller p value, for significance, varying by test (within this method, some tests in a set end up requiring a lower p value than others for significance). Three of the six differences between sequence length are statistically significant according to this test: grey versus red ($t(40)=5.25$, $p<0.001$, $\alpha=0.008$), grey versus blue ($t(31)=3.76$, $p<0.001$, $\alpha=0.017$), white versus red ($t(57)=2.98$, $p=0.004$, $\alpha=0.025$). A fourth was marginally significant, white versus blue ($t(48)=2.07$, $p=0.04$, $\alpha=0.03$). The remaining two tests were not significant, white versus grey ($t(37)=1.65$, $p=0.11$, $\alpha=0.042$) and blue versus red ($t(51)=0.49$, $p=0.63$, $\alpha=0.05$). This pattern of results is more clearly shown in Fig. 15.2.

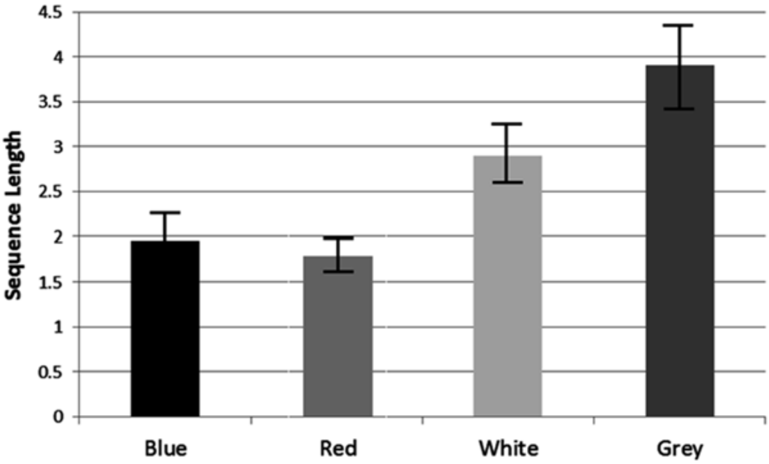


Fig. 15.2 The average sequence length for the student quickly clicking each color particle. Standard error *bars* shown

These findings show that players are markedly differentiating the particles in terms of their mass, which is consistent with an implicit understanding of Newton's second law of motion. In the game, the mass of the balls is near equal for the red and blue balls, and for the white and grey balls. Players' behavior in the game are consistent with their differentiating these masses; they treat the red and blue ball similarly, but click more (impart more force) to accelerate the white and grey balls. Furthermore, the grey ball has a smaller radius of any of the other balls (as if it were made of a much more dense material) yet players still distinguish the mass from size as the factor causing the acceleration, demonstrating possible evidence of implicit understanding that the two particles have different relative density.

A second potential test of this is how far players click from the various particle colors, as closer clicks create a greater force on the object. We can compute this by looking at the distance the player was away from the particle when he or she clicked, with that particle as a target, and then computing a two-sample *t*-test with Benjamini and Hochberg adjustment (e.g., the same test as conducted immediately above) to compare between particles colors. In this case, we find that three of the six statistical tests are significant: red versus white, ($t(89) = 5.17$, $p < 0.001$, $\alpha = 0.008$), grey versus white ($t(49) = 4.95$, $p < 0.001$, $\alpha = 0.017$), and blue versus white ($t(33) = 4.82$, $p < 0.001$, $\alpha = 0.025$). In other words, players always clicked further away from the white particle than the other particles. The remaining three tests were not significant: grey versus red ($t(68) = 1.36$, $p = 0.18$, $\alpha = 0.03$), blue versus red ($t(86) = 0.69$, $p = 0.49$, $\alpha = 0.042$), and blue versus grey ($t(46) = 0.65$, $p = 0.52$, $\alpha = 0.05$). Therefore, there were no differences in click distance from the other particles. Note that the degrees of freedom are higher for these tests than for the previous set of tests; more students clicked near a particle of a certain color at least once, than clicked near that particle in under four seconds. The pattern of results for particle distance is more clearly shown in Fig. 15.3.

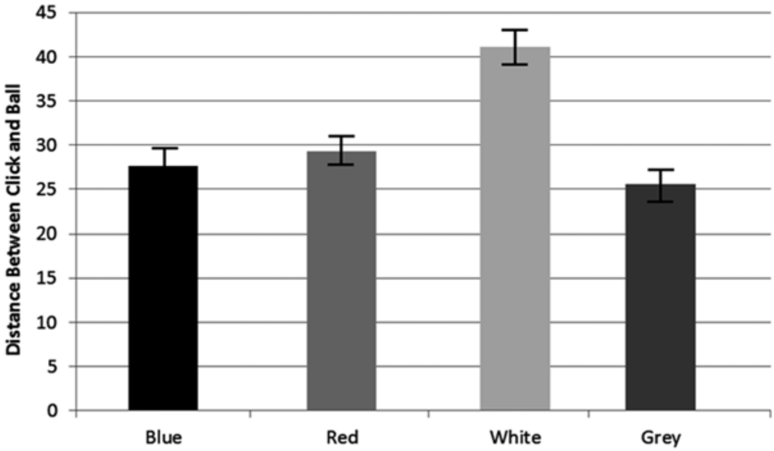


Fig. 15.3 The average distance (*pixels*) away that the student clicked each color particle. Standard error *bars* shown

Players treat most of the particles the same with regard to distance of the impulse, but the white particle appears to be an exception. This may likely be due to the larger radius of the white particle (it appears much larger than the other particles on the screen). This finding may be explained by the fact that the balls were in motion, so players' accuracy in distance may have been compromised. The finding further highlights players' ability to distinguish that it is the mass of the ball, rather than the size, that is important in the relationship between force and acceleration.

6 Discussion of this Approach for Serious Game Analytics

The results from this research provide a model set of methods to use game data logs to detect strategies that may be linked to foundational implicit knowledge that has previously gone unmeasured. We feel this emergent approach to developing a game-based assessment mechanic is particularly well suited to open-ended game spaces with large numbers of play patterns that could serve as evidence of implicit understanding. Table 15.5 provides a summary of how our method connects explicit learning outcomes to implicit game-based knowledge.

We have shown that we can reliably detect a series of strategic moves in *Impulse* data that players were observed using in their quests to get their particle to the goal while grappling with Newtonian mechanics. The use of float, stop, and clear path strategies may indicate players' implicit understanding that the particle will stay in constant motion in the absence of an external force (Newton's First Law).

Even more striking to these authors is players' differentiation between masses of the particles in *Impulse*. The notable difference between clicks near light and heavy particles is a strong indicator of possible implicit understanding of Newton's Second Law. Players use more force to accelerate the heavier particles—even when they are smaller in diameter.

Having built and validated these detectors, we are now applying these detectors to a larger sample of gameplay data from 388 students as part of a national implementation study of 39 classrooms (Rowe, Asbell-Clarke, Bardar, Kasman, & MacEachern, 2014).

Table 15.5 Connecting explicit and implicit science knowledge

Explicit learning outcome	Implicit game-based knowledge	Cognitive strategy	Game-based strategic move
Newton's first law	Each particle will keep moving on its path without an impulse or force from another particle	Slow particle down by using an opposing force	Consistently click in the path of a particle, close enough to stop or slow it down
Newton's second law	The different mass particles react differently to the same force	Impart more force to move heavier particles than lighter particles	Consistently click more frequently next to heavier particles than lighter particles

This user-generated data and distilled features will be inputted into RapidMiner, along with the previously generated W-J48 decision trees. The trees will be applied to the data, producing a prediction for every click of the probability that each of the relevant strategic moves in Table 15.3 was used. Every learner action in this game will be annotated with the probability that the learner was using each of the strategic moves.

We then plan to apply sequential pattern mining (Srikant & Agrawal, 1996) to the data set created by the application of the detector to all students' log data. The annotated logs will show us sequences of student strategic moves over time; sequential pattern mining will allow us to find out whether there are specific combinations of strategic moves that emerge over time and how those sequences are connected to broader learning of the physics concepts present in *Impulse*. Similar strategies have been used to infer whether students form strategies over time in Betty's Brain, a learning-by-teaching environment (Kinnebrew & Biswas, 2012).

Our ability to detect common strategies in the game data logs that are related to learning outcomes is a foundational step in research on implicit learning. Ultimately we are using these data along with many different instruments to measure engagement, attention, and non-cognitive factors that may be influencing the entire learning experience. In such, we are developing new models of learning in which data reveal learning that was previously invisible.

Acknowledgments We are grateful for NSF/EHR/DRK12 grant #1119144 and our research group, EdGE at TERC, which includes Erin Bardar, Teon Edwards, Jamie Larsen, Barbara MacEachern, Emily Kasman, and Katie McGrath. Our evaluators, the New Knowledge Organization, assisted with establishing the reliability of the coding.

References

Asbell-Clarke, J., & Rowe, E. (2014). Scientific inquiry in digital games. In F. Blumberg (Ed.), *Learning by playing: Video games in education*. New York: Oxford University Press.

Asbell-Clarke, J., Rowe, E., & Sylvan, E. (2013, April). Assessment design for emergent game-based learning. *Paper presented at the ACM SIGCHI conference on human factors in computing systems* (CHI'13). Paris, France.

Asbell-Clarke, J., Rowe, E., Sylvan, E., & Baker, R. (2013, June). Working through impulse: Assessment of emergent learning in a physics game. *Paper presented at the 9th annual meeting of the Games+Learning+Society (GLS) conference*, Madison, WI.

Baker, R. S., & Clarke-Midura, J. (2013). Predicting successful inquiry learning in a virtual performance assessment for science. In *User modeling, adaptation, and personalization* (pp. 203–214). Berlin: Springer.

Baker, R. S., Ocumpaugh, J., Gowda, S.M., Kamarainen, A., Metcalf, S.J. (2014) Extending log-based affect detection to a multi-user virtual environment for science. In *Proceedings of the 22nd conference on user modelling, adaptation, and personalization*, pp. 290–300 (To appear).

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*, 289–300.

Clark, D. B., Nelson, B., Chang, H., D'Angelo, C. M., Slack, K., & Martinez-Garza, M. (2011). Exploring Newtonian mechanics in a conceptually-integrated digital game: Comparison of learning and affective outcomes for students in Taiwan and the United States. *Computers and Education, 57*(3), 2178–2195.

Clearleft Ltd. (2013) Silverback (Version 2.0) [Software]. http://silverbackapp.com.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. doi:10.1177/001316446002000104.

Collins, H. (2010). *Tacit and explicit knowledge*. Chicago: University of Chicago Press.

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction, 10*(2/3), 105–225. doi:10.2307/3233725.

Fisch, S. M., Lesh, R., Motoki, E., Crespo, S., & Melfi, V. (2011). Children's mathematical reasoning in online games: Can data mining reveal strategic thinking? *Child Development Perspectives, 5*(2), 88–92.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy* (1st ed.). New York: Palgrave/Macmillan.

Gee, J. P. (2007). *What video games have to teach us about learning and literacy* (2nd ed.). New York: Palgrave/Macmillan.

GlassLab (2014). Psychometric considerations in game-based assessment. Institute of Play. Retrieved July 1, 2014, from http://www.instituteofplay.org/work/projects/glasslab-research/

Halverson, R., Wills, N., & Owen, E. (2012). CyberSTEM: Game-based learning telemetry model for assessment. Presentation at 8th Annual GLS, Madison, WI.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29–36. doi:10.1148/ radiology.143.1.7063747.

Kinnebrew, J. S., & Biswas, G. (2012). Identifying learning behaviors by contextualizing differential sequence mining with action features and performance evolution. In *Proceedings of the international conference on educational data mining*, pp. 57–64.

McCloskey, M. (1983). Intuitive physics. *Scientific American, 248*(4), 122–130.

Minstrell, J. (1982). Explaining the "at rest" condition of an object. *The Physics Teacher, 20*(1), 10–14.

Mislevy, R., & Haertel, G. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6–20.

Moore, A.W. (2003) Cross-validation for detecting and preventing overfitting. *Statistical Data Mining Tutorials*.

National Research Council. (2011). Learning science through computer games and simulations. In M. A. Honey & M. L. Hilton (Eds.), *Committee on science learning: Computer games, simulations, and Education*. Washington, DC: National Academies Press.

Pardos, Z.A., Baker, R.S.J.d., San Pedro, M.O.C.Z., & Gowda, S.M. (2013). Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. *Proceedings of the 3rd international conference on learning analytics and knowledge*, pp. 117–124.

Plass, J., Homer, B. D., Kinzer, C. K., Chang, Y. K., Frye, J., Kaczetow, W., et al. (2013). Metrics in simulations and games for learning. In M. Seif El-Nasr, A. Drachen, & A. Canossa (Eds.), *Game analytics: Maximizing the value of player data* (pp. 694–730). London: Springer.

Polanyi, M. (1966). *The tacit dimension*. Chicago, IL: University of Chicago Press.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco: Morgan Kaufmann.

Rowe, E., Asbell-Clarke, J., Bardar, E., Kasman, E., & MacEachern, B. (2014, June). Crossing the bridge: Connecting game-based implicit science learning to the classroom. *Paper presented at the 10th annual meeting of Games+Learning+Society*. Madison, WI.

Rowe, E., Baker, R., Asbell-Clarke, J., Kasman, E., & Hawkins, W. (2014, July). Building automated detectors of gameplay strategies to measure implicit science learning. *Poster presented at the 7th annual meeting of the international educational data mining society*, July 4–8, London.

Sabourin J, Mott B, Lester J (2011) Modeling learner affect with theoretically grounded dynamic Bayesian networks. In *Proceedings of the 4th international conference on affective computing and intelligent interaction*. Memphis, TN, pp. 286–295.

Sao Pedro, M. A., Baker, R. S. J., Gobert, J., Montalvo, O., & Nakama, A. (2013). Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. *User Modeling and User-Adapted Interaction, 23*(1), 1–39.

Sao Pedro, M., Baker, R.S.J.d., & Gobert, J. (2012) Improving construct validity yields better models of systematic inquiry, even with less information. In Proceedings of the 20th international conference on user modeling, adaptation and personalization (UMAP 2012), pp. 249–260.

Shute, V. J., Masduki, I., Donmez, O., Kim, Y. J., Dennen, V. P., Jeong, A. C., et al. (2010). Assessing key competencies within game environments. In D. Ifenthaler, P. Pirnay-Dummer, & N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 281–309). New York: Springer-Verlag.

Shute, V., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. Cambridge, MA: MIT Press.

Shute, V., Ventura, M., Bauer, M., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning? Flow and Grow. *Serious Games: Mechanisms and Effects, 1*(1), 1–33.

Shute, V., Ventura, M., & Kim, J. (2013). Assessment and learning of qualitative physics in Newton's playground. *Journal of Educational Research, 106*(6), 423–430. doi:10.1080/00220671.2013.832970.

Srikant, R., & Agrawal, R. (1996). *Mining sequential patterns: Generalizations and performance improvements* (pp. 1–17). Berlin, Germany: Springer.

Thomas, D., & Brown, J. S. (2011). *A new culture of learning: Cultivating the imagination for a world of constant change*. Lexington, KY: CreateSpace.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Part VI
Serious Games Analytics Design Showcases

Chapter 16
A Game Design Methodology for Generating a Psychological Profile of Players

Emmanuel Guardiola and Stephane Natkin

Abstract Can we track the psychological traits or the profile of a player during gameplay, at least when the player is engaged in a ludic experience? We propose a game design methodology dedicated to the generation of a psychological profile of the player. Our experiment, a vocational guidance game, was created in collaboration with academic experts and industry game developers. Our first results form the basis for exploration of a field at the crossover of computer science (in particular, game design), psychology, and cognitive science.

Keywords Game design • Player profiling • Player model • Psychological profile • Vocational guidance • Holland model

1 Introduction

Game design has a fundamental link with the creation of a player model (Guardiola & Natkin, 2010). We assume a connection between player models in game design and psychological models, but we do not have clear insight into the interaction between psychological traits and a profile established during a gameplay session. There is no such methodology that allows us to create games that generate psychological profiles based on a scientifically validated model. We propose a new game design methodology to fill this gap.

E. Guardiola (✉) • S. Natkin
Conservatoire National des Arts et Métiers (CNAM), CNAM-CEDRIC,
292 Rue St Martin, FR-75141, Paris, Cedex 03, France
e-mail: emmanuelguardiola@gmail.com; stephane.natkin@cnam.fr

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based Learning, DOI 10.1007/978-3-319-05834-4_16

1.1 Why Create Games to Observe and Build a Model of the Player?

In 1984, Sherry Turkle (2005) affirmed that video games can be used as projection tests in psychoanalysis. Practitioners use games in mediation therapy, as one can see in the recent work of Geoffroy Willo (2012) and Michael Stora (2006). Titles like *The Sims* (Maxxis, 2000) or *ICO* (ICO Team, 2002) become tools to illuminate the psychological process.

In the video game entertainment industry, player evaluation tools are embedded in games to enhance player experience, to adapt the content to given gameplay profile dimensions. *Left 4 dead* (Valve, 2008) is a notable example: the AI director maintains rhythm and tension by tracking player behavior and by adapting the configuration of future waves of enemies to this. *Silent Hill Shattered Memories* (Climax, 2009) is an example of a psychological approach, where the game designers try to manipulate players' emotions using the Five Factor Model (Costa & McCrae, 1992).

Some other motivations are ethically more questionable, such as monitoring the behavior of players/consumers in order to adapt a marketing process to them. Zynga, creator of the free-to-play game *Farmville* (Zynga, 2009), was founded by former Amazon.com staff. Zynga links a consumer model to their games to make their business model work.

Therapeutic games sometimes need to create a player profile in order to assure the efficacy of the desired effect. SGCogR, also known as *Le village aux oiseaux* (Tekneo, 2012), measures player attention to adapt content and reduce the difficulty of the play session (Mader, Natkin, & Levieux, 2012). As these embedding profiling tools are based on cognitive aspects, they do not help to measure the psychological dimensions.

More closely aligned to our goals, there are some serious games in the domain of human resources or assessment that attempt to create profiles of the player—in the context of a recruitment process, for instance, *America*'*s Army* (Secret Level, 2006) is a famous example of this.

As we can see in these examples, there is a clear tendency towards using player profiling linked to a psychological model, but these efforts remain empirical. We propose a scientific game design methodology dedicated to psychological profiling.

1.2 Establishing a Psychological Profile of the Player

On the one hand, gameplay is a succession of player decisions; on the other hand, analyzing decisions allows psychological measurement: player inputs can be transformed into psychometric items. To reach this objective, a new method of game design must be created that is scientifically challengeable.

If we look with a scientific approach at Bartle's player typology (Bartle, 1996), we can identify several biases, which are reproduced in the majority of typologies developed by other researchers since Bartle's work. The typology is not based on an existing and scientifically documented psychological model. The questionnaire takes place after the player's engagement in the gameplay. The typology is dedicated to a specific genre, the Multi User Dungeon, a MMORPG ancestor. We want our method to be based on data collected during the gameplay session using a scientific psychological model.

Entertainment games using psychological model such as *Silent Hill Shattered Memories* do not provide results even if they sometimes talk about their process (Mountain, 2010). In the *Silent Hill* case, the psychological model has no strong connection with the intended result. The lesson for us: The chosen psychological model must match the profiling intention.

Assessment games try to create a connection between player activity and skill profiling. We lack scientific studies on experimentation and efficiency of their methods. To succeed, a serious game must have a clear goal and a way to evaluate its impact. In the case of psychological profiling, we need to compare the results from the game to the ones based upon reliable data; for instance, traditional questionnaires.

We have the basis of a framework for our methodology:

- The psychological model must have been scientifically validated
- The data are collected during the gameplay session
- The psychological model must match the profiling intention
- We must be able to compare the result from the game to the ones from reliable means

2 Game Design Methodology

Following our framework, we propose a game design process in nine steps:

1. Identify a scientifically tested and documented psychological model.

   At this step, we just begin the exploration of the link between the player's decision and psychological traits. The model might have a referential test to compare results given by the game and the ones from the traditional process.
2. Identify the set of constraints bound to the type of audience and to the "useful" aim of the game.
3. Create a pipeline for content validation by experts.

   Concept development and production pipelines must allow collaboration between the game designer and experts from the model's domain, for instance, through joint working sessions or reviews.
4. Set the game concept and resolve conflicts between models.

Like in any game production, we need to set the type of game experience for the target audience. In the case of a psychological profiling game, the team must confront the player model with the psychological model to identify potential bias.

5. Define psychometric items with various game proprieties and include them in the gameplay loop.

   At this stage, we don't know which player activity data have the most chance to correlate with dimensions from the psychological model. Even if we can intuitively establish some link between activities and dimensions, we might adopt a cautious approach and determine a large spectrum of items, with different levels of connections with gameplay. This work is done at the same time as designing the gameplay loop.

6. Adopt a method to rationalize the connection between game design content and psychological dimensions.

   During the conception phase, the link between game design elements and the dimension to measure must be clearly established. We must use a rational method to evaluate game design elements in four major areas: game system, interaction staging, motivational structure, and narrative context. For instance, if the psychometric item in the game is the choice of an object among others by the player, we must know the effective link between each object and the psychological dimension it must increment.

7. Build a documented interface to provide an access to the results.

   In the evaluation of the game's "useful" aspect, we need to share the way a gameplay element becomes an item and how it is used into the calculation of the profile. This is required for the tuning of the measurement and the scoring algorithm. And finally, the experts using the game need access to the calculation process to deliver efficient feedback and reports to their subjects.

8. Use a test and analysis protocol to evaluate the quality of items.

   The traditional protocol to evaluate a new psychometric test consists of comparing the result of an existing test with the result of the new one. In our case, the subject fills out the questionnaire of the classical psychological test and plays the game. The evaluation result has for its first objective to identify the origin of potential failures and to help us to tune the game content or the scoring algorithm. Then, it will be used to evaluate if the game attains its "useful" objective.

9. Iterate with item definition (5) or content definition (6).

The iteration continues until the game shows a strong connection between profiles generated by the game and the result of the traditional questionnaire (Fig. 16.1).

The originality of the method is in the coexistence of both the player model and the psychological model: choosing the model (1); resolving conflicts between models during conception (4); linking the psychometric items with the gameplay loop (5); rationalization of the link between game elements and dimensions of the psychological model (6); the scientific evaluation protocol (8).

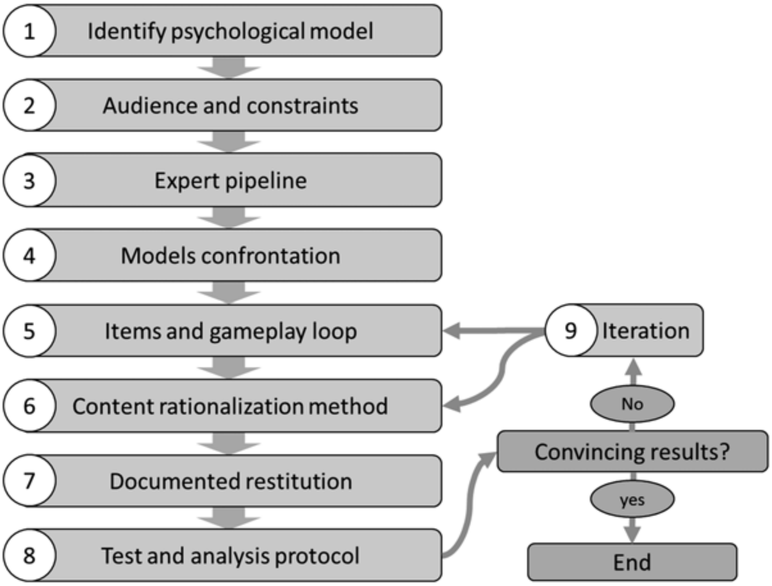

1 Identify psychological model
2 Audience and constraints
3 Expert pipeline
4 Models confrontation
5 Items and gameplay loop
6 Content rationalization method
7 Documented restitution
8 Test and analysis protocol
9 Iteration
No
Convincing results?
yes
End

Fig. 16.1 Game design method synthesis

3 Application to JEU SERAI Development

The main experiment was the development of a serious game. Its title JEU SERAI is a play on words in French. "Jeu" (game) and "Je" (I) have the same pronunciation. You hear "I will be" and read "Game will be".

The ambition of JEU SERAI is to help the vocational guidance of students and adults at an individual level. It proposes to use fundamental mechanisms of games to evaluate users' professional interest, motivation, and the way they take decisions. The goal is not to be a substitute for a career adviser, but to offer an engaging way to make self-assessment psychometric tests.

The game was developed in a consortium composed of two companies (Wizarbox and Seaside Agency), two universities (CNAM and UPOND), and an adult training association (ARCNAM Poitou-Charentes). As with many game productions, we had timing and budget constraints.

In this section, we report how the method was applied, step by step, during production.

3.1 Identify Psychological Model

We worked with experts in psychometric and vocational guidance (CNAM-INETOP and UPOND) to identify a suitable model. We chose the work of John Holland, considering that six types or dimensions (RIASEC: Realistic, Investigative, Artistic, Social, Enterprising, Conventional) are sufficient to get a profile of vocational interest (Holland, 1966) (Fig. 16.2).

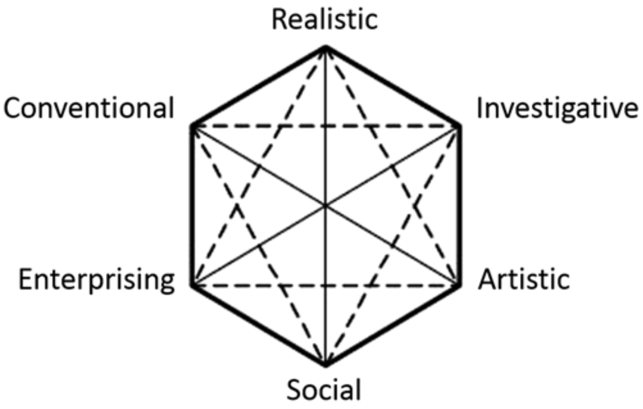


Fig. 16.2 Holland RIASEC model

Several points led us to this choice. This model has been used since the 1960s and was scientifically tested. Vocational guidance tests using this model have been used in educational institutions for decades. By using the two or three best dimensions of a subject, we can access a large typology of profiles (30–120 combinations), and it gives depth to the result. Also, the RIASEC types are interdependent. For instance, if you are strong in Artistic, you might be low in the opposite one, the Conventional: it helps the evaluation of the profile quality. Also, the Holland model measures the interest for a professional environment using profession lists, places, activity verbs, and sometimes images. These references are useful material to create a video game setting (gameplay actions, characters, environments…).

3.2 Audience and Constraints

The audience is heterogeneous, from middle school students to adults in a career change. Some can be familiar with computers or games, some others are not. Game design choices, including interaction design and game theme, must take into account this wide audience constraint.

3.3 Expert Pipeline

We began concept development for this game by sharing our fields of expertise: game developers were trained in vocational guidance tools and issues; experts were trained in game design principles. Then we established a pipeline including technical, gameplay, and expert review all along the development process (Fig. 16.3).

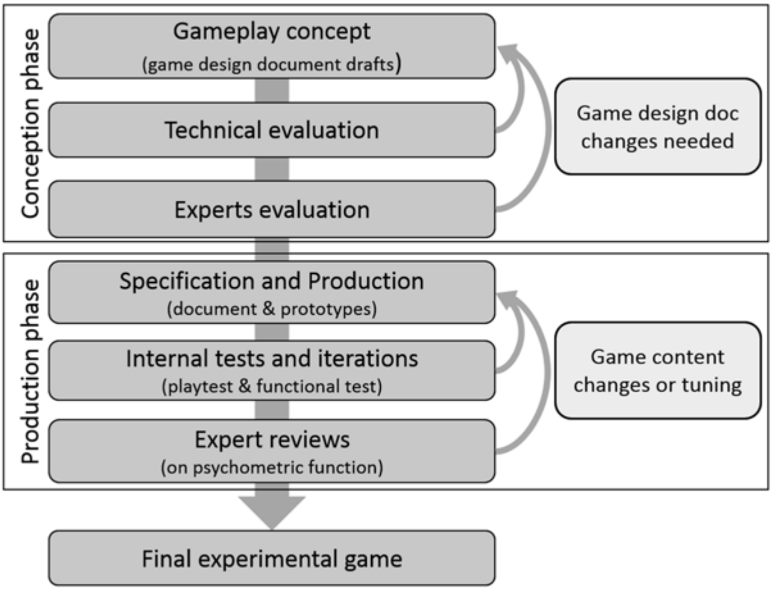

Conception phase
Gameplay concept
(game design document drafts)
Technical evaluation
Experts evaluation
Game design doc changes needed
Production phase
Specification and Production
(document & prototypes)
Internal tests and iterations
(playtest & functional test)
Expert reviews
(on psychometric function)
Game content changes or tuning
Final experimental game

Fig. 16.3 JEU SERAI expert pipeline

We were successful in respecting the pipeline during concept development and at the beginning of the production. As the production deadline got closer, some assets and gameplay final choices were implemented and tuned without having rigorous expert reviews.

3.4 Models Confrontation

During the concept development phase, the definition of the type of gameplay is a good example of the possible conflicts between game design and psychometric models.

The first game concept was based on the measure of strategic preferences of the player. We defined a succession of challenging situations you can solve by using a set of recurrent gameplay tools. For instance, you have to enter a place protected by a security guard. Do you try to convince (dialog tool) or fight him (combat gameplay)? Do you try to crack the security code on the back door (puzzle gameplay) or climb on the roof to find another entry (platform gameplay)? It appears that this approach might lead us to measure the gameplay optimization process of the player. Even if you are not a fan of fighting, you might use this strategy the second time because you master it and overcome challenges quicker. It was not the most effective approach to measure a preference in Holland theory.

We chose to use mini games to avoid the player optimizing gameplay. Each mini game is related to a Holland RIASEC type. It measures players' activity and their

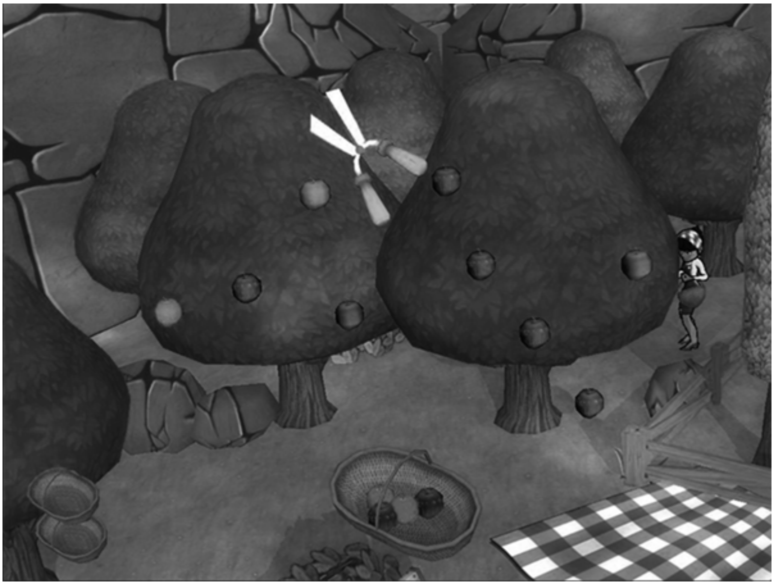


Fig. 16.4 Screenshot of the mini game Nice Apple

appreciation of the mini game. Life simulation games like Animal Crossing (Nintendo, 2008) or MySims (Electronic Arts, 2007) have structure, theme, and gameplay to provide this kind of content for a broad audience. We conceived 18 mini games, 3 for each type, dispatched in an open world, and distributed over three chapters (Fig. 16.4; Table 16.1).

3.5 Items and Gameplay Loop

JEU SERAI has lots of different types of gameplay, due to the mini game structure. For each mini game, at a micro gameplay level we set some psychometric items related to the activity and/or the performance of the players. For instance, for the apple harvesting mini game, we track if players fill baskets with ripe, rotten, or green apples, and how long they take to accomplish the task. We call the ensemble of these mini game items the "Score Item".

We also work higher up, at the game loop level, to enrich the items' harvesting. The life simulation genre offers us lots of other possibilities to collect data. We chose two other ways that were very different in terms of presentation to the players.

First, the players answer questionnaires integrated into dialog with non-playable characters. For instance, the mayor of the village needs to choose the next use of an abandoned house before its restoration. The players have to vote for the future use of this house: library, day nursery, police station… each choice is related to a

Table 16.1 JEU SERAI mini game list

RIASEC type	Mini game name/“French original name”	Metaphoric activity
Day 1 theme: welcome to the village		
R	Nice Apple/“La bonne pomme”	Apple harvesting
I	Invasion of the Ants/“L’invasion des fourmis”	Ant observation
A	Autumn fashion/“Création d’automne”	Designing a t-shirt
S	Ms. Petitpas shopping/“Les courses de Mme Petitpas”	Taking care of a senior lady
E	Omelet/“Omelette”	Directing a team in a workshop
C	Classified!/“Classe!”	Classifying stamps
Day 2 theme: the tempest		
R	Tile/“La tuile”	Repairing a roof
I	Light/“Lumière”	Finding the cause of a breakdown
A	Photos	Photo journalism
S	Take shelter!/“Aux abris”	Convincing people to take shelter
E	Sand and sweat/“Du sable et de la sueur”	Leading a team of porters
C	Stock	Stock management
Day 3 theme: the village party		
R	Assembly/“Assemblage”	Assembling theater set elements
I	Chemistry set/“Le petit chimiste”	Manipulating a chemistry set
A	Decoration/“Déco”	Art direction of a stage set
S	Reception/“Accueil”	Welcoming visitors
E	Promotion	Managing the ticket sellers
C	Good seat/“Chacun sa place”	Being the usher in the theater

Table 16.2 JEU SERAI recurrent activities

RIASEC	Recurrent activities the players can do
R	Water plants/juggle with soccer ball
I	Discover mushrooms/solve the weekly municipal puzzle
A	Sculpt shrub/draw with chalk stick
S	Donation to the red cross/talk with anxious inhabitants
E	Move inhabitants from place to place
C	Pick up detritus

RIASEC type. We distributed six of these narrative questionnaires in the village. They are presented as optional in the players’ to do lists.

Second, while exploring the village the players can interact with a lot of elements (peoples, plants, objects…). Each of these possible interactions were also linked to RIASEC types as reported in Table 16.2. This type of interaction is optional and not presented in the to do lists (Fig. 16.5).

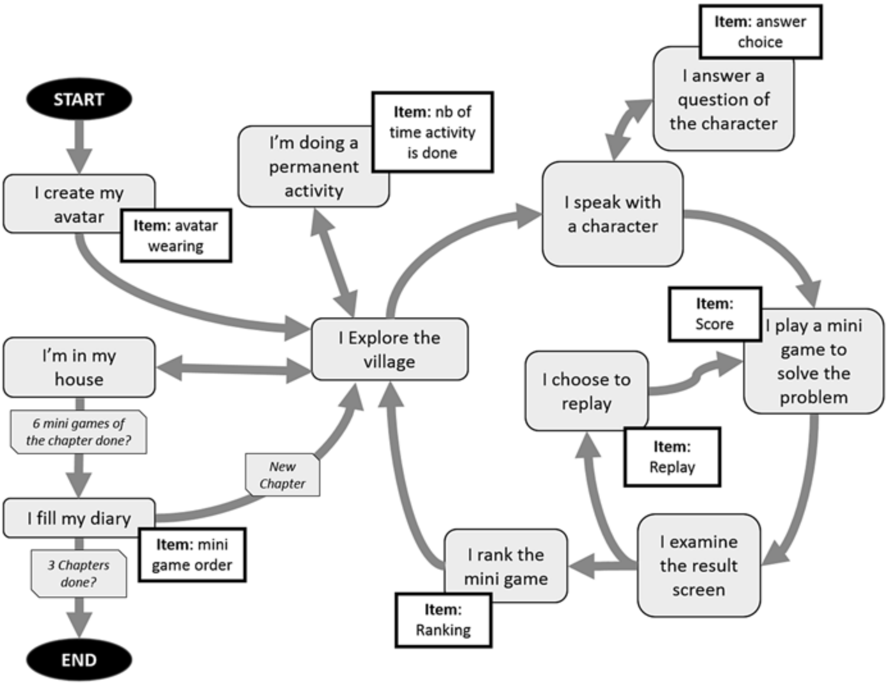


Fig. 16.5 Final game loop created during the conception phase

During the elaboration of the final game loop, we formalized all activities the players can do in the game, from launching the session to the end of it. We also saw the opportunity to add a new item to the game. At the very beginning of the game, the players create an avatar: we design the clothing choices in connection with RIASEC types.

The final list of psychometric items offers a large variety of presentation types and affiliations to categories of content (Table 16.3).

3.6 Content Rationalization Method

As we consider the preference of the players for mini games as a key element for the generation of a RIASEC profile, we must design game elements paying attention to the dimension they are associated with. For instance, in a mini game dedicated to the Realistic dimension measure, like "Nice Apple", what is the most adequate mouse cursor? A classical arrow or a tool, like pruning shears? By working with experts and references (Demangeon, 1984; Vrignaud & Cuvillier, 2006), we were able to create a RIASEC classification of game elements.

Table 16.3 Variation scope of items

Item	Gameplay		Ranking		Questionnaire	Simulated activity	
Presentation	Mini game score (regroup micro gameplay items)	Mini game replay	Mini game ranking	Mini game diary ordering	Answers to narrative questionnaires	Recurrent activities	Initial choice of wearing
Mandatory, in the player to do list	X	X	X	X			X
Optional, in the player to do list					X		
Optional, not in the to do list						X	
Explicit measure			X	X			
Hidden measure	X	X			X	X	X

First, we had to determine the types of game design elements to evaluate. We came up with a classification from our game design and empirical experience of production (Emmanuel Guardiola has 15 years of experience in game design methodology, used on more than 30 published titles). The evaluated elements are the ones we can link to the main interaction loop between the players and the game (Swink, 2009). They are elements the players can interact with, the ones they can perceive and/or are part of the gameplay understanding. We also add the action verbs for each mini game describing the gameplay, the type of objective, and the metaphor or simulated situation.

We used this tool during the entire production phase and synthesized in a table. It is an indicator of the connection quality between the mini game content and RIASEC dimensions.

The legend for the cell contents in Table 16.4 (below) refers to the Holland hexagon typology: "++" means we judge the element as matching the correct type, "+" means it matches with an adjacent type; "–" means matching with a nearly opposite type; "– –" means matching with the opposite type; an empty cell means a neutral or non-pertinent element.

Table 16.4 represents the state of the indicators at the end of the production. It shows that we evaluated that most of the elements are related to the right RIASEC type.

Table 16.4 RIASEC/game elements connection indicator synthesis

	Mini games	Simulated activity	Gameplay action verbs	Objective	Cursor	Handled elements' interaction target	Interaction feedback	Mini game environment	Non-interactive animated element	Triggering place	Triggering character
R	Nice apple	++	++	++	++	++	++	++	++	++	–
I	Invasion of the ants	++	++	++	++	+	++	+		–	++
A	Autumn fashion	++	++	++		++	++	++		++	++
S	Ms. Petitpas shopping	++	++	++		++	++	++	++		++
E	Omelet	++	++	++		++	++	++	++	–	–
C	Classified!	++	++	++		++	++	++	++	++	++
R	Tile	++	++	++	++	++	++	++	++	–	–
I	Light	++	++		+	+	++	++			++
A	Photos	++	++	++	++	++	++				–
S	Take shelters!	++	++	++		++	++			++	++
E	Sand and sweat	++	++			++	++	++	++		+
C	Stock	++	++	++		++	++	+	+	+	+
R	Assembly	++	+	++		++	++	–		++	++
I	Chemistry set	++	++	++		++	++	++		++	++
A	Decoration	++	++	++		++	++	++		++	++
S	Reception	++	++	++		+	++	++	++	+	+
E	Promotion	++	++	++		++	++	++		++	++
C	Good seat	++	++	++		+	+	–	–	++	++

3.7 Documented Results Interface

There are two types of restitution in *JEU SERAI*. First, there is one for the player at the end of the game/test session. It is a screen summing up the score using the RIASEC hexagon and short commentaries created by the experts. Second, we provide a data file containing all the items' values and easily usable in data analysis software or in a spreadsheet. Starting with the first version of the game, we maintained a document clearly explaining the link between in-game item value and the RIASEC dimension score.

3.8 Test and Analysis Protocol

In the first test session, our goal was not to officially validate the psychometric use of *JEU SERAI*. The chosen protocol must help us to sort out the efficient types of psychometrics items and to localize causes of failures.

In *JEU SERAI*, the statistical properties of the Holland model enable the external correlation method. It compares the *JEU SERAI* items' aggregations to results obtained by the subjects in a reference classical test, here IRMR3 (Holland interest test published by ECPA—PEARSON).

The main tool is the Bravais-Pearson correlation coefficient (r). It expresses intensity and direction (positive or negative) of a linear relation between two quantitative variables. The value r is between -1 and 1. The correlation is considered as significant over 0.4, medium between 0.2 and 0.4, insignificant below 0.2.

One hundred and forty people took the test between September 2011 and January 2012. One hundred and three were students from the UPOND University, 37 were adults in career changes from the ARCNAM-Poitou-Charentes. The recruitment was done through social networks and flyers. Students received a 20 Euro compensation. The total test duration, game and questionnaire, was close to 2.5 h. At the end of the test, each subject received their RIASEC profile through an IRMR3 questionnaire and through *JEU SERAI*. Commentaries and feedback was collected after the test.

Above are two examples of correlation coefficient tables produced during the analysis. The grey-tinted cells highlight where we originally supposed the correlation would be. The bold-italic values indicate where the correlation coefficient is actually the highest.

Table 16.5 represents the correlation between the subject's top dimension in the RIASEC questionnaire and the amount of clicks they did on the different recurrent activities. If in the questionnaire the player gets his best score in the Realistic dimension, we expect you to click more on the third and sixth recurrent activities, designed for your profile. As you can see, only one activity matches with his RIASEC type. This specific way to use recurrent activities in the evaluation of your profile is not appropriate.

Table 16.6 represents the correlation between the ranking the player gave to a mini game just after playing it and their top dimension in the RIASEC questionnaire. The analysis of this table shows that most "Mini Game Ranking" items seem promising. It also reveals an issue with a specific mini game. Correlation coefficients are

Table 16.5 Correlation coef.: recurrent activities and RIASEC type questionnaire

	Champignons Investigative	*Collecte* Social	*Plantes* Realistic	*Détritus* Conventional	*Buissons* Artistic	*Football* Realistic	*Ardoise* Artistic	*Suivre* Enterprising	*Consoler Calmer* Social	*Enigme* Investigative
R_IR Realistic	-0,016	0,002	-0,17	-0,141	***0,171***	0,093	-0,069	***0,263***	0,086	-0,352
I_IR Investigative	***0,171***	-0,059	***0,388***	-0,097	-0,082	***0,182***	***0,129***	-0,062	0,04	-0,101
A_IR Artistic	-0,241	***0,107***	-0,408	-0,076	-0,041	-0,001	0,107	-0,04	-0,239	0,019
S_IR Social	0,101	-0,235	-0,14	0	-0,081	-0,136	0,031	-0,052	-0,013	***0,26***
E_IR Enterprising	-0,024	-0,29	0,242	***0,102***	-0,136	-0,096	-0,2	-0,075	-0,019	0,223
C_IR Conventional	0,03	0,075	0,1	-0,006	-0,038	0,01	-0,228	0,026	***0,098***	0,084

Table 16.6 Correlation coef.: mini game ranking and questionnaire RIASEC type

Chapter 2 Mini game Ranking item	Tile__ranking Realistic	Light__ranking Investigative	Photos__ranking Artistic	Shelters__ranking Social	SandSweat__ranking Enterprising	Stock__ranking Conventionall
R_IR	***0,42***	0,3	-0,3	-0,12	***0,24***	0,17
I_IR	0,29	***0,35***	-0,07	0,02	0,2	-0,16
A_IR	-0,15	-0,19	***0,51***	0,03	-0,08	-0,31
S_IR	-0,05	-0,11	-0,04	***0,43***	0,01	-0,28
E_IR	-0,16	0,01	-0,29	-0,2	-0,37	0,16
C_IR	-0,03	0,18	-0,47	-0,32	-0,08	***0,44***

significant except for “Sand and Sweat”. In fact, during the whole analysis, most of the elements addressed to the Enterprising RIASEC type do not work correctly. We have identified two main reasons so far: the Enterprising type is underrepresented in our participants panel; and we do not express well the gameplay action verbs of Enterprising mini games during the content rationalization process.

The Bravais-Pearson correlation coefficient method allows us to sort out working items, to reveal issues, and to identify their causes.

3.9 Iteration

Following the tests, we were able to establish a list of modifications and recommendations to improve the game and optimize the profiling process. Some come from the analysis of correlation, like the need to rework the Enterprising type mini games. Some others come from the comments and observations collected during the test period. For instance, the mini game "Classified" (stamp ordering) was considered too easy and it corrupted the score items and ranking items.

The development team has already implemented some of the changes and we have finished a complete cycle of the method.

3.10 JEU SERAI Experimentation Conclusion

The chosen approach for *JEU SERAI* is to plunge the player into an environment where multiple mini games are available. These mini games reflect different dimensions of the RIASEC types. The game measures gameplay items and the interest of the player for each mini game. The village narrative context and the life simulation genre are favorable settings to implement more items as narratively hidden questionnaires or optional recurrent activities also linked to the RIASEC model.

The experiment confirms predictable issues such as the tension between psychological models and the player model during the concept development phase. It also raises some others such as the difficulty to define categories of gameplay elements like the description of player actions in the mini games. Finally, the test and analysis protocols from psychometrics integrate a scientific approach to evaluating content and item correlation.

4 Discussion

The proposed method succeeds in using the constraints of a psychological model to define the framework of a game design task. Despite the divergences between models, we established links between player choices and analyzable items. This work, at the boundaries of computer science and human sciences, can be used as reference material to set future player behavior analysis tools.

The total iteration took 24 months. During the entire production, we defined tools dedicated to the profiling purpose of the game, such as: the table of item presentation variation; the categorization of the elements related to gameplay and interaction; the rational evaluating system of the link between game design content and RIASEC.

Through the use of psychometric protocols, the method successfully ranked item type quality. "Mini game diary ordering," "Mini game ranking," and "Narrative questionnaires" are items with the best number of correct correlations (to the right RIASEC type). "Mini game diary ordering" tops the chart with 78 % success. In terms of how they are created, they are close to a classical psychometric questionnaire. They contain issues and need to be reworked, but are promising. All the other types of items get very bad results: 10–22 % correct correlations. They help us to reveal new questions.

4.1 The "Mini Game Replay" Item Case

JEU SERAI psychometric items like "Mini game score," "Mini game replay," ("Replay") or the completion of recurrent activities are close to classical data in video games. Player modeling during entertainment playtests includes performance and completion. They are not intentional preference marks from the player, such as one can produce using questionnaires.

These gameplay or behavioral items did not perform as well through the lens of the Bravais-Pearson correlation coefficient. On the other hand, the analysis result gives us an interesting way to use them in the future.

Originally, the "Replay" item looks like a perfect candidate to measure player preference for a mini game. You play it more because you like it. But the correlations tables demonstrate the contrary: for the "Replay" item, only two mini games out of 18 have the best correlation coefficient with the right RIASEC type. Another interesting fact from the analysis: the "Replay" item has its highest correlation coefficient 12 times out of 18 with the Realistic type of subjects, regardless of the anticipated RIASEC type of the mini game.

The "Replay" item value needs to be used in a completely different way to the one we originally thought of. From this new point of view, it is more significant to know if you replay mini games, than knowing which mini game you replay. The total number of times you replay any mini game seems to be an indicator of Realistic type affiliation. Another approach for measuring the RIASEC interest profile appears. A second observation confirmed it.

If we sort out the most replayed mini games during the tests, we notice they contain timing-related gameplay mechanisms. Does this mean that Realistic types are more sensitive to these specific kinds of gameplay? Additionally, the less replayed mini games are the ones related to Social types. Most of them are based on dialog tree choices with no real interest in replayability. In these cases, the "Replay" item value is meaningless for measuring Social type.

The analysis of the "Replay" item results leads us to the notion of gameplay patterns. There is a subtle connection between player gameplay behavior, game design elements, and the psychological dimensions. The item concepts and the rationalization of content process, step 5 and 6 of the method, need to integrate the gameplay pattern notion.

From the perspective of tracking the psychological profile from in-game activity, these observations mean we need to pay attention to the gameplay mechanisms in use. We can also suppose from a larger point of view that if a particular type from a psychological model is strongly connected to a specific gameplay mechanic, whatever the metaphorical activity, maybe another type is more connected to the metaphorical activity than a specific gameplay mechanic. The questions opened by the results of this first experiment are driving our next experimentation protocols, and they suggest going deeper into the relation between the constitutive elements of the game experience and psychological dimensions.

5 Conclusion

The primary contribution of this work is to have tested a scientific method to generate a psychological profile of the player using a video game. Although our problem is positioned in the line of research in psychology and sociology of trying to define what players are and how to model them, our approach is from game design.

We try to understand the player in order to reveal new game design methods and principles, to create better games. A main difference with previous approaches, including Bartle's, is that the evaluation of the player is integrated into the game and not done a posteriori. This framework, strongly connected to ludic aspects of the game experience, fits with the needs of entertainment games. A better understanding of the player and communities is a major goal for industry. Games are becoming online services and profiling the player is becoming a business concern. How do we decide which content must be produced to maintain an active community? A method that rationally links personality typologies and gameplay content should be an important decision-support tool.

The use of a player model finds application in fields other than entertainment. We created an interesting precedent with vocational guidance psychometrics in *JEU SERAI*. If we look at the case of training games, they usually assume that a learner model drives the progression in pedagogic content. Most of the time, this learner model relates to some aspects of the motivational model of the player (in good cases, more elaborate than just earning points). Another example is therapeutic games, which try to combine a patient recovery model with gameplay. We aim to test our method on other types of useful games to improve these processes.

Our contribution accords with statements made in game studies in human sciences. The psychologist Thomas Gaon (2010) highlights the link between specific gameplay patterns and anxiety motivation. At the boundary of computer and cognitive science, we explore the player engagement principle (Soriano, Erjavec, Natkin, & Durand, 2013). To progress in the study of the player, the world's primary consumer of cultural goods, we need to equip ourselves with new tools at the meeting of computer science and human sciences.

Acknowledgment Katharine Neil and Jurie Horneman for the revision work.

References

Bartle, R. A. (1996). Hearts, clubs, diamonds, spades: Players who suit MUDs. *Journal of MUD research* n°1. Republished in Salen, K., & Zimmerman, E. (Eds.). (2006). The game design reader. Cambridge, MA: MIT Press.

Costa, P. T., & McCrae, R. R. (1992). *NEO PI-R professional manual*. Odessa, FL: Psychological Assessment Resources.

Demangeon, M. (1984). *Intérêts, valeurs, personnalité en classe Terminale*. Paris: INETOP.

Gaon, T. (2010). Le ressort de l'angoisse dans les jeux vidéo. Les jeux vidéo au croisement du social, de l'art et de la culture. *Questions de communication. série acte 8/2010*. Nancy: Presses universitaires de Nancy.

Guardiola, E., & Natkin, S. (2010). Player's model: Criteria for a gameplay profile measure. In *Proceeding ICEC 2010, Seoul*. Berlin, Germany: Springer.

Holland, J. L. (1966). *The psychology of vocational choice. A theory of personality types and model environments*. Waltham, MA: Blaisdell.

Mader, S., Natkin S., & Levieux, G. (2012). How to analyse therapeutic games: The player/game/therapy model. In *Proceeding ICEC 2012, Bremen*. Berlin, Germany: Springer.

Mountain, G. (2010) Psychology profiling in Silent Hill Shattered memories. In *Communication, Paris game AI conference 2010*, Paris.

Soriano D., Erjavec G., Natkin S., & Durand M. (2013). Could the player's engagement in a video game increase his/her interest in science? In *Proceeding ACE 2013—Making New Knowledge, Netherlands.* Berlin, Germany: Springer.

Stora, M. (2006). Ico, conte de fée interactif: histoire d'un atelier jeu vidéo. *L'autre, 7*(2), 215–230.

Swink, S. (2009). *Game feel, a game designer's guide to virtual sensation*. Burlington, MA: Morgan Kaufmann.

Turkle, S. (2005). *The second self: Computers and the human spirit* (20th Anniversary ed.). Cambridge, MA: MIT Press

Vrignaud, P., & Cuvillier, B. (2006). *HEXA3D. Questionnaire d'évaluation des intérêts*, Paris: Editions du centre de psychologie appliquée.

Willo, G. (2012). Le "surgissement" cybernétique, un opérateur du transfert dans la psychose. *Adolescence 2012/1, n° 79.*

Games

Climax. (2009). Silent hill shattered memories. Nintendo Wii. Climax studios. Konami.

Electronic Arts. (2007). My sims. Nintendo Wii. Electronic Arts

ICO Team. (2002). ICO. Sony playstation. ICO team SCE Japan Studio. SCE.

Maxxis. (2000). Les sims. PC. Maxis. Electronic Arts.

Nintendo. (2008). Animal crossing let's go to the city. Nintendo Wii. Nintendo.

Secret Level. (2006). America's army rise of a soldier. Microsoft Xbox. Secret level. Ubisoft.

Tekneo. (2012). Le village aux oiseaux. PC. Consortium Tekneo, Seaside Agency, Cnam, Neo Factory, Spirops, Inserm.

Valve. (2008). Left 4 dead. PC. Valve.

Zynga. (2009). Farmville. PC en ligne sur Facebook. Zynga.

Chapter 17
Replay Analysis in Open-Ended Educational Games

Erik Harpstead, Christopher J. MacLellan, Vincent Aleven, and Brad A. Myers

Abstract Designers of serious games have an interest in understanding if their games are well-aligned, i.e., whether in-game rewards incentivize behaviors that will lead to learning. Few existing serious games analytics solutions exist to serve this need. Open-ended games in particular run into issues of alignment due to their affordances for wide player freedom. In this chapter, we first define open-ended games as games that have a complex functional solution spaces. Next, we describe our method for exploring alignment issues in an open-ended educational game using replay analysis. The method uses multiple data mining techniques to extract features from replays of player behavior. Focusing on replays rather than logging play-time metrics allows designers and researchers to run additional metric calculations and data transformations in a post hoc manner. We describe how we have applied this replay analysis methodology to explore and evaluate the design of the open-ended educational game *RumbleBlocks*. Using our approach, we were able to map out the solution space of the game and highlight some potential issues that the game's designers might consider in iteration. Finally, we discuss some of the limitations of the replay approach.

Keywords Alignment • Replay analysis • Open-ended games

1 Introduction

The field of serious games has grown to cover a diverse array of domains and subjects. Along with this growth, there has been a similar broadening of the design space of serious games to include many different game structures. These structures bring new questions. Initial work on serious games looked at whether games could

E. Harpstead (✉) • C.J. MacLellan • V. Aleven • B.A. Myers
Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA
e-mail: eharpste@cs.cmu.edu; cmaclell@cs.cmu.edu; aleven@cs.cmu.edu; bam@cs.cmu.edu

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based Learning, DOI 10.1007/978-3-319-05834-4_17

promote learning. It has generally been found that serious games are capable of accomplishing their goals of fostering learning and so the new questions in the field are what features of games contribute to this success and how can they be made better (Clark, Tanner-Smith, Killingsworth, & Bellamy, 2013). More broadly, as the potential design space of serious games grows, the field of serious games analytics must also grow to accommodate new structures and questions.

One question that has seen less discussion in serious games analytics is what we refer to as alignment. In serious games, alignment is the idea that the rewards of your game correspond to the goals you have in mind for it. One example of potential misalignment is *DragonBox*, a game designed to teach students algebra, where in a recent study it was found that while students succeed in the game, they do not improve on pre-posttests of algebra outside of the game (Long & Aleven, 2014). In dealing with misalignment, designers could benefit from having more ways of knowing whether or not their games are well-aligned and where potential sources of misalignment are. The field of serious games analytics is well-suited to fill this gap.

Common approaches in serious games only tangentially address issues of alignment. Techniques such as pre-posttests and AB testing can only provide designers insight into whether games are aligned or which of the tested features are better for alignment. These approaches require an explicit experimental design and the foresight of what features might be worth varying. While useful for advancing the science of serious games, these approaches are less helpful for informing design because they tend to come at the end of the game design process, rather than during the process. Serious games analytics would benefit from further focus on the formative context and the development of methods for identifying variables that would be worth AB testing.

Knowing if a game design is aligned is particularly difficult in open-ended games, whose structure provides a wide variety of actions for players to take. These games can easily become misaligned because there are more chances for the game to present conflicting feedback when players have more freedom of action. Providing players with freedom is a central design goal of using an open-ended structure and so we must rely on serious games analytics to create new methods that allow designers of open-ended serious games to understand the alignment of their game without compromising player freedom.

A difficulty in working with open-ended games is understanding players' actions in terms of the context in which they took place. To address this, we find potential in the use of replay systems. Replay systems have become a common idiom in game design, with many commercial games providing access to state-based replay files that can be analyzed by researchers (Weber & Mateas, 2009). Additionally, many common game engines provide facilities for quick recreation of game entities and whole states with little programming effort (Harpstead, Myers, & Aleven, 2013). By having access to a full replay of player actions, all actions can be considered within the context which they took place.

In this chapter, we will explore more about what we mean for a game to be open-ended and the kinds of alignment challenges open-ended games run into.

Then we discuss replay analysis, a technique we have applied to help us in analyzing open-ended games. Finally, we walk through a case study of how we apply this technique to the open-ended game *RumbleBlocks* (Christel et al., 2012).

2 Open-Ended Serious Games

We begin by defining what an open-ended serious game is and explore the challenges that open-ended serious games pose for design and serious game analytics. First we explore how other researchers have conceptualized this space before presenting our own definition of openness. Finally, we recount the prior serious games analytics work that is relevant to the open-ended serious games space.

Squire is one of the first researchers to use the term “open-ended games” in the context of serious games and to highlight their strong potential for serious applications (Squire, 2008). In his analysis he makes a distinction between two game genres that have potential for serious games. The first is persistent worlds, which have a large, socially shared, explorable world. Two widely known games in this category are *World of Warcraft* and *Everquest*. Squire mentions that some serious games have been designed to inhabit this space, notably *Quest Atlantis* (Barab, Thomas, Dodge, Carteaux, & Tuzun, 2005). The second genre that he discusses is open-ended simulation games, or sandbox games. Critical to this second category is the feature that there is no single correct pathway through the game to an end goal. Squire gives the examples of *Civilization III* and *Grand Theft Auto*: *San Andreas*. Both games contain large, non-social, spaces which the player can explore. Further, part of learning and playing in these games is to gain an understanding of the game’s overarching system in order to use it to the player’s advantage. It is this systems focus, Squire claims, that makes open-ended games attractive for serious applications.

Spring and Pellegrino have also discussed the idea of open games (Spring & Pellegrino, 2011). Influenced by Squire, Spring and Pellegrino also cite the importance of players coming to an understanding of a game’s system in order to succeed as a crucial component of why open games are attractive for serious applications. They present two insights regarding open games: they lend themselves primarily to conceptual knowledge and they allow students to learn through failure. These insights imply that while open games may be primarily sandbox-like experiences, they are not devoid of designer-implemented feedback. Squire argued that you cannot have a targeted open-ended game, but Spring and Pellegrino suggest that targeted feedback is still a component of open serious games.

While never explicitly discussing the openness of games, some of Gee’s learning principles apply to open-ended serious games (Gee, 2003). Paralleling Squire, Gee highlights the importance of having multiple solution paths through a game’s space so that students can learn through exploration. Similar to Spring and Pellegrino, Gee argues that players learn from the practice of probing open systems and receiving feedback, what he calls the probing principle.

In order to understand the number of possible solution paths, we can turn to Shell's concept of functional game spaces (Schell, 2008). In Schell's framing, a functional game space is the space in which a game *really* takes place rather than the physical or virtual space that might be seen by the player. For example, while the game of *Monopoly* is physically played on a two-dimensional board, *Monopoly*'s functional game space is a one-dimensional loop. The location of properties on the two-dimensional board has no meaning beyond conveying property adjacency in the one-dimensional loop. From this perspective, the space a player works in might be much larger than the functional space of a game. However, the complexity of the functional space is really what determines whether a game is open-ended.

Combining these prior conceptions, we define open-ended games as those with complex functional spaces containing many solution paths. In one regard, this definition is less inclusive than those proposed by prior researchers; i.e., games with limited functional spaces are less open-ended. Persistent worlds, such as *River City* (Ketelhut, 2006), are often thought of as open-ended because players are allowed to explore a large virtual world; however, from a functional space perspective, these virtual worlds are composed of a few relevant non-player character interactions and events spread through a vast space that contains little game meaning. In another regard, our definition is more inclusive than those proposed by others. For example, our definition includes games such as *Refraction* (Andersen, Gulwani, & Popovic, 2013), which have many functional solution paths, but do not allow for free exploration, which is a key feature of open-ended games according to prior conceptualizations. While others' definitions are focused heavily on the sandbox nature of open-ended serious games because of their pedagogical affordances, our definition centers more on the challenges that a games structure imposes on serious games analytics.

Taking the perspective that open-ended serious games are characterized by large functional solution spaces, the issue of alignment becomes a key challenge. As the complexity of the solution space increases, it becomes more important for designers to understand how their game is behaving and giving feedback in all portions of the space. As serious games are meant to guide players toward a set of learning objectives, designers need more insight into how this is working in practice. Are students meeting the objectives? What parts of the design are out of alignment with the objectives? We view providing such insight as a grand challenge for serious game analytics.

2.1 Prior Work in Open-Ended Serious Game Analytics

The work on Playtracer by Andersen, Liu, and colleagues is relevant to open-ended serious games (Andersen et al., 2013; Liu, Andersen, & Snider, 2011). This approach takes player log traces and aggregates them into a state graph for each game level, allowing designers to see the different ways that players move through the space of their game. The original Playtracer method aggregated states that were exactly equal but further work examined an alternative based on common game-relevant features that allowed for better aggregation. Feature-based projection approach was an improvement, but it still had trouble on very large continuous spaces such as the one present in *Foldit* (Liu et al., 2011). In terms of alignment, these analyses have benefits for helping designers see the functional solution space of their game; however, they do not provide strong guidance on how to explore that space while considering whether the game accomplishes design goals.

Another relevant body of work looks at using procedural content generation to create level designs in open-ended games that verifiably have no short cut solutions (Smith, Andersen, Mateas, & Popović, 2012; Smith, Butler, & Popović, 2013). This work looks at the game *Refraction* and explores framing the level design process as an answer set programming task. This approach allows designers to specify a series of constraints that all levels must hold and then asks a procedural content generation system to create level designs that are known to hold to those constraints. This attacks the issue of alignment by trying to make misalignment impossible. While this work is a very promising approach to the issue of alignment in educational games, it has limitations in that it requires designers to adopt a very particular formalization of their game design. It also has the limitation that the constraints the designer wants to hold must be known in advance for the system to work, a limitation shared by common AB testing paradigms (Lomas, Patel, Forlizzi, & Koedinger, 2013).

Spring and Pellegrino used an approach of counting the number of positive and negative plant interactions players used in the game *Hortus* and used the pattern of how this usage changed to infer player learning (Spring & Pellegrino, 2011). In a similar vein, Shute and colleagues used a classifier in the game *Newton*'*s Playground* to determine the kinds of simple machines players were using to solve physics puzzles. They found that players that had higher pretest scores also had higher usage of simple machines, suggesting the game is likely well-aligned. We would describe both of these approaches as traditional metric styles or feature counting as both approaches use a system in the game itself to report the presence or count of a feature. These approaches are afforded by game logging libraries, like the ADAGE system (Owen & Halverson, 2013). These systems provide the ability to record standard game metric telemetry, which can provide insight into the particular experience of an individual playing. However, these systems require knowing what you want to log in advance, which can be problematic.

3 Replay Analysis

Replay systems are becoming a more common design element in games, with many commercial games providing access to player replay files (Weber & Mateas, 2009). By replay, we are referring to a system included within a game that re-enacts player actions, recorded in a transaction log, so as to reproduce a player's session. This is distinct from a video replay in that the analyst has full access to running game code, opening up many avenues for analysis that would not be possible from a video recording or simple metrics logging.

We have developed an approach for using a replay system to aid in serious game analytics research, which we refer to as Replay Analysis. Replay Analysis is an approach to serious game analytics that tries to address some of the challenges of open-ended game design by logging repayable traces of players' sessions rather than a predefined set of metrics. Using this approach, player performance can be considered in the context in which it took place. Additionally, it allows researchers and analysts to prototype new measures of learning without having to commit early, smoothing the design process. In this section, we briefly describe the components involved in performing a replay analysis; for more implementation details on a replay analysis library that we developed for the Unity game engine see: (Harpstead, Myers & Aleven, 2013).

The approach entails both a particular schema for logging data and a system to replay logs through the game engine. Logs of student behavior are captured at the level of a basic action, defined as the smallest unit of meaningful action that a player can exert on the game world, what Schell would call operant actions (Schell, 2008). These actions are meant to be contextualized to the game world, e.g., picking up or dropping an object, rather than the raw input of the player, e.g., mouse down at position (*x*, *y*). The reason for recording the smallest possible actions is to allow for analysis at various grain sizes by capturing data at the finest grain size. Additionally, it is easy for game developers to see where logging calls need to be inserted into game code because the basic actions make up the base mechanics of the game. Each action is also paired with a description of the state of the game just before the action took place. This allows analyses to consider each action within the context in which it took place and to know the initial conditions of an action that may take time to broadly affect the game environment. Having paired states with each action also allows for logs to be replayed accurately without having to interpolate prior actions.

The second major component of the approach is a system for replaying actions, which we refer to as a Replay Analysis Engine (RAE). The RAE reads in a player's log file and reconstructs the player's play session action-by-action. For each action, the RAE constructs the state in which the action took place and then enacts the player action to let the game engine resolve the consequences of that action, using the same code that would normally handle such an action. Analyses can then be performed by running calculations on the results, with full access to any state attributes that would have been present at play time. These analyses represent an accurate reproduction of the player's own experience because the re-instantiated state is composed of exactly the same game elements, in terms of code.

One of the major benefits of the replay approach is that analyses are free to evolve along with the questions of the design team because no potential data is lost in logging. This paradigm also allows game design to proceed without having to agree on set of metrics to capture beforehand, reducing friction between designers and analysts during the development process and allowing analysts to explore multiple candidate metrics without having to commit to one.

The full benefit of this approach is felt most strongly in the context of classroom playtests. Because of the administrative and logistical processes involved in securing large populations of students in a classroom setting, such playtests are comparatively rare making the data captured in them all the more valuable. The high signal fidelity of replay logs ensures that datasets captured through classroom playtests throughout the design process can continue to benefit analysis and iteration.

4 RumbleBlocks

To demonstrate replay analysis, we describe a number of analyses that were facilitated by this technique on a game called *RumbleBlocks* (Christel et al., 2012). *RumbleBlocks* is an educational game designed to teach basic concepts of structural stability and balance to children in grades K-3 (ages 5–8 years old). The primary educational goals are for players to gain an understanding of three main principles of stability: objects with wider bases are more stable, objects that are symmetrical are more stable, and objects with lower centers of mass are more stable. These principles are derived from goals outlined in the National Research Council's Framework for New K-12 Science Education Standards (National Research Council, 2012), and other science education curricula for the target age group.

The game follows a sci-fi narrative where players help a group of stranded aliens on a number of foreign planets. Each level starts with the player finding an alien stranded on a cliff and a deactivated spaceship left off to the side of the world (see Fig. 17.1). The player's goal is to build a tower out of blocks that is tall enough

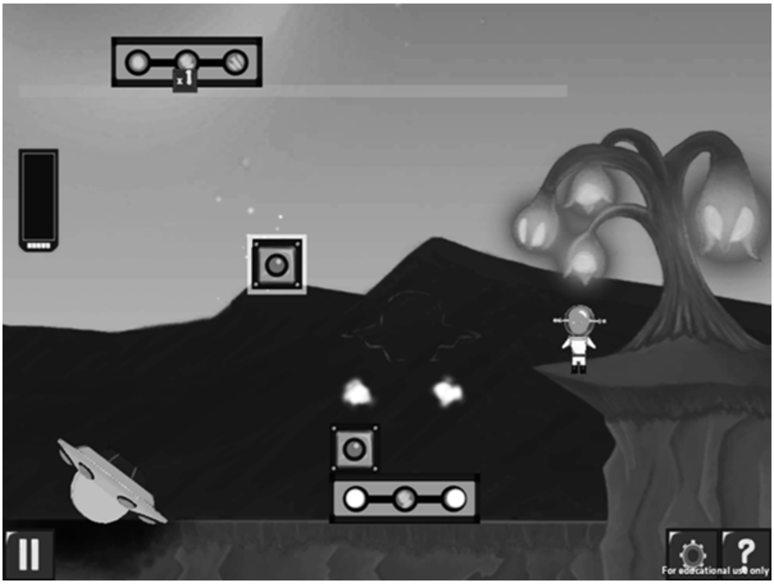


Fig. 17.1 A screenshot of *RumbleBlocks*

to reach the alien so that they can give the alien's ship back. In the process, they must also cover a series of energy balls with their tower, which are captured in orbs on the blocks and provide the ship with power. Once the player has placed the ship on top of the tower, the ship powers up triggering an earthquake; if the earthquake topples the tower, or knocks the ship off the top, then the player must restart the level; however, if the tower remains standing, with the ship on top, then the player succeeds and moves on to the next level.

We consider *RumbleBlocks* to be an open-ended game because, while its design may appear simple, it actually possesses a large functional solution space. Players are allowed to place blocks in free space leading to effectively infinitely many possible solutions. The space could alternatively be considered as a grid, based on the size of the cube and implied by the energy dot placement, but even then there are many situations where there is a combinatorial number of possible solution configurations. Another element worth pointing out is that the designers of *RumbleBlocks* had no a priori knowledge of how many solutions can successfully satisfy the constraints of each level. While it may have been possible to enumerate all possible solutions, in a method similar to (Smith et al., 2012) such analysis was beyond the resources of *RumbleBlocks*' design team.

RumbleBlocks was outfitted to use our replay logging framework and replay analysis engine (RAE) to log player actions (Harpstead, Myers & Aleven, 2013). The basic actions involved in the game correspond to players' actions with the blocks (i.e., picked up, rotated, or placed). The states being captured include a description of the position, orientation, and velocity of every block in the world as well as the spaceship. Logs can then be played back through the RAE and produce metrics, or other data encodings, as desired by our evolving analyses.

5 Analyses

Our analysis of RumbleBlocks progressed through a series of investigations, each facilitated by a different interpretation of players' log data. The data we discuss here were gathered as part of in-class playtests done with 174 students in two Pittsburgh area public schools, with players taken from the target demographic (ages 5–8). The goals of this playtest were to evaluate the current state of the game's design with a large group of players and to attempt to ground measures of learning within the game against out-of-game assessments of the game's educational objectives. Testing took place over four sessions: an external pretest, two 40 min sessions of play, and an external posttest.

Two sets of levels were selected to be used as in-game pre- and posttests counterbalanced across players. These levels were chosen out of the normal pool of levels, but were altered to remove the energy ball mechanic and to prevent players from retrying after a failed attempt. These special levels were placed after a short collection of tutorial levels, which explained the basic mechanics of the game, and at the end of the game. This design allows us to get a sense of how players built before and after they had experience with the game. In addition to the in-game evaluations, players also took out-of-game paper and pencil tests, before and after playing the game. These tests contained items relating to stability and construction, based on the three principles of base width, low center of mass, and symmetry.

5.1 Learning Results

The first question we sought to answer about *RumbleBlocks* was whether or not players were improving at the game's target concepts of structural stability and balance after playing the game. Turning first to the out-of-game tests, using a paired-samples *t*-test we measured a slight, yet significant, increase in performance from pretest to posttest, $t(173)=-2.13$, $p=.03$, $d=.16$. In looking at the difference in raw performance, i.e., the pass rate of the in-game pretest and posttest levels, a paired-samples *t*-test showed that there is a significant, medium-sized improvement in the passing rate from pre to post, $t(173)=-4.96$, $p<.001$, $d=.51$. While these results are encouraging, they can only tell us that players are getting better; they do not give us a sense of what they are getting better at.

We wanted to explore this question further by using the RAE and in-game log data to see if players were actually getting better at the specific principles targeted by the game. This would mean that from pre to post, players would build towers that showed a better awareness that (1) a structure with a wide base is more stable, (2) a structure with a lower center of mass is more stable, and (3) a structure that is symmetrical is more stable. It is important to note that looking at a difference in metrics related to learning goals is different from looking at the difference in player success rate. If we entertain the possibility that the game is not necessarily well-aligned, then it is possible that players could improve in their pass rate in the game but not in metrics related to the game's goals.

To find out whether players are learning the physics principles targeted by the game, we instrumented the RAE to read logs from the in-game pre-post levels and calculate a variety of metrics based on each player's final state of each level. We refer to these metrics as Principle-Relevant Metrics (PRMs). These metrics were: the width of the tower's base (Fig. 17.2a), the height of the tower's center of mass relative to the ground (Fig. 17.2b), and a measure of symmetry defined as the angle formed by a ray from the center of the base to the center of mass and 90° (Fig. 17.2c). These measures were then compared to values calculated across all other players for that same level in order to create a relative score for each player. This was done to account for the nuanced difference in level design, making it difficult to compare metrics across levels. In the case of base width, the relative score was calculated relative to the maximum observed width for that level; for center of mass position, the score is relative to the lowest observed position for that level; and for symmetry a score is already relative to 90°, which would represent perfect symmetry.

To see if there was any improvement on the use of principles between the pre- and posttests, we compared the related pre and posttest metrics for each student (averaged across the levels of the pre and posttest) using a paired-samples *t*-test. Looking at the results in Table 17.1, we saw a significant improvement for the base

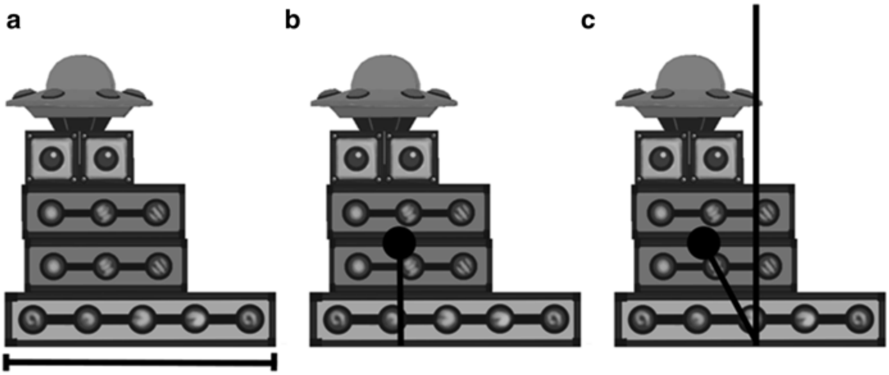


Fig. 17.2 A visual depiction of each of the three Principle-Relevant Metrics used in analysis. (**a**) Base width, (**b**) center of mass height, and (**c**) symmetry angle

Table 17.1 *t*-Test results for average Principle-Relevant Metrics from pretest to posttest

	Pretest		Posttest				
Metric	*M*	SD	*M*	SD	*t*(173)	*p*	*d*
Base width	.60	.01	.64	.01	−2.77	.006	.30
Center of mass height	1.61	.02	1.63	.02	−.66	.501	.08
Symmetry angle	5.98	.34	5.20	.27	1.98	.050	.19

width and symmetry metrics, meaning that at the end of playing the game, students were beginning to design towers that had wider bases and more symmetrical layouts. However, we did not see any significant difference in terms of center of mass height, meaning that students did not seem to attempt to lower the center of mass of their structures. This result would suggest that the current version of the game may possess a misalignment in how it handles the low center of mass principle.

5.2 *Metric Alignment Analysis*

Knowing from the pre-post level analysis that there were likely some design issues with *RumbleBlocks*, our next goal became understanding what those design issues were and how they might be addressed. The next step in analysis was to determine if the game was properly incentivizing players to act in a way that corresponds to the goals for the game. If the game is knocking over towers that are well-designed or letting poorly designed towers remain standing, players will not know what to make of the feedback they are given and improve toward better understanding. Such cases would be examples of misalignment.

To facilitate this analysis, we augmented the RAE to calculate the same PRMs from the pre-post analysis, except this time to do it for all levels. We wanted to

Table 17.2 The results of a logistic regression of success of solution on Principle-Relevant Metrics for levels targeting each of the three principles

Group	Coefficient	*B*	*SE B*	*β*	*p*
Symmetry levels (*df* = 1,788)	(Intercept)	1.044	.061	17.250	<.001
	Base width	.449	.054	8.368	<.001
	Center of mass height	.418	.089	4.700	<.001
	Symmetry angle	−.205	.069	−2.969	.003
Center of mass levels (*df* = 2,107)	(Intercept)	1.379	.063	22.042	<.001
	Base width	.022	.066	.326	.745
	Center of mass height	−.046	.047	−.975	.330
	Symmetry angle	−.165	.043	−3.803	<.001
Wide base levels (*df* = 1,997)	(Intercept)	1.729	.074	23.463	<.001
	Base width	.221	.069	3.229	.001
	Center of mass height	−.113	.097	−1.164	.245
	Symmetry angle	.011	.078	.135	.893

explore how well metrics that should indicate a well-constructed tower actually corresponded to a player passing a given level. To do this, we created three groups by collecting together all student solutions to levels that target each of the three principles, i.e., all levels targeting the wide base principle together, all levels targeting the low center of mass principles together, and all levels targeting the symmetry principle together. We performed logistic regressions using each of the metrics of the players' towers to predict success on a level, with all metrics normalized to mean 0 and standard deviation 1 to aid in interpretation of the relative strengths of their coefficient. What we would expect from this analysis is that the principle which is targeted by a level has a strong predictive relationship with success on that level. It is important to note that this analysis is concerned primarily with the behavior of the game and not with student performance. In this context, students are merely providing the test data for our analysis of the game's system.

The results of the regression analyses can be found in Table 17.2. When looking at the PRMs for base width and symmetry, there is a significant relationship between the PRM and success on the level, which is what would be expected. The relationship for the center of mass PRM, however, was not found to be significant. This would mean that, counter to what the target principles suggest, players that build with lower centers of mass are not any more likely to succeed on levels that target the center of mass principle than those who do not. This could not have been the *RumbleBlocks* designers' intent.

5.3 Solution Clustering Analysis

The logistic regression analysis agrees with the previous findings that players do not seem to be improving at the center of mass principle because they are not being given consistent feedback in terms of the principle. The next question that arises out of this is: if players are not getting consistent feedback on the center of mass principle, then what are they doing? Answering this question requires a different view of the players' performance.

Rather than looking at how well players' creations do in terms of the games' rules or the pedagogical principles, we wanted to get a more qualitative sense for what players were creating and how different features of these creations affected in-game performance (as measured by success on each level—does the tower stand or fall?). We hoped that analyzing properties of individual student solutions might provide insights into possible new design directions. However, there were over 6,000 student solutions and combing through them all by hand would be impractical. To make understanding players' patterns of play more manageable, we wanted to cluster the solutions into groups that essentially embody the same solution. This way our analysis could proceed at the group level, which has many advantages. For example, we could determine whether students were creating the solutions originally envisioned by the designer. Furthermore, we could determine the specific structural features of a solution that contribute to its success or failure. This allowed us to explore how certain mechanics of the game affect the kinds of solutions that get rewarded, subsequently affecting what students do and learn. This type of analysis is useful for reflecting on educational goals and game mechanics on subsequent design iterations.

To perform these analyses, we first had to convert the solutions into a representation that captured their essential structural features. For example, many students might build a tower that uses an arch pattern, whereas others might build an inverted "T" shape. We wanted a representation that captured elements of these basic structural patterns. To build this new representation, we first instrumented the RAE to produce representations of student towers aligned to a two-dimensional grid. This process makes use of a number of capabilities exposed by the game engine in the RAE such as collider information and individual block dimensions.

Next, we employed two-dimensional grammar induction to learn a set of patterns that can be used to describe all of the student solutions (Harpstead et al., 2013). A two-dimensional grammar consists of three components: terminal symbols, which represent the blocks, spaceship, and space (in this context); non-terminal symbols, which represent structural patterns; and, rules which map non-terminal symbols to pairs of other non-terminal symbols oriented in a certain direction (horizontal or vertical). Rules can also map non-terminals to terminal symbols (a unary relationship). Figure 17.3 shows an example grammar (u, h, and v represent unary, horizontal, and vertical, respectively) and the parses for two different towers. To learn a grammar, we generated an exhaustive set of rules that describe every possible way to parse all of the solutions. We then computed all the possible parses of each solution. Given the parses for each solution, we then created a vector for each solution which contained a 1 for every non-terminal present in the solution and a 0 for every non-terminal not present in the solution. The resulting feature vector contains information about all of the structural patterns present in each solution.

For each level, we clustered the featurized solutions using *g*-means, a variant of the *k*-means algorithm that chooses a value for *k* optimizing for a Gaussian distribution

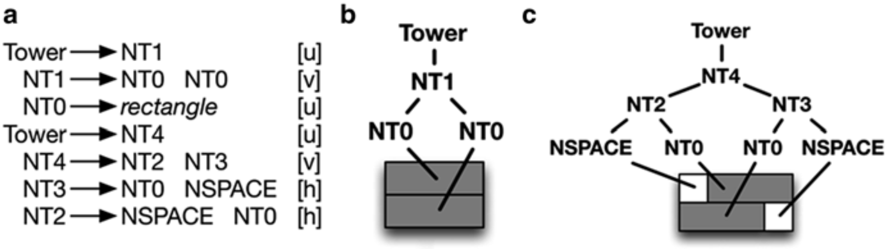


Fig. 17.3 A two-dimensional grammar (**a**) and the parse trees generated by applying this grammar to two solutions (**b**, **c**)

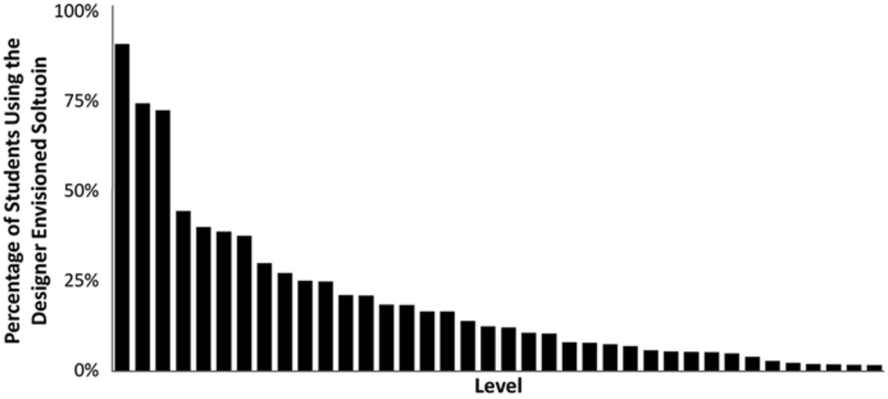


Fig. 17.4 The percentage of students who used the solution envisioned by the designers on each level (sorted by percentage)

within clusters (Hamerly & Elkan, 2004). This produced a set of different groups for each level, where each group represents solutions that share structural similarity. The resultant clusters allow us to get a picture of what the solution space to the game looks like and begin to tease apart the different patterns.

As a first pass we wanted to see how often players conformed to the designers' expectations for a level. When designing levels, game designers often have a particular solution in mind, even if they intend the level to allow for a number of different solutions. We had a member of the game design team create a player trace that represented the designer-envisioned solutions to each of the levels. We compared this trace to the clusters of other player solutions to identify which cluster each envisioned solution belonged to. Finally, we looked at how many players fell into the envisioned cluster as opposed to any other cluster. This analysis resulted in the pattern visible in Fig. 17.4. A number of levels have a high degree of agreement with designers; these levels are mostly simple early levels or tutorial levels. Other levels have very little agreement with the designers' vision; these tend to be later levels that have a higher level of complexity.

The pattern shown in Fig. 17.4 helps to illustrate how open-ended *RumbleBlocks* can be, as the levels in the tail of the graph have a wider space of possibility than the designers may have envisioned themselves. It should be noted that this analysis does not necessarily result in a value judgment of the game, as players, particularly in an open-ended game, are likely to diverge from the designers' perspective. Instead, it helps designers to understand where their own intuitions are likely to be the weakest.

Another way of using this clustering formulation is to explore the alignment of individual clusters, allowing designers to have a picture of specific examples of misalignment (Harpstead, MacLellan, Aleven, & Myers, 2014). To do this we looked at the metrics, calculated for the previous regression analysis, within each cluster and looked at the patterns between those metrics and success on levels. For each cluster, we created a representative solution by averaging the PRM scores within the cluster and assuming the majority success value. This gives us the ability to think about common patterns of solutions through a single representative solution rather than individual solutions. We then examined levels by looking at plots such as Fig. 17.5. These plots show each representative solution plotted with its frequency percentage along the *x*-axis and its relative principled-ness, in terms of a normalized PRM score, along the *y*-axis. The lighter squares represent solutions that are mostly successful where the darker diamonds show solutions that are mostly unsuccessful.

When examining plots like Fig. 17.5, two different patterns are primarily of interest: principled failures and unprincipled successes, which both represent the game generally giving feedback contrary to the target principle for the level. When visually inspecting these misaligned cases across levels, a pattern started to emerge; failure seemed related to towers with small platforms for the spaceship to sit on top, particularly in levels targeting the low center of mass principle. Such a pattern

Fig. 17.5 A plot of representative solutions' PRM score versus frequency

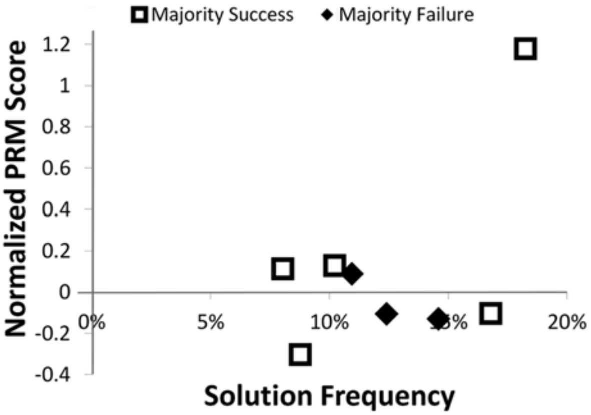


Table 17.3 χ^2 results of students' solutions' structural features against success, with Bonferroni-corrected p-values and logistic regression direction

Symbol	χ^2	P	Direction
NT 1,633	20.95	0.028	–
NT 319[a]	21.00	0.028	–
NT 1,083[a]	22.10	0.016	–
NT 154[a]	26.27	0.002	–
NT 2,458[a]	26.65	0.001	–
NT 1,166[a]	30.22	<.001	–
NT 1,537[a]	30.25	<.001	–
NT 284	30.43	<.001	–
NT 329[b]	33.35	<.001	–
NT 10,897[a]	35.83	<.001	–
NT 37[a]	37.07	<.001	–
NT 13	38.55	<.001	+
NT 1,538[a]	39.18	<.001	–
NT 40[a]	39.53	<.001	–
NT 11,914[a]	39.55	<.001	–
NT 150[a]	39.55	<.001	–
NT 1,541[a]	68.50	<.001	–
NT 44[b]	71.13	<.001	–
NT 151[a]	73.11	<.001	–

Note. [a]Symbol visually similar to NT37
[b]Symbol visually similar to NT44

is likely to lead to the spaceship falling off the tower, regardless of how well the design of the rest of the tower adheres to the target principles. While there is a component of stability, we were concerned that the secondary success criterion of the spaceship staying on the tower may be overshadowing the rest of the features of a tower, leading students to focus on a small element of their design rather than the design as whole.

To test this hypothesis, we turned again to RAE and the features that were produced for clustering. These symbols represent binary relationships of different block types and so encode different sub-structural patterns present in player towers. We performed a χ^2 test of each of the 6,010 symbols against solution success to see which patterns were most strongly related to success of a tower. We applied a Bonferroni correction to the χ^2 results to account for the number of statistical tests (6,010). Results for the χ^2 analysis can be found in Table 17.3, and visually rendered in Fig. 17.6. Overall we found 19 symbols that were significantly related to success, after correction. Due to idiosyncrasies in our grammar learning algorithm, it happens that 14 of these symbols would ground out to be the same, NT37 in Fig. 17.6, and two of the symbols, NT329 and NT44, share a similar relationship. This happens because the grammar learning process learns multiple rules to represent adjacent empty space and we have omitted these for clarity. Once we had a selection of

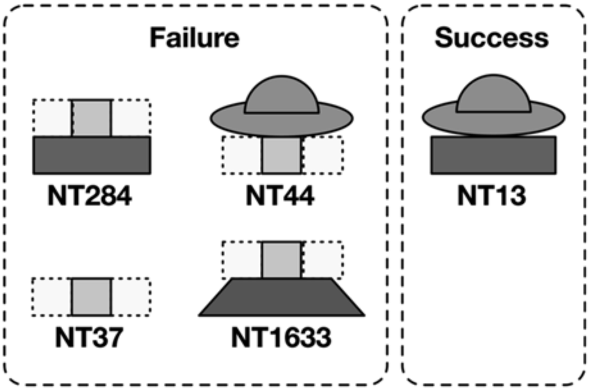


Fig. 17.6 Visual renderings of the symbols from Table 17.3 found to be significantly related to success and failure

significant features, we performed a logistic regression of those 19 features against solution success to understand the direction of the relationship, i.e., does each feature predict success or failure. We only present the directional results of this regression and not the actual coefficients.

The pattern that arises from the χ^2 analysis confirms that towers containing the spaceship on top of a lone square, or substructures containing squares without other supports, are more likely to fail than towers containing the spaceship on top of a wide platform. This analysis demonstrates that the spaceship remaining on top of the tower represents a secondary success criterion. While the designers were aware that the spaceship served such a purpose in the design, they did not think it would be such a strong determining factor to the potential detriment of other learning goals. When pursuing iteration, the designers of *RumbleBlocks* will have to consider if this result represents a flaw in the game's mechanics, which contradicts the message, or an opportunity to teach a nuanced aspect of stability and balance with some new feedback.

6 Discussion

In this chapter, we have demonstrated a number of analyses of the design of the game *RumbleBlocks*, all of which were facilitated by the use of replay analysis. Each of these analyses has contributed to an evolving understanding of the design of *RumbleBlocks* and how well it accomplishes its stated goals. As work on the game moves forward, similar assessments will be made of future versions to ensure existing design problems have been properly addressed.

While we have described a particular case study here, we believe that the issue of alignment is an important one for serious game analytics to consider more broadly. Players approach games from different perspectives than their designers (Hunicke, Leblanc, & Zubek, 2004). When designing games for serious purposes, it is important that designers be equipped with tools and techniques that help them understand whether their goals are being met. We have presented replay analysis as one means of filling, this need but we view this as just a first step into what we expect will be a larger space within serious game analytics.

Another use of replay analysis, which we have not discussed here, is to use replayed sessions as a means of exploring the implications of alternative designs. For example, the designers of *RumbleBlocks* might want to entertain the idea of having the spaceship stick to the top block of a tower to reduce its effect as a secondary success criterion. The RAE could be instrumented to recreate players' final towers with this system in place and run the earthquake to see how the game reacts differently. While this particular use of replay can provide interesting perspectives, it has limitations. For example, it could be argued that players would have played differently had certain rule changes been in effect. Additionally, exploring alternatives with log files from different iterations of the game requires more attention to detail in version control to ensure replayed results reflect the right version of the game.

Another issue with the use of a replay paradigm is that it produces comparatively large log files. The entire *RumbleBlocks* dataset that we have discussed here (containing approximately 80 min of play from 174 players) is roughly 850 MB in size. This does not include the residual files output by the RAE for use in statistical analysis or clustering. Additionally, while the replay process can be sped up, it does require some amount of real time to re-simulate the game environment. When taken to the scale seen in some other serious games work that involves millions of players over the course of months, this could quickly become intractable (Andersen et al., 2012; Lomas et al., 2013). A possible way of addressing this could be to only record replay fidelity logs for a subset of the overall player population. Another approach might be to use replay logs in the earlier stages of design, before a final set of metrics has been decided, as a way of prototyping measures of learning or other desirable outcomes.

7 Conclusion

The increase in the number of open-ended serious games is an exciting trend in the field of serious game analytics. While this trend is encouraging, it is important to keep in mind how the open-ended nature of some games can undermine their intended serious purposes. Replay analysis is just one of what we hope will become many analytics techniques for studying how players are navigating open-ended domains and interacting with content. We hope that other researchers can find utility in our approach and look forward to what new challenges the future of the field brings.

Acknowledgments This work was supported in part by a Graduate Training Grant awarded to Carnegie Mellon University by the Department of Education #R305B090023 and by the DARPA ENGAGE research program under ONR Contract Number N00014-12-C-0284. All opinions expressed in this article are those of the authors and do not necessarily reflect the position of the sponsoring agency.

References

Andersen, E., Gulwani, S., & Popovic, Z. (2013). A trace-based framework for analyzing and synthesizing educational progressions. In *Proceedings of the 31st International ACM SIGCHI Conference on Human Factors in Computing Systems—CHI '13* (pp. 773–782). New York: ACM Press. doi:10.1145/2470654.2470764

Andersen, E., O'Rourke, E., Liu, Y., Snider, R., Lowdermilk, J., Truong, D., et al. (2012). The impact of tutorials on games of varying complexity. In *Proceedings of the 30th International ACM SIGCHI Conference on Human Factors in Computing Systems—CHI '12* (pp. 59–68). doi:10.1145/2207676.2207687

Barab, S., Thomas, M., Dodge, T., Carteaux, R., & Tuzun, H. (2005). Making learning fun: Quest Atlantis, a game without guns. *Educational Technology Research and Development*. doi:10.1007/BF02504859.

Christel, M. G., Stevens, S. M., Maher, B. S., Brice, S., Champer, M., Jayapalan, L., et al. (2012). RumbleBlocks: Teaching science concepts to young children through a unity game. In *Proceedings of CGAMES'2012 USA—17th International Conference on Computer Games: AI, Animation, Mobile, Interactive Multimedia, Educational and Serious Games* (pp. 162–166). IEEE. doi:10.1109/CGames.2012.6314570

Clark, D. B., Tanner-Smith, E. E., Killingsworth, S., & Bellamy, S. (2013). *Digital games for learning: A systematic review and meta-analysis (executive summary)*. Menlo Park, CA: SRI International.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy. Computers in entertainment*. New York: Palgrave Macmillan. doi:10.1145/950566.950595.

Hamerly, G., & Elkan, C. (2004). Learning the k in k-means. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference* (pp. 281–288). Cambridge, MA: MIT Press.

Harpstead, E., MacLellan, C. J., Aleven, V., & Myers, B. A. (2014). Using extracted features to inform alignment-driven design ideas in an educational game. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems—CHI '14* (pp. 3329–3338). New York: ACM Press. doi:10.1145/2556288.2557393

Harpstead, E., Maclellan, C. J., Koedinger, K. R., Aleven, V., Dow, S. P., & Myers, B. A. (2013). Investigating the solution space of an open-ended educational game using conceptual feature extraction. In *Proceedings of the International Conference on Educational Data Mining—EDM '13* (pp. 51–58).

Harpstead, E., Myers, B., & Aleven, V. (2013). In search of learning: facilitating data analysis in educational games. In *Proceedings of the 31st International ACM SIGCHI Conference on Human Factors in Computing Systems—CHI '13* (pp. 79–88). Paris: ACM Press. doi:10.1145/2470654.2470667

Hunicke, R., Leblanc, M., & Zubek, R. (2004). MDA: A formal approach to game design and game research. In *Proceedings of the AAAI Workshop on Challenges in Game AI* (pp. 1–5).

Ketelhut, D. J. (2006). The impact of student self-efficacy on scientific inquiry skills: An exploratory investigation in river city, a multi-user virtual environment. *Journal of Science Education and Technology, 16*(1), 99–111. doi:10.1007/s10956-006-9038-y.

Liu, Y., Andersen, E., & Snider, R. (2011). Feature-based projections for effective playtrace analysis. In *Proceedings of the 6th International Conference on Foundations of Digital Games—FDG '11* (pp. 69–76). ACM Press. doi:10.1145/2159365.2159375

Lomas, D., Patel, K., Forlizzi, J. L., & Koedinger, K. R. (2013). Optimizing challenge in an educational game using large-scale design experiments. In *Proceedings of the 31st International ACM SIGCHI Conference on Human Factors in Computing Systems—CHI '13* (pp. 89–98). ACM Press. doi:10.1145/2470654.2470668

Long, Y., & Aleven, V. (2014). Gamification of joint student/system control over problem selection in a linear equation tutor. In *Proceedings of the 12th International Conference on Intelligent Tutoring Systems* (pp. 378–387). doi:10.1007/978-3-319-07221-0_47

National Research Council. (2012). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. In H. Quinn, H. Schweingruber, & T. Keller (Eds.), *Social sciences*. Washington, DC: The National Academies Press.

Owen, V. E., & Halverson, R. (2013). ADAGE (Assessment Data Aggregator for Game Environments): A click-stream data framework for assessment of learning in play. In *Proceedings of the 9th Games + Learning + Society Conference—GLS 9.0* (Vol. 9, pp. 248–254). ETC Press.

Schell, J. (2008). *The art of game design: A book of lenses* (1st ed.). Burlington, MA: Morgan Kaufmann.

Smith, A. M., Andersen, E., Mateas, M., & Popović, Z. (2012). A case study of expressively constrainable level design automation tools for a puzzle game. In *Proceedings of the 7th International Conference on Foundations of Digital Games—FDG '12* (p. 156). ACM Press. doi:10.1145/2282338.2282370

Smith, A. M., Butler, E., & Popović, Z. (2013). Quantifying over play: Constraining undesirable solutions in puzzle design. In *Proceedings of the 8th International Conference on Foundations of Digital Games—FDG '13* (pp. 221–228).

Spring, F., & Pellegrino, J. W. (2011). The challenge of assessing learning in open games: HORTUS as a case study. In *Proceedings of the 8th Games + Learning + Society Conference—GLS 8.0* (pp. 209–217).

Squire, K. (2008). Open-ended video games: A model for developing learning for the interactive age. In K. Salen (Ed.), *The ecology of games: Connecting youth, games, and learning* (pp. 167–198). Cambridge, MA: MIT Press. doi:10.1162/dmal.9780262693646. 167.

Weber, B. G., & Mateas, M. (2009). A data mining approach to strategy prediction. In *2009 IEEE Symposium on Computational Intelligence and Games* (pp. 140–147). IEEE. doi:10.1109/CIG.2009.5286483

Chapter 18
Using the Startle Eye-Blink to Measure Affect in Players

Keith Nesbitt, Karen Blackmore, Geoffrey Hookham, Frances Kay-Lambkin, and Peter Walla

Abstract The startle eye-blink is part of a non-voluntary response that typically occurs when an individual encounters a sudden and unexpected stimulus, such as a loud noise or increase in light. Modulations of the startle reflex can be used to infer affective processing in players. The response can be elicited using simple auditory, visual, electric, or mechanical stimuli. The magnitude of the startle eye-blink is used to infer the unconscious positive (pleasant) or negative (unpleasant) emotional state of the player. It is frequently used in psychology where variations in the magnitude, latency, and duration of the startle response are used to understand attention, workload, affective processing, and psychopathologies such as schizophrenia. By comparison, there has been limited use of this objective measure for studying games. As such, there are opportunities to adapt this measure to studies of player affect in the context of game design. We provide a review of the concepts of "affect" and "affective computing" as they relate to game design and also explain in detail the use of the startle eye-blink for objectively measuring player affect. Finally, the use of the approach is illustrated in a case study for evaluating a serious game design.

Keywords Affective processing • Emotion • Startle reflex • Startle eye-blink

K. Nesbitt (✉) • K. Blackmore • G. Hookham
University of Newcastle, University Drive, Callaghan, NSW 2308, Australia
e-mail: keith.nesbitt@newcastle.edu.au; karen.blackmore@newcastle.edu.au; geoffrey.hookham@newcastle.edu.au

F. Kay-Lambkin
University of New South Wales, 22-32 King Street, Randwick, NSW 2031, Australia
e-mail: f.kaylambkin@unsw.edu.au

P. Walla
Webster Vienna Private University, Palais Wenkeim, Praterstrasse 23, Vienna 1020, Austria
e-mail: peter.walla@webster.ac.at

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based Learning, DOI 10.1007/978-3-319-05834-4_18

1 Introduction

The startle response is a non-invasive measure of central nervous activity that typically occurs when an individual encounters a sudden, surprising change in their environment (Blumenthal et al., 2005). It is a reflex reaction that occurs without voluntary control and is characterised by protective body reactions, such as eye-blinks and stiffening of neck muscles. In evolutionary terms, it likely provides a defensive response to threatening stimuli and is associated with the fight or flight response. Elements of the startle response, such as the eye-blink, are modulated with an individual's emotional state; eye-blinks being larger when individuals are highly aroused and in unpleasant emotional states compared to blinks associated with low arousal and positive affect (Witvliet & Vrana, 1995).

The eye-blink component of the startle response is a reflex transmitted by the facial nerve that controls a number of facial muscles including those responsible for eye movement. Historically, the eye-blink reflex has been observed in studies since 1874 (Dawson, Schell, & Böhmelt, 1999). Variations in the amplitude, duration, onset (latency), and probability of the responses have been used to study a variety of psychological phenomena including attention (Filion, Dawson, & Schell, 1993), workload (Neumann, 2002), affective processing (Witvliet & Vrana, 1995), and psychopathologies such as schizophrenia (Swerdlow, Weber, Qu, Light, & Braff, 2008). Apart from variations in amplitude of the response, another common protocol involves determining how the startle response is inhibited when an additional stimuli, often referred to as a prepulse, is presented just prior to the startle stimulus (Swerdlow et al., 2008).

In terms of emotion, it is usually changes in the magnitude of the eye-blink that occur with negative or positive affect that are of interest (Witvliet & Vrana, 1995). Typically, affects are described in a two-dimensional model using arousal and valence (Lang, 1995). Arousal might range from sleepy and relaxed to excited or agitated. Valence, on the other hand, describes the pleasant or unpleasant aspect of an affect. For example, negative valence is generated under conditions that invoke fear or anger and are associated with stronger eye-blinks than those related to positive valence, such as those measured in happy or contented states. Negative and positive valence can be measured with the startle eye-blink and may be combined with other physiological measures of arousal, such as heart rate or skin conductance to classify more distinct emotional states within a two-dimensional model of affect (Witvliet & Vrana, 1995).

Importantly for serious game designers, both positive and negative valence have been strongly associated with positive and negative learning effects (Sabourin & Lester, 2014). For example, positive emotional states, such as engagement, joy, and happiness, lead to increased learning (Bless et al., 1996; Kanfer & Ackerman, 1989; Pekrun, Goetz, Titz, & Perry, 2002; Raghunathan & Trope, 2002). By contrast, negative experiences, such as frustration, anger, and boredom, lead to decreased effort, reduced motivation, and disengagement from learning activities (Meyer & Turner, 2006; Pekrun et al., 2002; Ramirez & Dockweiler, 1987; Sabourin, Rowe, Mott, & Lester, 2011). The startle eye-blink suggests itself as a measure that can be used in serious game design to evaluate the affect generated by gameplay. In simple terms, a positive affect should lead to better learning outcomes in serious games.

Although we believe the startle response holds much promise as a tool to support more objective evaluation of game design, much more work still needs to be done to apply this measure and understand its limitations. Therefore, at this stage, the use of the startle reflex measure for evaluating game design in terms of player affect needs to be approached carefully and tested in more studies. In this mood of cautious optimism, this chapter introduces information about how to use the startle response measure and summarises existing technical guidelines related to collecting, analysing and reporting results with the measure. To complement this review, we present in detail a case study where we have used the response to assist in the design of a serious game to assist in psychological counselling.

Understanding how player affect can be manipulated could impact directly the success of many serious games. Fortunately, the serious game community is not alone in considering the role of emotion in usability criteria such as effectiveness. Understanding, detecting, and responding to emotions and affective user responses are an issue at the forefront of the design of many modern computer systems. The cross-discipline field of study that interprets and simulates human emotions in terms of system design is known as "Affective Computing". Thus, we begin the chapter with a discussion of the concept of emotion, the importance of affect in interface design, and common approaches to performing affect detection.

2 Affective Computing

Affective computing concerns the practical development of computer systems that are able to detect and respond to human moods and emotions (Calvo & D'Mello, 2010). These systems might recognize the emotions of humans, respond by expressing an emotion in a way that a human can understand and, most ambitiously, even be able to "feel" in the way humans do (Picard, 1997).

Computer games likewise often have a design goal that includes manipulating human affect. For example, it may be desirable to produce an engaging game, dominated by positive affect that better supports learning or cognitive therapy. In a first-person horror game like *Slender*: *The Eight Pages* (Hadley, 2012), the intention may be to produce a negative affect such as fear, if that is the experience desired by the player and the intention of the designer (Coppins, 2014). Thus, a good question for any game designer is "What aspects of games make them enjoyable, addictive or engaging, and how do games, or their interactivity, elicit emotional involvement from players?" This area of enquiry involves an understanding of human emotions and emotional responses.

Typically, more subjective approaches are taken to assess player responses to design choices, yet the startle eye-blink provides a more objective possibility for evaluating these design choices and perhaps even adapting gameplay based on a player's recognising emotional state. As well as dynamic difficulty balancing, that is balancing gameplay difficulty with player ability, we might ideally provide dynamic mood balancing, whereby players' emotional states are balanced with game mechanics.

Unfortunately, the goal of recognising emotional state is extremely challenging as the concept of emotion is difficult to define let alone measure and even common moods, feelings, and attitudes vary significantly between individuals, both in how they are experienced and how they are expressed (Calvo & D'Mello, 2010). Indeed the mechanisms of emotion are still not agreed on and, in the next section we review some alternative theories of emotion and consider how they relate to the goal of detecting human affective states in computer games.

2.1 Emotional Theories

Emotions are biologically based action dispositions, theorised to be systematic responses that occur when a highly motivated action is delayed (Lang, Greenwald, Bradley, & Hamm, 1993). It has also been proposed that an emotion is the result of a novel circumstance preventing the completion of behaviour (Hebb, 1949). Essentially, emotions can be considered an involuntary response with primitive origins.

The behaviour of very primitive organisms can be categorised into two distinct categories: a direct approach to appetitive stimuli, and a withdrawal from noxious stimuli (Schneirla, 1959). It is theorised that humans follow the same two directives of behaviour, but elaborate acts, delays, and inhibitions have evolved to facilitate more complex, goal-directed paths to achieve withdrawal or approach (Lang, 1995). Thus, while involuntary and principally biphasic (pleasant or unpleasant), the expression of an emotion is mediated by higher level, goal-directed behaviours. This greatly complicates the measurement and even definition of emotion.

Early, more traditional emotion theories tend to focus on emotion as either a means of expression, a form of embodiment, a type of cognitive appraisal, or a social construct (Calvo & D'Mello, 2010). Not surprisingly, it was Darwin who first considered the evolutionary role of emotion in terms of behaviour (Darwin, 2002). Notably, emotions such as interest, joy, surprise, sadness, anger, discuss, contempt, fear, and shyness are considered to be universally recognised (Izard, 1994). As such, detection of these emotional states frequently underpins facial expression and body recognition systems that try to detect emotions.

In contrast, other traditional emotion theories would argue that emotion is more than just a form of expression, being also accompanied by a distinctly embodied physiological state (James, 1884). Assuming a typical physiological response to standard emotions like joy, anger, and fear implies that common patterns of physiological changes could be used to detect common emotional states. Indeed this assumption underlies the use of many objective systems based on detecting emotions using physiological measures.

While emotional expression is incredibly varied and complex, most theorists endorse an approach to emotion that features three components; “subjective feeling”, “expressive behavior”, and “physiological arousal” (Scherer, 1993). Additionally, some add “motivational state”, “action tendency”, and/or “cognitive processing” (Scherer, 1993). While these multiple emotional components are noted, simpler models to capture the motivational basis of emotion have evolved.

Physiological models usually consider the motivational basis of emotion using a very simple, two-factor model featuring affective valence and arousal (Lang, 1995). This dimensional theory of emotion holds that all emotions can be located on a two-dimensional space, as a function of valence and arousal (Ravaja, Saari, Salminen, Laarni, & Kallinen, 2006). In this two-dimensional model, valence represents a user’s emotional reaction to a stimulus, reflecting the degree to which it is a pleasant or unpleasant experience (positive and negative valence, respectively). Arousal indicates the level of activation associated with the experience, from very excited and energised, to sleepy, calm, and/or disinterested (Ravaja et al., 2006). This frequently used model typically uses the startle eye-blink to measure valence and other physiological indicators, such as heart rate or skin conductance to determine arousal.

Although valence and arousal provide the simplest and most commonly used model in affective computing, it has been argued that four dimensions are needed to satisfactorily represent similarities and differences in emotional experience (Fontaine, Scherer, Roesch & Ellsworth, 2007). These four dimensions are: valence, arousal, potency-control, and unpredictability (Fontaine et al., 2007). These four were identified based on the applicability of 144 features, representing six major components of emotion; appraisal of events, psychophysiological changes, motor expressions, action tendencies, subjective experiences, and emotion regulation (Fontaine et al., 2007).

A further, more cognitive-based approach considers emotions as something experienced in relation to the unconscious appraisal of an object or event (Scherer, Schorr, & Johnstone, 2001). This appraisal process may take into account a persons’ experience, their goals, and their ability to take action (Dalgleish, Dunn, & Mobbs, 2009). Cognitive approaches to understanding emotions have generally provided the basis of computational models of emotion used in agent-based systems (Reisenzein et al., 2013).

Considering the role that social interaction plays in the world of emotions means that the context of culture (Salovey, 2003) and society (Kemper, 1991) also impact the understanding of emotions. Calvo and D’Mello (2010) point out that this social construct view of emotions is somewhat underrepresented in the study of affective interface design.

More recently, the underlying neural circuitry of emotions has also come under study by neuroscientists, highlighting the complex overlap of emotion and cognition (Dalgleish et al., 2009), where emotion continually interacts with cognitive processes such as remembering, reasoning, goal setting, and planning. This work in neuroscience highlights that some emotional phenomenon may act below our normal level of consciousness and that emotions are states that emerge from the underlying complex system of underlying affective processes (Coan, 2010).

A recent, alternative model based on neuroscience emphasises a clear distinction between affective processing and emotion such that affective processing generates emotions (Walla & Panksepp, 2013). This model suggests that affective processing forms the neurophysiological basis for emotions, which are behavioural output and thus not a direct measure of processing itself. For example, behaviours such as facial expressions produced by facial muscle contractions are indicative of an emotion generated by underlying affective processing. Thus, if neural activity within affective processing circuits codes for unpleasant, the generated facial expression is negative. If, on the other hand, the neural activity codes for pleasant, the respective facial expression is positive. One consequence of this model is that affective processing can take place without necessarily generating an emotion in an individual. Another consequence is that a measurement approach such as the startle eye-blink records affective processing as distinct from an emotion that may be experienced and reported by a player.

While this chapter will not consider the various emotional theories in more detail, there are a number of good reviews related to affective computing that are available. These include reviews of detection approaches (Pantic & Rothkrantz, 2003; Sebe, Cohen, & Huang, 2005; Zeng, Pantic, Roisman, & Huang, 2009) and the various emotional theories that underpin this work (Barrett, Mesquita, Ochsner, & Gross, 2007; Dalgleish et al., 2009; Russell, 2003). In the next section we consider how emotions are currently detected for applications of affective computing.

2.2 *Detecting Emotion*

Computer systems designed to detect and respond to human emotional states must trade-off against a number of criteria, including reliability, speed, cost, intrusiveness, and validity (Calvo & D'Mello, 2010). As such, a number of different approaches have been tried that focus on replicating human abilities for interpreting facial expressions, speech, body language, or a combination of these signals.

Detecting emotions from facial expressions assumes that standard expressions (Ekman, 1992) are automatically triggered in response to an affective state being experienced. The Facial Action Coding System (Ekman & Friesen, 1978) was developed to standardise the recognition of the common emotions of joy, sadness, surprise, fear, disgust, and anger. These facial expressions are broken down to smaller units of facial motion that can be identified by trained human observers. While this manual decoding process is expensive, there are ongoing efforts to automate this process using a range of algorithmic classifiers such as Bayesian networks (Gunes & Piccardi, 2007), discriminant analysis (McDaniel et al., 2007), and support vector machines (Bartlett et al., 2006). This approach has been used for educational support (McDaniel et al., 2007) both alone and also in combination with other types of physiological sensors (Arroyo et al., 2009). However, while automated techniques continue to improve, they are generally not yet as effective as manually decoded approaches as most fail to operate in real time or take into account the context in which the interactions are occurring (Zeng et al., 2009).

Another promising approach that relies on the innate expression of emotion is to detect changes in body posture or movement that reflect underlying emotional states (Calvo & D'Mello, 2010). Unlike facial expressions, body movement is usually less prone to conscious control and disguise, and so may provide a more reliable channel of information (Ekman & Friesen, 1969a, 1969b). Posture analysis has previously been used to classify interest levels in children during 20 min of serious gameplay (Mota & Picard, 2003). In this experiment nine posture positions: *sitting on the edge*, *leaning forward*, *leaning forward right*, *leaning forward left*, *sitting upright*, *leaning back*, *leaning back right*, *leaning back left,* and *slumping back* were used to three levels of interest (low, medium, high) and the further states of, *taking a break* and *bored*. A similar posture detection system, based on measuring the distribution of body pressure in a chair, was used to categorise *boredom*, *confusion*, *delight*, *flow*, and *frustration*, from *neutral* while college students used an intelligent tutoring systems designed to teach Newtonian mechanics (D'Mello & Graesser, 2009).

The rhythm, stress, and intonation of speech, along with other vocalizations, such as, sighs and laughter have been used extensively to try and detect emotional states (Juslin & Scherer, 2005; Russell, Bachorowski, & Fernandez-Dols, 2003; Zeng et al., 2009). These systems tend to focus only on detecting basic emotions, but they do have the advantage of being nonintrusive, low-cost, fast, and suitable for working with spontaneous real-world speech (Calvo & D'Mello, 2010). Semantic emotional cues can also be extracted from text or speech content using associations between words and affective dimensions such as good or bad, active or passive, and strong or weak (Osgood, May, & Miron, 1975). Furthermore, analysis of word counts and structured sets of words such as Wordnet (Strapparava & Valitutti, 2004) and ANEW (Bradley & Lang, 1999) allow for automatic semantic analysis of text to detect affective states. This approach has been extended to allow for categorising sentiment and opinion analysis of larger populations into emotional categories such as good/bad or angry/sad (Pang & Lee, 2008).

Many non-invasive techniques based on measuring physiological signals or brain activity monitoring and brain imaging have been developed in fields such as psychophysiology and neuroscience (Calvo & D'Mello, 2010). Assuming physiological state and brain activity are appropriate measures of affect, all of these approaches suggest promise in terms of providing objective measures of a user's emotional state. Typical measures include skin conductance (GSR), brain activity (EEG, MRI), heart activity (ECG), and muscle activity (EMG). The specificity of particular patterns of physiology for detecting specific emotions using such measures of the autonomic nervous system (Ekman, Levenson, & Friesen, 1983) needs to be balanced against significant variations that are known to occur between individuals (Andreassi, 2007).

A number of physiology-based systems have been used to categorise different emotions (Alzoubi, Calvo, & Stevens, 2009; Calvo, Brown, & Scheding, 2009; Nasoz, Alvarez, Lisetti, & Finkelstein, 2004; Picard, Vyzas, & Healey, 2001; Vyzas & Picard, 1998). However, the two key dimensions that can be distinguished using physiology are arousal and valence. High levels of arousal are categorised with faster heart rate and other physiology changes that are activated for human actions

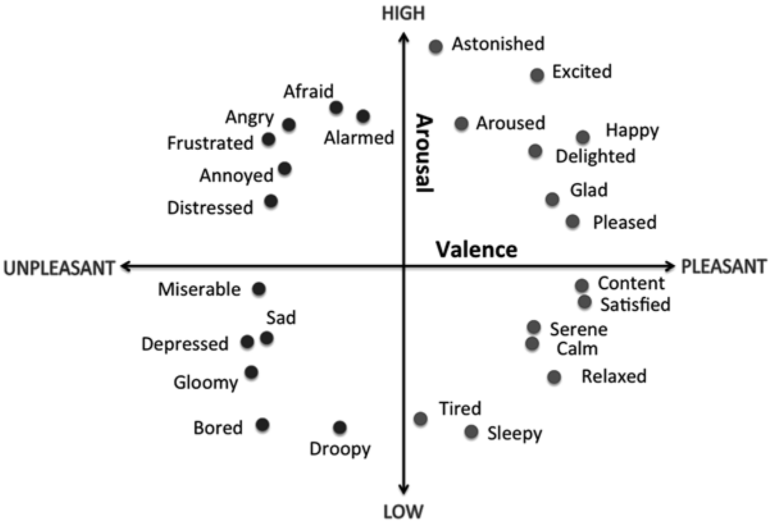


Fig. 18.1 Two-dimensional model of valence and arousal (adapted from Russell, 1980)

such as fright, flight, and fight. Valence, by contrast, refers to either positive or negative association of affect, for example, happy and sad feelings. Modulations in the startle reflex are typically used to measure valence.

This two-dimensional model, using valence and arousal, was originally described in a circular structure that categorised "core affects" (Russell, 1980) (see Fig. 18.1). More recently, a relationship between core-affect and general product experience has been described (Desmet & Hekkert, 2007). Likewise, the two dimensions of arousal and valence were used to develop the Psychophysiological Emotional Map (Villon & Lisetti, 2006) (see Fig. 18.2). This map was developed using 28 measures extracted from heart rate and skin conductance sensors. In this map, *sadness* is categorised as low arousal and low valence, while *fear* is associated with low valence and high arousal (Villon & Lisetti, 2006). By contrast, *happiness* is distinguished by high valence with a range of higher arousal levels, while *calmness* has a high valence but low arousal (Villon & Lisetti, 2006).

In terms of game design, we might expect that people actively seek out and purchase games that deliver positive emotional experiences and enjoyment. However, this needs to be considered in light of the player's intent, as an enjoyable game may be one that intentionally elicits negative emotions. This is due to the possible enriching effect of negative emotions embedded within positive experiences and products (Fokkinga, Desmet, & Hoonhout, 2010). It is therefore possible that games featuring what are putatively negative actions may prompt positive responses (Ravaja et al., 2006). This may be due to the threats within the game appearing as a challenge to the player rather than a real threat, or that the player finds surviving in an environment perceived to be dangerous as rewarding (Ravaja et al., 2006).

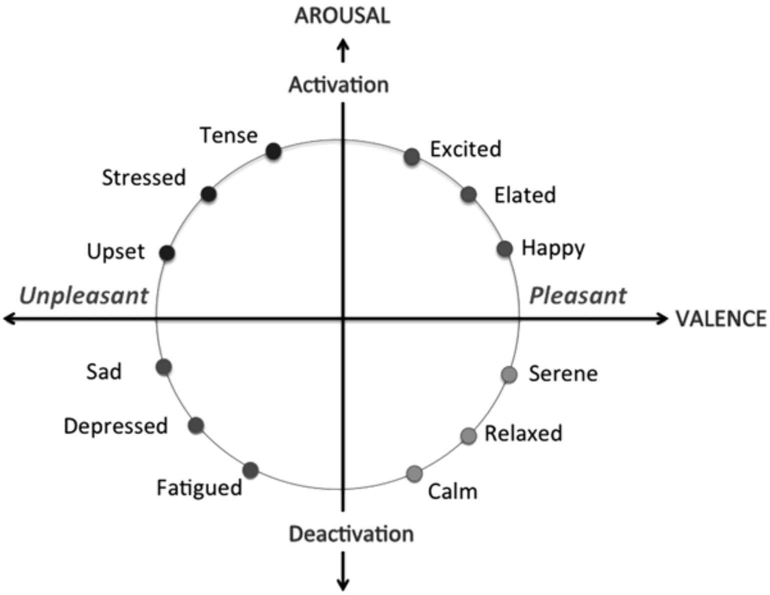


Fig. 18.2 The psychophysiological emotional map based on valence and arousal (adapted from Villon & Lisetti, 2006)

Regardless of the positive or negative emotional reaction, computer games, like other affective interfaces, can act as a stimulus for *affective processing*, which results in associated feelings or emotions in the user. When a serious game designer can nominate desirable affective states that relate to the serious intention of the game, it is feasible to use affective computing tools to evaluate the design. It is, however, important when evaluating a game design that the intention of an event or game scenario has clearly defined expectations around what emotion the designer is trying to elicit in the player.

The time or game state at which player affect is measured is also critical. Affect, or affective processing, is bound in time to the experience of the game world and the resulting emotional effect that this has (Barrett et al., 2007). This suggests other game analytics should be used in conjunction with affective measures so that player affect is carefully correlated with the game state.

In summary, for interface design, the term affective processing is perhaps a preferable construct to emotion, as the latter is more prone to confusing and arbitrary definitions (Scherer, 2005). Additionally, affect is subconscious and is a more reliable indicator of a person's core emotional state than self-reported emotion (Filion, Dawson, & Schell, 1998). While subjective ratings from players provide useful information about their perceived emotion, the subconscious nature of affect offers further opportunities for measurement through the collection of physiological data. The startle eye-blink is one such physiological measure that can help to determine the participant's affective processing, and in particular, measure the valence of a player's reaction to a startling event. In the next section, we describe previous uses of this measure in game research and provide detailed guidelines for eliciting, recording, and analysing this measure.

3 The Startle Reflex

The concept of a reflex is well known. It is an automatic direct motor response to a stimulus above a certain threshold. Perhaps, the most well-known reflex is the knee-jerk (patellar reflex), but there are various other such automatic motor responses, one of which is the so-called eye-blink reflex. When one is startled by, for instance, a loud noise like a gunshot, bright flash, or a sudden explosion, an involuntary eye-blink is elicited. Although an eye-blink occurring as a startle reflex is an automatic response, its magnitude varies as a function of affective state (Filion et al., 1998). The more positive the current state of affect, the smaller the eye-blink magnitude. The more negative the current affective state, the larger the eye-blink magnitude. This simple correlation forms the very basis for the startle eye-blink to be an excellent measure of affective processing related to any given stimulus, situation, or game being played. Following, we provide more detailed background information, including example studies, which demonstrate the potential of this measure for the serious game community.

3.1 Previous Uses of the Startle Reflex

One of the more interesting applications of the startle reflex has been in the study of people's responses to commercial products for marketing purposes (Walla, Brenner, & Koller, 2011). This study found significantly reduced eye-blink amplitudes related to "liked" brand names compared to "disliked" brand names. In another marketing study, the startle reflex was used to measure significant differences in preference for bottle shape (Grahl, Greiner, & Walla, 2012). Likewise, the amplitude of the startle response was shown to be stronger when individuals experienced unpleasant versus pleasant odors (Kaviani, Wilson, Checkley, Kumari, & Gray, 1998). Measures of the eye-blink amplitude have also been used to distinguish different affective responses associated with eating different foods (Walla, Richter, Färber, Leodolter, & Bauer, 2010). Compared to eating yoghurt and chocolate, eating ice cream results in the lowest startle responses, or the most positive affect.

In terms of multimedia, a traditional use of the startle reflex in psychology involves grading pleasant versus unpleasant images (Allen, Trinder, & Brennan, 1999; Vrana, Spence, & Lang, 1988). This work typically relates startle results with standardised image libraries such as the International Affective Picture System (Bradley & Lang, 2007) and the Geneva affective picture database (Dan-Glauser & Scherer, 2011). These standard databases are well-correlated with both valence and arousal and can form a useful baseline to study the variations in startle response between individuals.

The startle reflex has also been used to study responses to other media, with the amplitude of the responses shown to be stronger when listening to unpleasant versus pleasant music (Roy, Mailhot, Gosselin, Paquette, & Peretz, 2009). A similar result has been found in emotionally toned film clips (Kaviani, Gray, Checkley, Kumari, & Wilson, 1999), and the response has been used to measure the viewer's emotional response to television content (Bradley, 2007).

Virtual realities have much in common with computer games and they have also been used in conjunction with the startle response. For example, a study comparing real-world effects with virtual environments used the startle response to determine that participants actively driving through virtual tunnels experienced more negative feelings while in the darker parts of the virtual tunnel (Muehlberger, Wieser, & Pauli, 2008). In a further example, the startle response was used in conjunction with Google Street View to objectively assess affective processing associated with different urban environments (Geiser & Walla, 2011). In this study participants had to virtually walk through six districts of Paris with different median real estate prices. The eye-blink magnitudes of participants were recorded during these walk-throughs. Real estate price was strongly correlated with explicit pleasantness ratings, and the startle measures confirmed affective differences between the most expensive and cheapest districts (Geiser & Walla, 2011). In a further study, a virtual environment viewed from the perspective of the driver of a Humvee was used to examine variations in eye-blink responses in both low-threat and high-threat zones, under immersive and non-immersive conditions, while driving through a virtual Iraqi city (Parsons, Rizzo, Courtney, & Dawson, 2012). The participant's eye-blink amplitudes increased in the high-threat zone under the high immersion conditions.

Much of the prior research using the startle response in relation to video games examines the tendency for video games to encourage violent behaviour (Wood, Griffiths, Chappell, & Davies, 2004). For example, a recent doctoral dissertation examined the effect of violent video gameplay on modulation of the startle reflex (Elmore, 2012). The study found that participants who played violent video games before being shown unpleasant images elicited lower eye-blink responses (Elmore, 2012). The results were used to support the idea that violent video games desensitize players to violence.

Another related example is the investigation of the effects of violent video games using psychophysical measures such as facial electromyography, skin conductance level, and heart rate (Ravaja, Turpeinen, Saari, Puttonen, & Keltikangas-Järvinen, 2008). In this experiment participants' real-time emotional responses to playing violent video games were recorded. The study found that all violence within the game either perpetrated by their character or on their character resulted in an increase in arousal. However, violence perpetrated against the player's character was associated with negative emotion measures, while violence perpetrated by the player's character was associated with positive emotional measures (Ravaja et al., 2008).

While the use of the startle reflex for studying affect in relation to game design elements has previously been proposed (Lang, 1995; Nacke, 2009; Ravaja &

Kivikangas, 2008; Sasse, 2008), few definitive results seem to be reported. In essence, most previously reported research efforts using the startle reflex have attempted to answer the fundamental question of whether video games are "good" or "bad" in terms of influencing future behaviour. While video game designers appear to follow trending design decisions and internal rules, little evidence exists to suggest that developers are using psychophysiological measures such as the startle reflex to make their games more appealing (Wood et al., 2004) or in the case of serious games, more useful.

There are a few exceptions where studies have reported results using the startle reflex to study game design elements. For example, in one study researchers used various physiological indicators such as heart rate, skin conductance, and the startle reflex to gauge the immersion of participants while playing a bespoke level of Half-Life 2 (Grimshaw, Lindley, & Nacke, 2008). The information gathered was associated with the participant's sense of immersion, with a higher magnitude response indicating that the player was more engaged with the game at the moment of startle pulse (Grimshaw et al., 2008). In a more recent project, the startle response was used to gauge the immersion related to sound on and off conditions in a commercial horror game (Coppins, 2014). As such games are designed to create a sense of fear, it is in theory reasonable to use the startle reflex measure to evaluate how well a negative valence associated with the emotion of fear is generated. Although no significant differences were found in the startle amplitude with the two sound conditions, a significant variation was detected when participants actively played the game as opposed to the situation where they simply watched a replay of the game.

One possible reason for the still limited use of startle response in the game industry, and in particular in the development of games, is that designs tend to be subjectively evaluated. Arguably, this is also the case in the film industry where the manipulation of emotion in viewers is a well-honed skill. Despite this, there are a number of studies that illustrate why a subconscious measure like the startle reflex may be of use in quantifying player's responses in serious games.

Principally, affect is subconscious, and thus the startle reflex is a more reliable indicator of a person's core emotional state than self-reported emotion (Filion et al., 1998). For example, in a study investigating the modulation of the startle reflex in depressed versus healthy populations (Allen et al., 1999), it was found that while the self-reported pleasantness measure related to picture presentations was largely similar, the startle reflex data showed clear differences between depressed and non-depressed participants (Allen et al., 1999). The depressed group did not show the typical finding that pleasant images elicit a significantly reduced startle reflex compared to unpleasant images, which indicates that internally, depressed people responded rather negatively to positive image presentations. Such a discrepancy demonstrates how misleading self-reported data can be, especially when related to affective content.

In another study, psychopaths demonstrated normal self-reported responses to emotional images, whereas they did not show typical startle response enhancement as a consequence of unpleasant image presentation (Patrick, Bradley, & Lang, 1993). Once again, this clinical investigation suggests that the startle reflex may provide important information about the inner state of affect of a person that may be more reliable that any explicit response. In some more industry-related studies, similar discrepancies between explicit and implicit measures of affective processing have also been found (Geiser & Walla, 2011; Grahl et al., 2012). Thus, the startle reflex measure may tell us more about the actual state of affect of a person than the person is actually able to do by themselves. While we do not discount the importance of subjective feedback in game design, we do believe an objective measure like the startle response suggests itself as a useful adjunct that can be used in the analysis of game designs.

3.2 Using the Startle Reflex

A number of measurement techniques have been developed for studying different aspects of the startle response (Dawson et al., 1999). The simplest, cheapest, and most frequently used approach for research into affect involves surface electromyographic (EMG) recording of action potentials generated by the orbicularis oculi muscle. Using two electrodes placed on the skin just below the eye, and an isolated earth electrode placed on the forehead, it is possible to reliably detect even small voltage changes produced by the orbicularis oculi muscle (see Fig. 18.3).

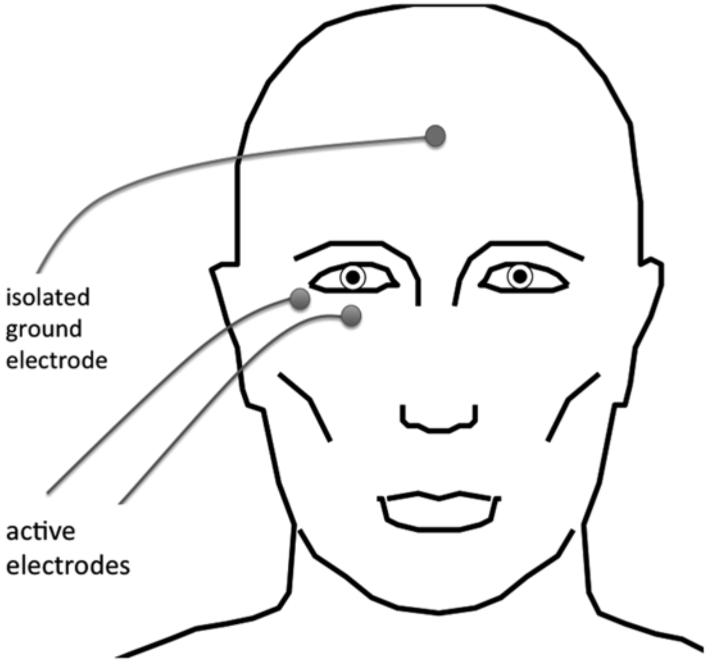


Fig. 18.3 Electrode positions used to record eye-blink magnitude

These changes in voltage are associated with contraction of this muscle during an eye-blink. While eye-blinks can be detected using video processing with frame rates of over 500 Hz, a distinct advantage of surface EMG is that even weak responses that do not result in discernable blinks can be detected using this approach. This provides an advantage over techniques that rely on directly measuring physical movement of the eyelid. One potential disadvantage of the approach is that it requires sensors to be attached to an individual's face. Furthermore, sensors need to be tethered by wires to recording equipment, making EMG problematic for interfaces that require the player to make large movements.

EMG measures of eye-blink can also be prone to some noise as the changes to surface potentials are small and external electromagnetic interference is common in most environments. Where precise measures are required, magnetic search coils can be placed on the skin to detect subtle changes in magnetic field associated with electrical activity in muscles (Evinger & Manning, 1993). A disadvantage of this approach is the requirement for even larger, more intrusive sensors than required for EMG.

The measurements used in startle eye-blink studies are normally taken from the orbicularis oculi muscle, a muscle that causes a blink (among other functions). An eye-blink reflex is transmitted by the facial nerve. However, the facial nerve also innervates other key facial muscles that are sometimes studied as part of affect research such as the zygomatic and corrugator supercilli muscles. The zygomatic major and minor muscles are associated with facial expressions involving the lips such as smiling, while the corrugator supercilli muscle, sometimes called the frowning muscle, is associated with wrinkling of the forehead. Positive affect has been shown to increase activity in the zygomatic muscles, while negative emotions cause an increase in activity of the corrugator supercilii (Dimberg, Thunberg, & Elmehed, 2000).

It was in the late 1980s, after many pioneering investigations in rodents, that it was found that humans demonstrate a modulated startle reflex as a function of degree of pleasantness (Vrana et al., 1988). Since then, the magnitude of an eye-blink as a response to loud and short acoustic white noise, containing a broad spectrum of frequencies for about 50 ms at a sound level of 105 dB and with a rapid onset, has been used to study affective valence (Mavratzakis, Molloy, & Walla, 2013; Walla et al., 2011). Guidelines on the use of human startle eye-blink EMG studies provide clear direction and consensus on the appropriate use of the technique (Blumenthal et al., 2005). The process of measuring and further detail about analysing the startle eye-blink is available elsewhere (Blumenthal et al., 2005). However, for convenience, the key steps of that process, preparation, eliciting, processing, analyzing, and reporting, are summarised here.

3.2.1 Preparing for Measurement

The eye-blink startle is usually measured by recording changes in surface potential using two electrodes placed below one of the eyes (see Fig. 18.3). Two electrodes are used to independently measure voltage changes and ensure that noise on either electrode can be accounted for. A sudden change in potential is indicative of the brief electrical signal, called an action potential that causes contraction of all, or parts of the orbicularis oculi muscle. The magnitude of the current measured is small, in the order of 0–300 μV, so careful preparation is required to ensure a reliable measurement (Blumenthal et al., 2005).

It is vital that the skin is carefully cleaned before placing the electrodes to help reduce impedance to the electrical signal. This can be done by rubbing the skin with gauze and cleaning with soap and water or alcohol. To further improve impedance, a small amount of electrode gel can be applied to the specific surface of the site of each electrode. However, care must be taken to ensure that the electrode gel does not complete a circuit between the two electrodes. Due to the sensitive nature of skin below the eye, care also needs to be taken that no abrasive materials are used in the preparation and the participant's eyes are closed so that alcohol fumes do not become a source of irritation (Blumenthal et al., 2005).

The orbicularis oculi surrounds the eye. While the eye-blink response is more precisely discerned on the top of the eye, this is an uncomfortable position for electrode placement and the motion of the upper eyelid can introduce artifacts into the detected signal. The recommended type of electrodes are AG/AgCl miniature electrodes, smaller than 5 mm, contained in a recessed plastic casing with external diameter of less than 15 mm and filled with electrode gel (Blumenthal et al., 2005).

An isolated ground electrode is typically attached to an electrically inactive site such as the middle of the forehead or temple. One active electrode is typically positioned in line with the center of the pupil while the participant looks directly ahead, and a second about 1–2 cm lateral to the first active electrode. The electrodes can be attached using double-sided adhesive collars. It is important to avoid overlapping of the electrode attachment and that the electrodes are placed to ensure they do not interfere with normal eye movement (Blumenthal et al., 2005). It is advisable to check the signal that is being detected before proceeding by asking the participant to perform a voluntary blink. Where the EMG signal is not clear, it may be necessary to reposition or reapply the electrodes.

The two active electrodes need to allow for the same level of conductance to ensure a consistent measure, and as previously noted, high impedance on either electrode can limit the ability to record an accurate signal. The baseline signal should also be inspected for high levels of background noise. Interference from background power lines and equipment in the 50–60 Hz range can be a common problem in some environments and should be avoided if possible. EMG signals from the two active electrodes are amplified differentially, so noise can be reduced by braiding the cables of the two electrodes together and ensuring they are picking the same level of noise (Blumenthal et al., 2005). Shielding equipment may also be used or a specialised environment set aside that is free of excessive electromagnetic interference. However, in computer game studies this is not always possible, so as a fall back, a notch filter in the 50–60 Hz range can be used to reduce noise in the signal. However, use of such a filter will also reduce the measured EMG signal from the eye-blink response that occurs in this 50–60 Hz frequency range.

Another source of noise in EMG measurement can come from large head and eye movements of the participant. This can be controlled in some experiments where participants can be asked to focus on a stationary point and avoid movement. However, this is more difficult to control with active game interfaces. It may be necessary to monitor participants for such movement during the study. Startle responses corrupted by movement of the electrodes may need to be excluded from the study during the analysis phase.

3.2.2 Eliciting the Startle Response

To measure the magnitude and latency of the eye-blink response, the response must first be elicited. Eye-blinks can be elicited by a range of acoustic, visual, electrical, magnetic, and mechanical stimuli, each of which may create variations in the measured response (Blumenthal et al., 2005). Indeed, variations in the response can be caused by the number of factors, such as the frequency of presentation, the background conditions, the composition of the stimulus, as well as the way it is presented. The most commonly used approach is to use an acoustic startle, and white noise is generally the most effective stimuli. This suggests a sound that consists of broadband noise containing frequencies in the range of 20 Hz to 20 kHz.

The magnitude of response, the speed of onset, as well as the probability of elicitation are increased with higher intensity sounds. The response can be influenced by the intensity of the sound and other properties of the sound envelope such as the rise time and duration (Blumenthal et al., 2005). A typical acoustic stimulus is characterised by a maximum amplitude of 100 dB(A) SPL, a rapid rise time, and a duration of around 50 ms (Blumenthal et al., 2005). In summary, sudden, short, loud sounds are more startling.

Another factor that is known to affect responses to an acoustic startle stimulus is the level and nature of other background sounds. For example, pulsing sounds can inhibit the response, while consistent background noise can help facilitate the startle response (Hoffman & Fleshier, 1963). Indeed variations in startle response are often studied by using a prepulse sound prior to the pulse of startle sound. The slightly weaker prepulse sound normally inhibits the stronger startling stimulus with a maximum inhibition typically observed with a 120 ms interval (Graham, 1975). Prepulse inhibition is used in the study of a range of psychological disorders such as schizophrenia (Swerdlow et al., 2008) and conditions that impact on attention (Filion et al., 1993).

An acoustic startle stimulus can be presented either by headphones or loudspeakers. In both cases the intensity of the presentation signal needs to be calibrated using a sound level meter. Properly fitted headphones can ensure a more consistent delivery of the startle stimulus, but can also interfere with other equipment and electrodes. By contrast, the use of loudspeakers may require targeted positioning of the participant between loudspeakers to ensure a consistent presentation of the startle stimulus.

3.2.3 Processing the EMG Signal

The EMG signal related to the eye-blink response oscillates between both positive and negative values around a zero value, in the frequency range of 28–500 Hz (Blumenthal et al., 2005). This suggests that the EMG signal should be recorded at a minimum sampling rate of 1,000 Hz. The time frame of interest in the startle blink is in the order of 0–500 ms. The raw EMG signal, measured on the surface of the skin, is a low voltage signal, typically in the order of a few microvolts (μV), where 1 V is equivalent to 1,000,000 μV.

For analysing the startle response measure, there are a number of key parts of the surface EMG signal that may need to be considered. These include: latency, amplitude, baseline amplitude, peak amplitude, duration, and the integrated EMG (IEMG).

Latency, measured in milliseconds (ms), is the time between the presentation of the startle stimulus and the onset of a significant change in surface EMG that indicates activation of the muscle fibers underlying the active electrodes. The two challenges in recording this signal are to ensure the timing of the stimulus presentation is synchronised with the raw EMG signal and identifying the onset of the response. The first challenge can be overcome by triggering the presentation stimulus electronically using an output channel from the same recording equipment that is monitoring the EMG signal. Conversely, the actual audio startle or an externally generated trigger can be fed into the recording device to accurately mark the raw EMG signal with the exact time the startle stimulus is presented.

Amplitude is a measure of the magnitude in microvolts (μV) of the average EMG signal at a point or interval of the EMG signal. For the startle response, this is usually reported as a magnitude of the rectified EMG signal. The rectified signal is the absolute value of the raw signal and so only contains positive values. This is in contrast to the raw EMG signal that oscillates between positive and negative values. The baseline amplitude is a measure of background electrical activity being detected during an interval of muscle inactivity. For the startle response, it is typically calculated as a mean value for a period of around 150 ms just prior to the startle stimulus. This mean calculation should include the positive and negative variations in the EMG signal. The baseline amplitude can be subtracted from other amplitude measures to help quantify EMG activity that is specific to a muscle response.

The peak amplitude, also measured in microvolts (μV), is the maximum amplitude in an interval of the EMG signal. For the startle probe, the interval of interest is typically taken between the onset of the startle response and the return to baseline of the signal. This peak value minus the baseline amplitude is the measure of most interest for inferring the valence associated with affect.

The duration of the startle response would typically be the interval between the response onset and the return to baseline of the signal. The onset and end of the response is often identified by visually inspecting at the raw EMG trace. This inspection process can be simplified by using a smoothed EMG signal that is cleaner to inspect for key features. This smoothing can be achieved, for example, by a technique like a moving average filter. Longer time filters create more smoothing, but also tend to lower the observed variations in the signal amplitude. Regardless, the selection of onset and end point requires some experience from the observer and can introduce some subjectivity. This subjectivity can be partially offset by using a group of independent observers to select and reject key features and then combining the results. An alternative is to automate this process by considering the standard deviation of the signal from the EMG baseline. For example, the onset might be selected when the mean of a short interval of the signal exceeds two standard deviations of the baseline. The end of the response might be gauged by an interval where the mean returns to within one standard deviation of the baseline.

While peak amplitude is commonly used to assess startle response, an alternative is to use the area under the curve of the rectified EMG signal for a specified interval, such as the duration of the response. During a startle-blink, not all muscle fibers involved may be activated simultaneously. The measured surface EMG may be a composite of the electrical changes due to multiple contractions occurring in different muscle fibers. The integrated EMG can provide a measure of the force of the combined responses, being dependent on both the magnitude and duration of the response. This integrated value is measured in units of microvolt per second.

A detectable startle response is not always elicited in response to a startle stimulus. This is because some individuals may not have a normal response, short- or long-term habituation of the startle response (Valsamis & Schmid, 2011), experimental conditions, or the treatments of interest. Therefore, another calculation that is often reported for a study is the probability that the presentation of the startle stimuli produces an actual startle response. This involves detecting what are called zero or non-responses in relation to the startle stimuli. Zero responses are identified by no significant change in the baseline of the raw EMG signal in a short interval following the stimulus. The onset interval can vary with experimental conditions, but should be identifiable within 20–150 ms of the startle stimulus (Blumenthal et al., 2005). Using a short onset window helps distinguish real responses from background activity and voluntary or spontaneous blinks. To avoid short-term habituation to the response, it is recommended that intervals between startle stimuli are randomised and at least 30–60 s apart (Valsamis & Schmid, 2011).

In general, the processing of the EMG signal can be considered in four distinct steps: amplification, filtering, rectification, and finally either a smoothing or integrating step (Blumenthal et al., 2005).

Amplification of the raw surface EMG is required because it is low voltage signal. The two closely placed and active skin electrodes, located on the orbital muscles, measure underlying changes in muscle voltage related to the action potentials that signal the muscles to contract. The raw signal from these two active electrodes used to measure the startle response is usually differentially amplified. This requires an isolated AC-amplifier with high impedance (>100 MΩ), a high common rejection ratio (>100 dB), and low input noise (Blumenthal et al., 2005). Large individual variations in amplitude can occur between participants and also vary with stimulus conditions and trials. Adjusting the amplification needs to avoid clipping that can occur in the Analog to Digital conversion process, and also be wary of missing small but significant amplitude changes in the signal. For this reason, it is advised that the highest possible resolution be used in the Analog to Digital conversion process. In the order of 16–24 bits is advised, with values sampled at least 1,000 times per second (1 kHz) (Blumenthal et al., 2005).

Filtering of the raw EMG signal is designed to maximize the signal to noise ratio and allow for better detection of the eye-blink response. Background interference is first removed by filtering out frequencies below 28 Hz and above 500 Hz. The low frequency noise can be due to motion of the electrodes or other biological sources such as eye movements, retina activity, or the contraction of other facial muscles. To remove these low frequency artifacts, a digital high-pass filter with an infinite impulse response and 3 dB cutoff at 28 Hz is recommended (Blumenthal et al., 2005). Higher frequency noise due to electrical instruments and background electromagnetic interference can be removed with a low-pass filter. A low-pass, finite impulse filter with a roll-off of 24 dB per octave is one recommended configuration (Blumenthal et al., 2005).

Rectification of the filtered signal is achieved by taking the absolute value of the raw EMG values. This removes problems when averaging the signal that oscillates between positive and negative values. This assumes that the output DC level of the amplifier is centered on zero. During the rectification process, it is also normal to subtract the mean baseline value of the signal. This baseline value, as previously described, can be obtained during a selected pre-stimulus interval, by calculating the mean of the raw EMG values recorded during this interval.

The final step in processing the signal involves the application of smoothing filters and/or the calculation of an area under the curve of the signal amplitude. Various approaches for smoothing are possible and include a simple moving average filter or a variable weight filter if it is desirable to avoid phase shift and multiple peaks in the response (Blumenthal et al., 2005). The calculation of the integral of the signal requires the selection of onset and end points for the response and can be automated with various signal processing techniques, such as a contour following integrator.

3.2.4 Analysing Responses

The final critical step in this process is to analyze the processed signals to identify and quantify the key elements of the startle signal. These include previously discussed values such as the peak amplitude, latency of response, and the probability that a response is elicited after the presentation of the startle stimulus.

The analysis of the processed EMG signal is often performed manually, but might be computer-assisted or fully automated to avoid some subjective bias in the process. The first step of the analysis process involves deciding if each startle response can be discriminated or not. This may not be possible if the signal is contaminated by noise, or if a spontaneous or voluntary eye-blink has occurred around the same time as the stimulus. Movement of the electrodes can sometimes generate artifacts on the EMG signal that exceed the amplitude of any startle response, which prevents reliable identification of the startle response. Furthermore, the startle response should only be elicited within a 20–150 ms time window after the startle stimulus.

Thus, a response that occurs outside this onset window should also be rejected. Any rejected trials should be excluded from all further calculations.

Once a response is accepted, it is possible that the response is too small to include and should be categorised as a zero response (non-response). A value of zero is then recorded for the amplitude of this non-response. For any response that is deemed significant, the key characteristics of the response need to be quantified. For example, of those characteristics, the peak amplitude (measured in microvolts) is typically of most interest in startle studies related to measuring affective valence.

In the next section, we briefly introduce some possible uses of the startle eye-blink as an analytical tool to support game design. The next section also reports on a preliminary case study that uses the startle eye-blink to measure the player's affective valence when interacting with three key parts of a serious game. This case study serves to illustrate the use and reporting of a study using the startle eye-blink.

4 The SHADOW Case Study

There are two basic approaches for integrating game analytics into the game design process; one is summative in nature and the other more formative in intent. A summative evaluation is designed to test a clear design hypothesis in a finished game. For example, the study of Coppins (2014) uses the startle reflex to assess the role of sound in eliciting player affect in a commercial horror game. The intention of this study was to better understand how sound is used in a completed game to create immersion. The second approach involves a more formative evaluation during game development and the intention is to guide or refine a game element. This assumes an iterative development approach and implies a less structured use of the startle reflex to measure player affect surrounding some key design elements of the game.

In the SHADOW case study, we describe such a formative use of the startle eye-blink measure to examine the affective response of players to the three key sections of a serious game. The game is being designed to support psychological counselling where the efficacy of the game is dependent on players learning new skills to manage their own behaviour.

4.1 Background to SHADOW

SHADOW, the serious game, is designed to support online psychological counselling of younger adults aged 18–30, of both genders (Hookham, Deady, Kay-Lambkin, & Nesbitt, 2013). The SHADOW game builds on a more traditional web-based counselling tool called SHADE, which was designed as a clinician-assisted intervention program for the treatment of comorbid depression and alcohol or other drug use problems (Kay-Lambkin, Baker, Kelly, & Lewin, 2011).

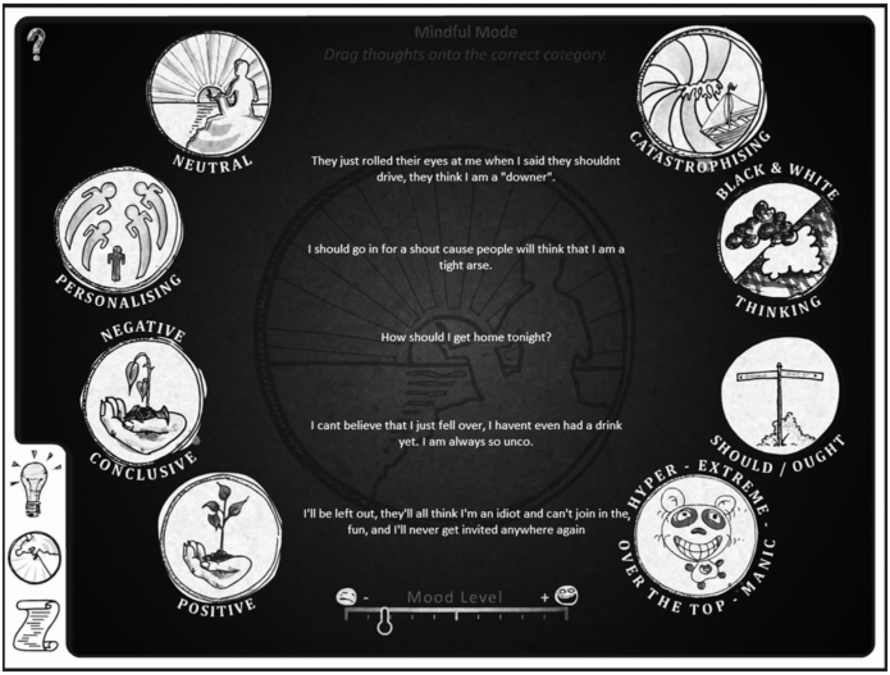
SHADE, the precursor to SHADOW, is an internet-delivered, evidence-based, psychological treatment that uses the principles of Cognitive Behavioural Therapy (CBT), mindfulness meditation, and motivation enhancement to target these conditions in an integrated way. A major objective of CBT is to identify and challenge the unrealistic beliefs that maintain a person's problematic patterns of thinking and behaviour (Beck, Rush, Shaw, & Emery, 1979). In combination, "mindfulness" is an important skill, particularly when learning how to cope with negative automatic thoughts that are associated with depression and drinking alcohol. The central idea of mindfulness is not to prevent these thoughts from occurring, but rather to stop these thoughts from setting in and taking control when they are triggered (Segal, Williams, & Teasdale, 2002).

The efficacy of the SHADE intervention program has previously been demonstrated in a large clinical trial (Kay-Lambkin et al., 2011). However, the efficacy of the program requires participants to complete the 10-week program (Kay-Lambkin et al., 2011) and develop the key skill of mindfulness and thought management. These two skills are considered critical to the efficacy of the SHADE program, but the introduction of these skill-training components at the half way stage of the SHADE program also coincides with the time when most participants choose to leave the program.

Thus, a key motivation for developing the SHADOW game is to try and make this training in mindfulness and thought management a more engaging and positive experience for participants. As such, a key design element of the game is the "Mindfulness" task (see Fig. 18.4). This task challenges players to be mindful of their thoughts, their underlying mood level, and to subsequently manage unproductive thoughts. In this case study, we use the startle reflex to measure the player's affective response to this key game element. We do this by comparing it to the player's affect in two other key parts of the game; the instruction section (see Fig. 18.5) and the general game play (see Fig. 18.6). Apart from the assumption that more positive affect leads to better learning outcomes, a further assumption of this study is that the most positive valence in the game needs to be associated with the mindfulness task as this is where the key skill training occurs.

4.2 Method

Seven participants, 5 male and 2 female, within the ages of 18–30 were recruited for the study using poster and word of mouth. The participants in the study were mainly recruited through convenience sampling, and as such consisted primarily of students at the University of Newcastle. Participants were required to have normal or corrected to normal vision. All participants were informed through a participation information statement about the intention and methods to be used in the experiment, including the fact that occasional startling noises would be played while they played the SHADOW game.



Mindful Mode

Drag thoughts onto the correct category.

They just rolled their eyes at me when I said they shouldnt drive, they think I am a "downer".

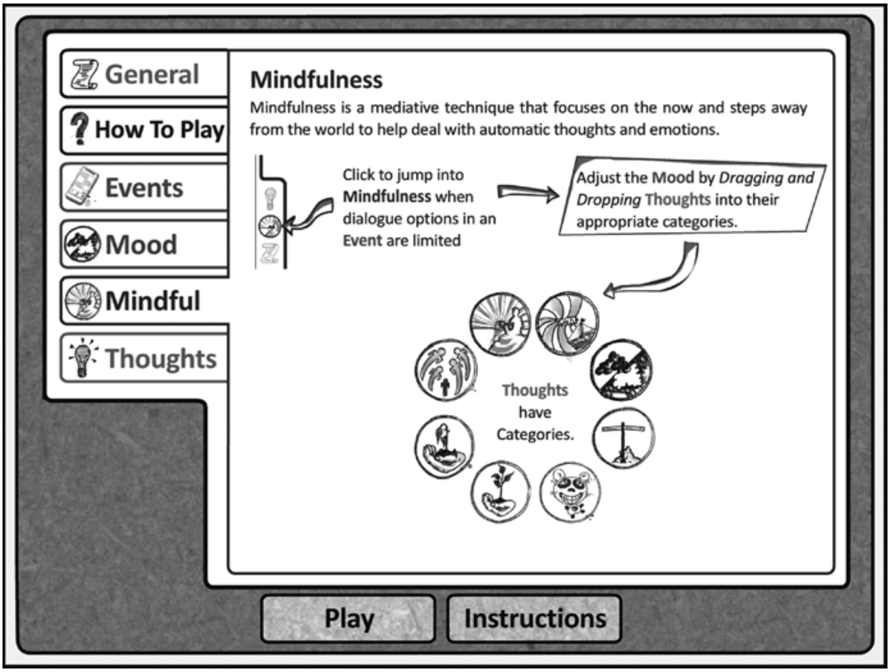
I should go in for a shout cause people will think that I am a tight arse.

How should I get home tonight?

I cant believe that I just fell over, I havent even had a drink yet. I am always so unco.

I'll be left out, they'll all think I'm an idiot and can't join in the fun, and I'll never get invited anywhere again

Fig. 18.4 The mindfulness component of the SHADOW game



Mindfulness

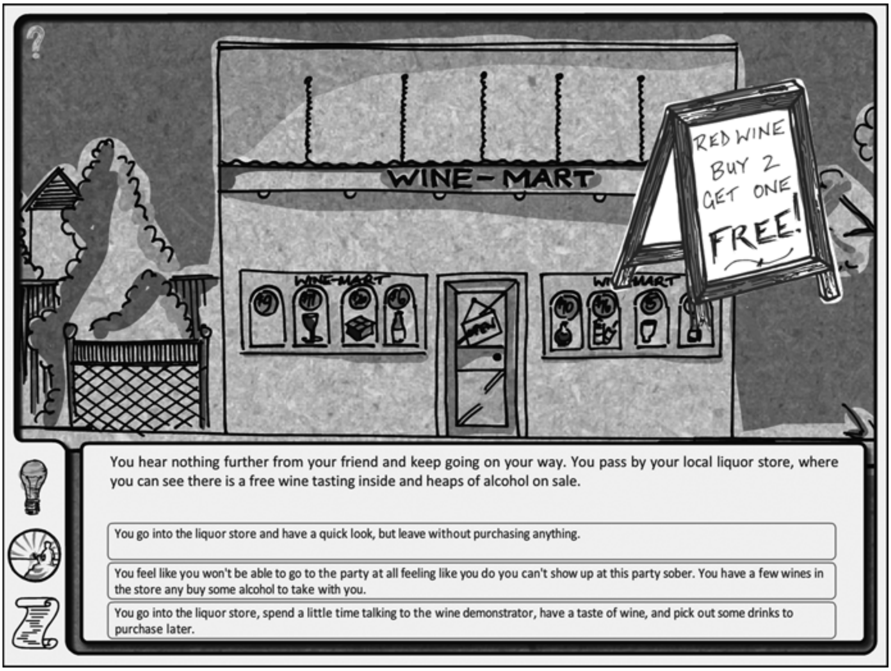
Mindfulness is a mediative technique that focuses on the now and steps away from the world to help deal with automatic thoughts and emotions.

Click to jump into **Mindfulness** when dialogue options in an Event are limited

Adjust the Mood by *Dragging and Dropping* Thoughts into their appropriate categories.

Thoughts have Categories.

Fig. 18.5 The instruction component of SHADOW, introducing game rules



You hear nothing further from your friend and keep going on your way. You pass by your local liquor store, where you can see there is a free wine tasting inside and heaps of alcohol on sale.

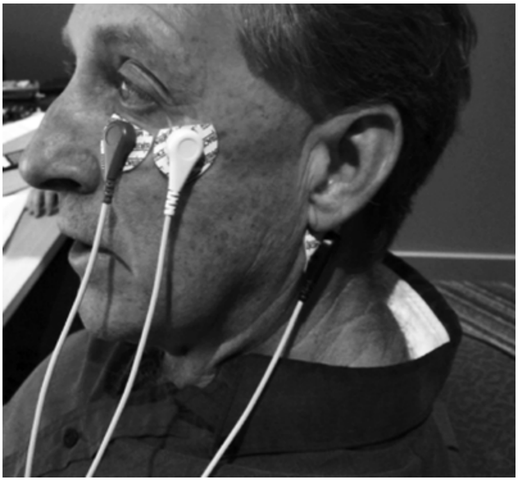
You go into the liquor store and have a quick look, but leave without purchasing anything.

You feel like you won't be able to go to the party at all feeling like you do you can't show up at this party sober. You have a few wines in the store any buy some alcohol to take with you.

You go into the liquor store, spend a little time talking to the wine demonstrator, have a taste of wine, and pick out some drinks to purchase later.

Fig. 18.6 A general gameplay scenario component of SHADOW

Fig. 18.7 Electrodes fitted for testing of the startle reflex



After completing an initial demographic survey, participants were fitted with three Kendall Medi-Trace Mini Ag/Ag electrodes (see Fig. 18.7) for surface EMG recording of activity in the orbicularis oculi. These particular electrodes were used rather than the standard 4 mm Ag/Ag electrodes commonly recommended

(Blumenthal et al., 2005) as the combined electrode/lead setup produced substantially less lead noise during player movement, with no discernible interference to the EMG signal capture. The fitting of the electrodes involved first cleaning the skin using disposable alcohol wipes where the electrodes were to be attached. One disposable ECG electrode was placed under the left eye, below the lower eyelid in line with the pupil in a forward gaze over the orbicularis oculi. A second electrode was also placed 1–2 cm laterally to the first, and a third "isolated ground" electrode was placed behind the ear (see Fig. 18.7). The electrodes wires were then connected to an ADInstruments Bio Amp (FE132) feeding into a PowerLab 8/35 for recording. Participants were asked to perform three voluntary blinks and the recording of these responses was checked to ensure they could be detected. Where required, the electrode attachment was adjusted until a reliable signal was being detected with voluntary blinks.

The basic premise and gameplay of SHADOW was explained to participants, who were then asked to play the SHADOW game without instruction for 12 min. The game was played in a standard computing set up: seated, with a keyboard and mouse on a desk surface, with a single monitor and sound delivered via speakers. While playing the game, 11 acoustic startle stimulus were presented using sudden acoustic white noise pulses containing frequencies in the range 20 Hz to 20 kHz at 105 dB, with a rise time of less than 10 μs. An Arcam FMJ Amplifier was used to convey the game sound and deliver the startle stimulus. The stimulus was presented at random intervals across the 12 min experiment period, at no less than 30 s apart. The gameplay was also recorded on video to confirm the locations in the game that the player was engaged when each startle stimulus was presented.

The EMG signal was recorded using an ADInstruments Bio Amp and PowerLab 8/35, in conjunction with Labchart 8 software. It was sampled at 1,000 Hz with a range of 500 μV. A low pass band filter at 50 Hz and a high pass filter at 0.3 Hz were used. Finally, a 50 Hz notch filter was applied, and the signal was inverted. The EMG response and the acoustic stimulation pulses were recorded in Labchart using Channels 2 and 4, respectively. A macro was used within Labchart to start the recordings simultaneously.

4.3 Results

At completion of the experiment, the recorded EMG signals were exported from Labchart in an appropriate format for data analysis in Matlab R2014b (8.4.0.150421) from Mathworks Inc. (2014). The raw EMG signal imported into Matlab was expressed in volts (V) at a sample rate of 1 ms. The raw signal was converted to (mV) to conform to reporting standards (Blumenthal et al., 2005; International Society of Electrophysiological Kinesiology, 1980). Following this, full wave rectification (absolute value) was applied to the biphasic signal.

A time interval of 150 ms before the time window of each startle acoustic pulse was used to derive a mean EMG baseline for each response. Individual baselines for each response were used to negate the effect that disturbance in the electrode lines might have over the duration of an experiment (Blumenthal et al., 2005; van Bedaf, Heesink, & Geuze, 2014), with 150 ms considered adequate for obtaining a reliable measurement (van Bedaf et al., 2014). The rectified EMG signal was also smoothed using a 30 sample moving average to allow for visual inspection.

The final step was to classify each startle response by visually inspecting the EMG data using the scoring approach detailed by Blumenthal et al. (2005). For each startle impulse, a decision was made as to whether: (a) the baseline period was contaminated with noise or movement artifact, or an involuntary blink, and thus a stimulated blink could not be quantified and the trial should be rejected, or (b) if a response within the 20–150 ms latency window occurred, and if not, the trial should be recorded as a nonresponse, or (c) a valid response occurred in a trial that has not been rejected and is thus scored as valid.

All rejected startles were excluded from the statistical analysis. The peak amplitudes for non-responses were recorded as zeroes for use in the analysis. For responses identified as valid, the rectified signal was processed to extract the peak amplitude in the period from 20 to 150 ms following the end of each startle acoustic stimulus. The corrected amplitude for each valid response was calculated by subtracting the mean EMG baseline amplitude from the raw peak amplitude. These values were recorded for each valid startle to be used in the final statistical analysis. The corrected peak amplitudes were compared as a measure of affective valence, with higher values indicative of unpleasant affect and lower values of pleasant affect (Figs. 18.8 and 18.9).

Finally, the video of each participant's interaction with the game was used to manually detect the game component (instruction, mindfulness, or scenario) related to each startle presentation. This information was used for grouping startle responses into three discrete treatment conditions for comparing in the statistical analysis (Fig. 18.10).

In total, 77 startle responses were recorded; 11 startle responses for each of the 7 subjects. Upon inspection, 14 % (11/77) were rejected due to signal noise or from being contaminated with responses that occurred prior to the startle stimulus. Of the valid responses, 95.4 % of the startle responses invoked a response, with 4.54 % (3/66) being judged as zero responses.

The player self-paced through the game and so the number of startles recorded in the three key sections varied. In total 24 startle responses were recorded in the instruction element, 18 in the scenario screens and 16 in the mindful challenge. Another four of the valid startles occurred in areas of the game that could not be clearly identified as one of the three key components and so these startles were also excluded from the statistical analysis.

The Instruction component recorded the highest mean amplitude ($M = 168.6$, SD = 4,551.9), the scenario screens recorded the next lowest mean amplitude ($M = 144.6$, SD = 11,204.7), and the mindful challenge recorded the lowest mean response amplitude ($M = 122.3$, SD = 1,510.6). There were large individual variations in the recorded amplitudes and this is reflected in high standard deviations.

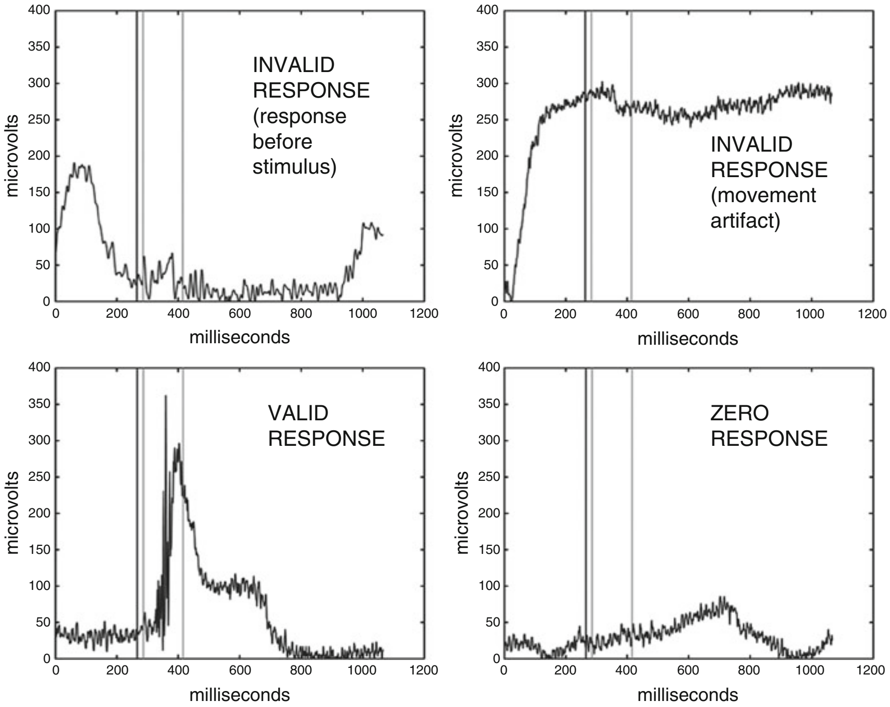
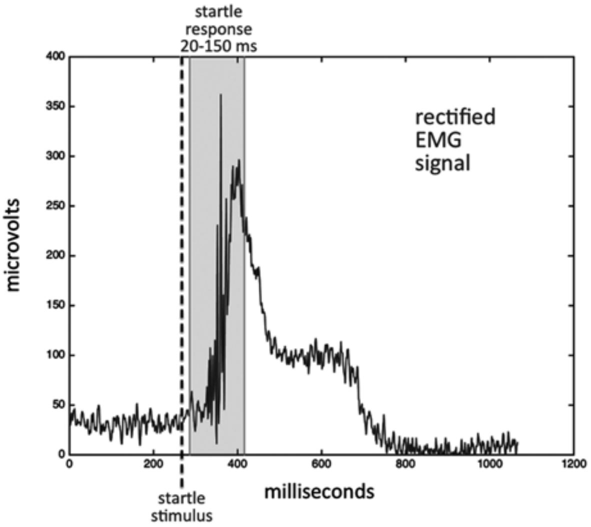


Fig. 18.8 Examples of valid and invalid startle responses as well as a zero response on a recorded EMG signal

Fig. 18.9 Rectified EMG signal showing the temporal relationship of a response to the startle impulse



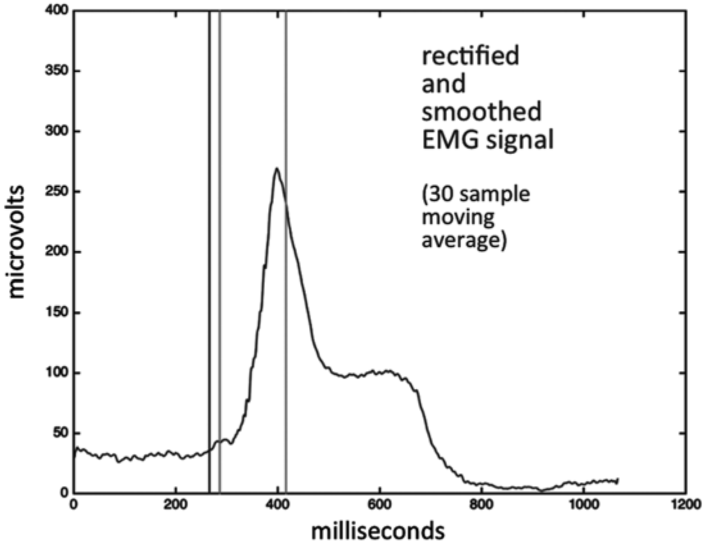


Fig. 18.10 Rectified and smoothed EMG signal (30 sample moving average)

We assumed unequal variances and then compared the means for the three conditions using a *t*-test. Mean peak amplitude (valence) was significantly different for the instruction and mindful conditions; $t(37)=2.03$, $p=.009$*. No significant difference was found between the instruction and the scenario component; $t(27)=2.05$, $p=.406$. No significant difference was found between the mindful and scenario components; $t(22)=2.07$, $p=.414$.

4.4 Discussion

The SHADOW game is being designed to engage players in developing two key skills of the SHADE program, namely, managing negative thoughts and mindfulness. In this study we examined the valence of three key components of the game, the instruction, mindfulness, and scenario elements. The study was intended as a formative evaluation to support the game design process and the intent was to measure the player's affective response to each of these elements. Our intention is that the most positive valence needs to be associated with the mindfulness challenge, as this is where critical new skills are imparted to players. This assumes that a positive affect supports improved learning, an assumption that is supported by previous research (Bless et al., 1996; Kanfer & Ackerman, 1989; Pekrun et al., 2002; Raghunathan & Trope, 2002).

The results indicate that the lowest amplitudes were recorded in the mindful mode, indicating that this component was associated with the most pleasant affect in players. A significant difference was found between the mindful challenge and the instructions section of the game. While these results are pleasing, some care must be taken when trying to interpret this outcome in terms of what it means for evaluating the design of the mindful challenge. The SHADOW case study has principally been provided to illustrate the use of the startle eye-blink measure in a game study. It should be remembered that the use of this study for serious game analytics is still in its infancy. For this reason, we will focus the discussion on the limitations of this study as they relate to general use of the startle eye-blink for assessing game designs and player affect in general.

Large individual variations between amplitude measures are common when using the startle response. This suggests larger sample sizes and more startle events are required. Some care also needs to be taken to ensure that startle responses occur both randomly, to avoid habituation, but also at the targeted design elements in the game. For example, in this study most startles occurred in the instructions section as participant's read about using the game. While the look and feel of the instruction component was consistent with the other game elements, it contained no actual gameplay. By contrast, the least number of startles occurred in the mindful component and this was the design element we were most interested in studying. This is indicative of the challenge of striving for both randomness and predictability in startle reflex experiments.

It is not possible to directly relate peak amplitude to a precise valence. This is particularly true as no real baseline of valence related to individual pleasantness or unpleasantness has been established. This lack of reference to valence is something that could be partially addressed by collecting suitable baseline data for each individual. This could be achieved by using an image library such as the International Affective Picture System (IAPS) that has been well-studied in relation to the startle response (Bradley & Lang, 2007). In future work, we intend to incorporate this feature into the preliminary part of our studies.

Indeed, most previous studies with studying emotional response with the startle reflex use static images (Bradley & Lang, 2007; Dan-Glauser & Scherer, 2011) or plain text (Witvliet & Vrana, 1995) in carefully controlled laboratory conditions. Games by contrast are highly interactive, partially random, and often dynamic environments. Indeed, one study has shown that active interaction with a virtual environment generates significantly different affect responses compared to a purely passive participation in the environment (Muehlberger et al., 2008). This suggests that the level of interactivity may also need to be considered when using the startle reflex measure. In our study we used a video recorder to also record the game play as a context to when the startle stimulus was presented. Importantly, the startle response quickly adapts to changing grades of pleasantness (Vrana et al., 1988), and although it is also studied in terms of workload (Neumann, 2002) and attention (Filion et al., 1998), it is generally considered independent from cognitive influence. This should make the startle eye-blink an ideal tool to quantify raw affective processing as it occurs while playing a game.

The startle reflex is used to measure valence. In terms of affective computing, most studies include an additional physiological measure such as skin conductance or heart rate to indicate arousal. This is because many consider the motivational basis of emotion using a very simple, two-factor model featuring affective valence and arousal (Lang, 1995). This dimensional theory of emotion holds that all emotions can be located on a two-dimensional space, as a function of valence and arousal (Ravaja et al., 2006).

5 Conclusion

Manipulating player emotions, whether it is for serious purposes or just to enhance general game experience, is a key responsibility for game designers. A distinguishing feature of the startle eye-blink measure is that it is has been frequently used as a way of objectively evaluating the positive and negative valence associated with a person's affective state. Importantly, the startle reflex is sensitive to affective processing and not to emotion. This means that it measures raw, basic affective responses, or in other words, the grade of negativity or positivity (or pleasantness) of any stimulus, situation, or environment a subject is exposed to. Eye-blink amplitude is reduced in the case of affective processing coding for pleasantness, whereas it is increased in the case of affective processing coding for unpleasant (Lang, Bradley, & Cuthbert, 1990, 1998).

Thus, the startle eye-blink suggests itself as an objective tool that can be adapted for assessing a player's affective response to various aspects of game design. Although the measure is relatively new to the serious games community, it has been well established in the field of psychology. However, in fairness many of these previous studies occur in well-controlled conditions with methods that may require some adaption for use with dynamic gaming environments. As a result, there is much need for further foundational work in applying this approach to game evaluation.

The startle response measure is not without some complexities, both in terms of collecting and processing captured data, and interpreting results. Indeed, the relationship between attention, emotion, and cognitive workload raises some complex issues in terms of game design, player perception, and cognition as well as their emotional state. For example, games relying on negative valence or stressful cognitive workloads to engage players are not easily translated to the common two-dimensional spaces used to explain valance and arousal.

Despite these complexities, we believe the startle reflex provides a useful adjunct to other approaches in assessing subconscious player responses to game elements. We also believe that this approach can successfully be adapted for assessing a player's emotional response to various aspects of game design and predict that the investigation of non-conscious information processing using the startle eye-blink will soon provide a useful new approach for analysing and improving serious game design.

References

Allen, N. B., Trinder, J., & Brennan, C. (1999). Affective startle modulation in clinical depression: Preliminary findings. *Biological Psychiatry, 46*(4), 542–550.

Alzoubi, O., Calvo, R. A., & Stevens, R. H. (2009). Classification of EEG for emotion recognition: An adaptive approach. In *Proceedings of 22nd Australasian Joint Conference on Artificial Intelligence* (pp. 52–61), Melbourne.

Andreassi, J. J. (2007). *Human behaviour and physiological response*. New York: Taylor & Francis.

Arroyo, I., Cooper, D. G., Burleson, W., Woolf, B. P., Muldner, K., & Christopherson, R. (2009). Emotion sensors go to school. *Proceedings of Artificial Intelligence in Education, 200*, 17–24.

Barrett, L. F., Mesquita, B., Ochsner, K. N., & Gross, J. J. (2007). The experience of emotion. *Annual Review of Psychology, 58*, 373.

Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J. (2006, April). Fully automatic facial action recognition in spontaneous behavior. In *7th International Conference on Automatic Face and Gesture Recognition, 2006. FGR 2006.* (pp. 223–230). IEEE.

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression*. New York: Guilford.

Bless, H., Clore, G., Schwarz, N., Golisano, V., Rabe, C., & Wolk, M. (1996). Mood and the use of scripts: Does a happy mood really lead to mindlessness? *Journal of Personality and Social Psychology, 71*, 665–679.

Blumenthal, T. D., Cuthbert, B. N., Filion, D. L., Hackley, S., Lipp, O. V., & Van Boxtel, A. (2005). Committee report: Guidelines for human startle eyeblink electromyographic studies. *Psychophysiology, 42*(1), 1–15.

Bradley, S. D. (2007). Examining the eyeblink startle reflex as a measure of emotion and motivation to television programming. *Communication Methods and Measures, 1*(1), 7–30.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Technical manual and affective ratings*. Gainesville, FL: The Center for Research in Psychophysiology, University of Florida.

Bradley, M. M., & Lang, P. J. (2007). The international affective picture system (IAPS) in the study of emotion and attention. In J. A. Coan & J. J. B. Allen (Eds.), *Handbook of emotion elicitation and assessment* (pp. 29–46). New York: Oxford University Press.

Calvo, R. A., Brown, I., & S. Scheding, S. (2009). Effect of Experimental factors on the recognition of affective mental states through physiological measures. In *Proceedings of 22nd Australasian Joint Conference on Artificial Intelligence*, Melbourne.

Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models. methods, and their applications. *IEEE Transactions on Affective Computing, 1*(1), 18–37.

Coan, J. (2010). Emergent ghosts of the emotion machine. *Emotion Review, 2*, 274.

Coppins, W. (2014). *Measuring the effect of sound on the emotional and immersive experience of players in a video game: A case study in the horror genre.* Honours thesis, The University of Newcastle, Australia.

D'Mello, S., & Graesser, A. (2009). Automatic detection of learner's affect from gross body language. *Applied Artificial Intelligence, 23*(2), 123–150.

Dalgleish, T., Dunn, B., & Mobbs, D. (2009). Affective neuroscience: Past present, and future. *Emotion Review, 1*, 355–368.

Dan-Glauser, E. S., & Scherer, K. R. (2011). The Geneva affective picture database (GAPED): A new 730-picture database focusing on valence and normative significance. *Behavior Research Methods, 43*(2), 468–477.

Darwin, C. (2002). *Expression of the emotions in man and animals*. New York: Oxford University Press.

Dawson, M. E., Schell, A. M., & Böhmelt, A. H. (1999). *Startle modification: Implications for neuroscience, cognitive science, and clinical science*. New York: Cambridge University Press.

Desmet, P., & Hekkert, P. (2007). Framework of product experience. *International Journal of Design, 1*(1), 2007.

Dimberg, U., Thunberg, M., & Elmehed, K. (2000). Unconscious facial reactions to emotional facial expressions. *Psychological Science, 11*(1), 86–89.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*, 169–200.

Ekman, P., & Friesen, W. (1969a). Nonverbal leakage and clues to deception. *Psychiatry, 32*, 88–106.

Ekman, P., & Friesen, W. V. (1969b). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica, 1*, 49–98.

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists.

Ekman, P., Levenson, R., & Friesen, W. (1983). Autonomic nervous system activity distinguishes among emotions. *Science, 221*, 1208–1210.

Elmore, W. R. (2012). *The effect of violent video game play on emotion modulation of startle*. Doctoral dissertation, University of Missouri–Kansas City, Kansas City, MO.

Evinger, C., & Manning, K. A. (1993). Pattern of extraocular muscle activation during reflex blinking. *Experimental Brain Research, 92*(3), 502–506.

Filion, D. L., Dawson, M. E., & Schell, A. M. (1993). Modification of the acoustic startle-reflex eyeblink: A tool for investigating early and late attentional processes. *Biological Psychology, 35*, 185–200.

Filion, D. L., Dawson, M. E., & Schell, A. M. (1998). The psychological significance of human startle eyeblink modification: A review. *Biological Psychology, 47*, 1–43.

Fokkinga, S. F., Desmet, P. M. A., & Hoonhout, J. (2010). The dark side of enjoyment: Using negative emotions to design for rich user experiences. In *Proceedings of the 7th International Conference of Design and Emotion Society*. Chicago, IL: Spertus Institute.

Fontaine, J. R., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological Science, 18*(12), 1050–1057.

Geiser, M., & Walla, P. (2011). Objective measures of emotion during virtual walks through urban environments. *Applied Sciences, 1*, 1–11.

Graham, F. K. (1975). The more or less startling effects of weak prestimulation. *Psychophysiology, 12*, 238–248.

Grahl, A., Greiner, U., & Walla, P. (2012). Bottle shape elicits gender-specific emotion: A startle reflex modulation study. *Psychology, 7*, 548–554.

Grimshaw, M., Lindley, C. A., & Nacke, L. (2008, October). Sound and immersion in the first-person shooter: mixed measurement of the player's sonic experience. In *Proceedings of Audio Mostly Conference*, pp. 1–7.

Gunes, H., & Piccardi, M. (2007). Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications, 30*, 1334–1345.

Hadley, M. J. (2012). *Slender: The Eight Pages* [PC game]. Parsec Productions.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. New York: Wiley.

Hoffman, H. S., & Fleshier, M. (1963). Startle reaction: Modification by background stimulation. *Science, 141*, 928–930.

Hookham, G., Deady, M., Kay-Lambkin, F., & Nesbitt, K. (2013) Training for life: Designing a game to engage younger people in a psychological counselling program. *Australian Journal of Intelligent Information Processing Systems, 13*(3): Special issue on Edutainment 2013.

International Society of Electrophysiological Kinesiology. (1980). *Units, terms and standards in the reporting of EMG research*. Carbondale, IL: Southern Illinois University School of Medicine.

Izard, C. (1994). Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin, 115*, 288–299.

James, W. (1884). What is an emotion? *Mind, 9*, 188–205.

Juslin, P. N., & Scherer, K. R. (2005). Vocal expression of affect. In J. A. Harrigan, R. Rosenthal, & K. R. Scherer (Eds.), *The new handbook of methods in nonverbal behavior research* (pp. 65–135). Oxford, MA: Oxford University Press.

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology, 74*, 657–690.

Kaviani, H., Gray, J. A., Checkley, S. A., Kumari, V., & Wilson, G. D. (1999). Modulation of the acoustic startle reflex by emotionally toned film-clips. *International Journal of Psychophysiology, 32*(1), 47–54.

Kaviani, H., Wilson, G. D., Checkley, S. A., Kumari, V., & Gray, J. A. (1998). Modulation of the human acoustic startle reflex by pleasant and unpleasant odors. *Journal of Psychophysiology, 12*, 352–361.

Kay-Lambkin, F., Baker, A. L., Kelly, B., & Lewin, T. J. (2011). Clinician-assisted computerised versus therapist-delivered treatment for depressive and addictive disorders: A randomised controlled trial. *Medical Journal of Australia, 195*(3), S44–S50.

Kemper, T. D. (1991). Predicting emotions from social-relations. *Social Psychology Quarterly, 54*, 330–342.

Lang, P. J. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist, 50*(5), 372–385.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological Review, 97*, 377–395.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1998). Emotion, motivation, and anxiety: Brain mechanisms and psychophysiology. *Biological Psychiatry, 44*, 1248–1263.

Lang, P. J., Greenwald, M. K., Bradley, M. M., & Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology, 30*, 261–273.

MATLAB and Statistics Toolbox Release R2014b. (2014). Natick, MA: The MathWorks, Inc.

Mavratzakis, A., Molloy, E., & Walla, P. (2013). Modulation of the startle reflex during brief and sustained exposure to emotional pictures. *Psychology, 4*, 389–395.

McDaniel, B. T., D'Mello, S., King, B. G., Chipman, P., Tapp, K., & Graesser, A. C. (2007). Facial features for affective state detection in learning environments. In *Proceedings of the 29th Annual Cognitive Science Society* (pp. 467–472). Austin, TX: Cognitive Science Society.

Meyer, D. K., & Turner, J. C. (2006). Re-conceptualizing emotion and motivation to learn in classroom contexts. *Educational Psychology Review, 18*(4), 377–390.

Mota, S., & Picard, R. W. (2003). Automated posture analysis for detecting learner's interest level. In *Computer Vision and Pattern Recognition Workshop, 2003* (Vol. 5, p. 49). IEEE.

Muehlberger, A., Wieser, M. J., & Pauli, P. (2008). Darkness-enhanced startle responses in ecologically valid environments: A virtual tunnel driving experiment. *Biological Psychology, 77*, 47–52. doi:10.1016/j.biopsycho.2007.09.004.

Nacke L. (2009). *Affective Ludology: Scientific measurement of user experience in interactive entertainment*. Blekinge Institute of Technology, Game Systems and Interaction Research Laboratory, School of Computing, Blekinge Institute of Technology, Doctoral Dissertation Series No 2009:04.

Nasoz, F., Alvarez, K., Lisetti, L., & Finkelstein, N. (2004). Emotion recognition from physiological signals using wireless sensors for presence technologies. *Cognition, Technology and Work, 6*(1), 4–14.

Neumann, D. L. (2002). Effect of varying levels of mental workload on startle eyeblink modulation. *Ergonomics, 45*(8), 583–602.

Osgood, C. E., May, W. H., & Miron, M. S. (1975). *Cross-cultural universals of affective meaning*. Urbana, IL: University of Illinois Press.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*, 1–135.

Pantic, M., & Rothkrantz, L. (2003). Toward an affect-sensitive multi-modal human-computer interaction. *Proceedings of the IEEE, 91*(9), 1370–1390.

Parsons, T. D., Rizzo, A. A., Courtney, C. G., & Dawson, M. E. (2012). Psychophysiology to assess impact of varying levels of simulation fidelity in a threat environment. *Advances in Human-Computer Interaction, 5*, 1–9.

Patrick, C. J., Bradley, M. M., & Lang, P. J. (1993). Emotion in the criminal psychopath: Startle reflex modulation. *Journal of Abnormal Psychology, 102*(1), 82–92.

Pekrun, R., Goetz, T., Titz, W., & Perry, R. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist, 37*(2), 91–105.

Picard, R. W. (1997). *Affective computing*. Cambridge, MA: MIT Press.

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(10), 1175–1191.

Raghunathan, R., & Trope, Y. (2002). Walking the tightrope between feeling good and being accurate: Mood as a resource in processing persuasive messages. *Journal of Personality and Social Psychology, 83*, 510–525.

Ramirez, O. M., & Dockweiler, C. J. (1987). Mathematics anxiety: A systematic review. In R. Schwarzer, H. M. Ploeg, & C. D. Spielberger (Eds.), *Advances in test anxiety research* (pp. 157–175). Hillsdale, NJ: Erlbaum.

Ravaja, N., & Kivikangas, J. M. (2008, August). Psychophysiology of digital game playing: The relationship of self-reported emotions with phasic physiological responses. In *Proceedings of Measuring Behavior* (pp. 26–29), Maastricht, The Netherland.

Ravaja, N., Saari, T., Salminen, M., Laarni, J., & Kallinen, K. (2006). Phasic emotional reactions to video game events: A psychophysiological investigation. *Media Psychology, 8*(4), 343–367.

Ravaja, N., Turpeinen, M., Saari, T., Puttonen, S., & Keltikangas-Järvinen, L. (2008). The psychophysiology of James Bond: Phasic emotional responses to violent video game events. *Emotion, 8*(1), 114.

Reisenzein, R., Hudicka, E., Dastani, M., Gratch, J., Hindriks, K. L., & Meyer, J. (2013). Computational modeling of emotion: Toward improving the inter- and interdisciplinary exchange. *IEEE Transactions on Affective Computing, 4*(3), 246–266.

Roy, M., Mailhot, J. P., Gosselin, N., Paquette, S., & Peretz, I. (2009). Modulation of the startle reflex by pleasant and unpleasant music. *International Journal of Psychophysiology, 71*(1), 37–42.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*, 145–172.

Russell, J. A., Bachorowski, J. A., & Fernandez-Dols, J. M. (2003). Facial and vocal expressions of emotion. *Annual Review of Psychology, 54*, 329–349.

Sabourin, J. L., & Lester, J. C. (2014). Affect and engagement in game-based learning environments. *IEEE Transactions on Affective Computing, 5*(1), 45–56.

Sabourin, J., Rowe, J. P., Mott, B. W., & Lester, J. C. (2011). When off-task is on-task: The affective role of off-task behavior in narrative-centered learning environments. In *Proceedings of 15th International Conference on Artificial Intelligence in Education* (pp. 523–536).

Salovey, P. (2003). Introduction: Emotion and social processes. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 747–751). Oxford, UK: Oxford University Press.

Sasse, D. (2008). *A framework for psychophysiological data acquisition in digital games*. Master's thesis, Otto-von-Guericke-University Magdeburg, Magdeburg, Germany.

Scherer, K. R. (1993). Neuroscience projections to current debates in emotion psychology. *Cognition and Emotion, 7*(1), 1–41.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information, 44*(4), 695–729.

Scherer, K. R., Schorr, A. E., & Johnstone, T. E. (2001). *Appraisal processes in emotion: Theory, methods, research*. Oxford, UK: Oxford University Press.

Schneirla, T. C. (1959). An evolutionary and developmental theory of biphasic processes underlying approach and withdrawal. In M. R. Jones (Ed.), *Nebraska symposium on motivation* (Vol. 7, pp. 1–43). Lincoln, NE: University of Nebraska Press.

Sebe, N., Cohen, I., & Huang, T. S. (2005). Multimodal emotion recognition. *Handbook of Pattern Recognition and Computer Vision, 4*, 387–419.

Segal, Z., Williams, J. M. G., & Teasdale, J. D. (2002). *Mindfulness-based cognitive therapy for depression: A new approach to preventing relapse*. New York: Guilford.

Strapparava, C., & Valitutti, A. (2004). WordNet affect: An affective extension of WordNet. *LREC, 4*, 1083–1086.

Swerdlow, N. R., Weber, M., Qu, Y., Light, G. A., & Braff, D. L. (2008). Realistic expectations of prepulse inhibition in translational models for schizophrenia research. *Psychopharmacology, 199*(3), 331–388.

Valsamis, B., & Schmid, S. (2011). Habituation and prepulse inhibition of acoustic startle in rodents. *Journal of Visualized Experiments, 55*, 3446. doi:10.3791/3446.

van Bedaf, L. R., Heesink, L., & Geuze, E. (2014, August 27–29). Pre-processing of electromyography startle data: A novel semi-automatic method. In: *Proceedings of Measuring Behavior*, Wageningen, The Netherlands.

Villon, O., & Lisetti, C. (2006). A user-modeling approach to build user's psycho-physiological maps of emotions using bio-sensors. In *Proceedings of IEEE RO-MAN 2006, 15th IEEE International Symposium on Robot and Human Interactive Communication, Session Emotional Cues in Human-Robot Interaction* (pp. 269–276).

Vrana, S. R., Spence, E. L., & Lang, P. J. (1988). The startle probe response: A new measure of emotion? *Journal of Abnormal Psychology, 97*, 487–491.

Vyzas E., & Picard, R. W. (1998). Affective pattern classification. In: *Proceedings of AAAI Fall Symposium Series: Emotional and Intelligent: The Tangled Knot of Cognition* (pp. 176–182), Orlando, FL.

Walla, P., Brenner, G., & Koller, M. (2011). Objective measures of emotion related to brand attitude: A new way to quantify emotion-related aspects relevant to marketing. *PLoS One, 6*(11), e26782. doi:10.1371/journal.pone.0026782.

Walla, P., & Panksepp, J. (2013). Neuroimaging helps to clarify brain affective processing without necessarily clarifying emotions. In K. N. Fountas (Ed.), *Novel Frontiers of Advanced Neuroimaging*. InTech. ISBN: 978-953-51-0923-5, doi:10.5772/51761.

Walla, P., Richter, M., Färber, S., Leodolter, U., & Bauer, H. (2010). Food evoked changes in humans: Startle response modulation and event-related brain potentials (ERPs). *Journal of Psychophysiology, 24*, 25–32.

Witvliet, C. V., & Vrana, S. R. (1995). Psychophysiological responses as indices of affective dimensions. *Psychophysiology, 32*(5), 436–443.

Wood, R. T., Griffiths, M. D., Chappell, D., & Davies, M. N. (2004). The structural characteristics of video games: A psycho-structural analysis. *CyberPsychology & Behavior, 7*(1), 1–10.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(1), 39–58.

Chapter 19
Using Pattern Matching to Assess Gameplay

Rodney D. Myers and Theodore W. Frick

Abstract In this chapter we describe Analysis of Patterns in Time (APT) and how it can be used to analyze gameplay choices to provide evidence of a play-learner's understanding of concepts modeled in a game. APT is an empirical approach to observing and coding phenomena as mutually exclusive and exhaustive categories within classifications. These data form a temporal map of joint and sequential patterns. We examine the case of the online *Diffusion Simulation Game*. An algorithm calculates scores for gameplay data patterns and compares them with scores for patterns based on optimal strategies derived from the game's conceptual model. We discuss the results of using APT for analysis of game sessions for three play-learners. We describe how APT can be included as part of a serious game to conduct formative assessment and determine appropriate hints, coaching, or other forms of scaffolding during gameplay. We conclude by discussing APT methods for summative assessment.

Keywords Pattern matching • Gameplay strategies • Assessment • Models • Scaffolding

1 Introduction

In this chapter, our goal is to illustrate the potential of Analysis of Patterns in Time (APT) as a way of measuring and analyzing play-learner interactions with a serious game, the *Diffusion Simulation Game* (DSG). First, we discuss APT and provide examples of a temporal map and APT queries. Next, we provide a brief overview of diffusion of innovations theory (DoI) and a description of the DSG. We then describe

R.D. Myers (✉)
Unlock Learning, 1726 East Thornton Drive, Bloomington, IN 47401, USA
e-mail: rod@webgrok.com

T.W. Frick
Indiana University, 201 North Rose Avenue, Bloomington, IN 47405-1006, USA
e-mail: frick@indiana.edu

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based Learning, DOI 10.1007/978-3-319-05834-4_19

our procedure for applying APT to DSG play-learner data; we analyze temporal maps of multiple DSG games played by three different play-learners of varying proficiency, in order to illustrate how APT can detect patterns of play-learner moves and determine how consistent those patterns of play are with expert strategies based on DoI theory.

Finally, we discuss the potential of APT to measure what play-learners are learning over time as they interact with a simulation or game, and how such pattern analysis could be used by an intelligent agent in the game to determine appropriate hints, coaching, or other forms of scaffolding during gameplay to improve learning and performance.

2 Overview of *MAPSAT*: Map & Analyze Patterns & Structures Across Time

MAPSAT is a different approach to measurement and analysis of data, when compared to traditional methods. Compare these two sets of findings:

(a) MAPSAT: Students in elementary schools are about 13 times more likely to be off-task during non-interactive classroom instruction, when compared with their engagement during interactive instruction.
(b) Linear Models Approach (LMA): The amount of interactive classroom instruction predicts 32 % of the variance in student task engagement, leaving 68 % of the variance unexplained.

These results are based on the *same* classroom observation data (see Frick, 1990). What's the difference? The short answer: MAPSAT *measures the relation*. The LMA *relates the measures*.

We first discuss the traditional methods, which should be familiar to most readers. Next we address the theoretical background of MAPSAT, why and how it is different from the LMA, and why it is theoretically impossible to derive *a* from *b* above. We conclude with an example of a specific temporal map and then illustrate APT queries for counting patterns.

2.1 Traditional Quantitative Methods of Measurement and Analysis of Data

In traditional quantitative research methods that are based on algebraic linear models, we typically obtain *separate measures of variables*, and then statistically analyze relations among measures (e.g., linear, curvilinear, or logistic regression analysis). That is, we *relate measures*. This approach, which assumes linear and additive models, can result in aggregation aggravation—that is, obfuscation of important relationships due to assumptions in this approach (Frick, 1983, 1990; Frick, Myers, Thompson, & York, 2008).

Table 19.1 Example of a typical spreadsheet for traditional quantitative analyses

Case	Gender	Height in inches	Weight in pounds
1	Male	70	200
2	Female	60	120

In traditional measurement we aggregate units when we obtain a value for a variable. For example, we aggregate (count) the number of inches when we measure a person's height, or we count the number of pounds when we measure someone's weight. We repeat this process of *independent* aggregations for more persons' heights and weights. Then we attempt a statistical analysis of these sets of independent measures, such as correlation or linear regression. This kind of thinking stems from algebra, for example, $y=Bx+C$, where variable y is measured separately from variable x, and a *functional relationship is assumed* to exist between x and y, where B is the slope and C is a constant.

Specifically, imagine a spreadsheet of data. Normally each row in the spreadsheet would contain data on a single case, columns are for variable names, and in each cell a value for each variable is entered. See Table 19.1 for an example.

Notice that a variable has a *single* value for each case, and these values are typically determined by separate measures for each case. In order to determine a relation between two variables, we would do a correlational analysis such as a Pearson Product Moment Correlation. This would be a statistical relation between separate measures, for example, between a person's height and his/her weight. It could also be an analysis of variance (ANOVA) to determine a statistical relationship between values of gender and height. Or we might perform a multiple regression analysis in order to predict a person's weight from knowledge of a person's gender and height.

2.2 *Theoretical Foundations of MAPSAT Methods*

While investigating the SIGGS theory model (Maccia & Maccia, 1966), Frick discovered that the measures of uncertainty in information theory were inadequate for predicting specific temporal patterns (Frick, 1983). SIGGS is grounded in *set* (S), *information* (I), *di-graph* (G), and *general systems* (GS) theories. SIGGS is a complex theory model with precise definitions of systems' dynamic and structural properties such as toput, strongness, adaptability, stress, wholeness, and so forth. SIGGS was used to develop a theory of education, consisting of 201 hypotheses. Space does not permit further description here. See https://www.indiana.edu/~tedfrick/siggs.html.

In SIGGS, *information* is defined as a "characterization of occurrences" (Maccia & Maccia, 1966, p. 40), and in turn is further defined mathematically via set theory and probability theory (pp. 10–23, 40–53). Frick (1983) interpreted these occurrences as temporal events, characterized by classifications and categories used when observing empirical phenomena.

Determination of values of SIGGS properties of *feedin*, *feedout*, *feedthrough,* and *feedback* requires measures of temporal patterns. More specifically, *feedin* is defined as transmission of information (occurrences of elements) from *toput* at *time 1* to *input* at *time 2*. For example, the distribution of students who *apply* to various degree programs at a university in the spring are part of *toput*, and those students who are subsequently *admitted and attend* in the fall then become part of the *input* distribution of students in those degree programs in that particular education system. Similarly, *feedthrough* is defined in SIGGS as *feedin* followed later by *feedout*. For example, students matriculate (*feedin*), and later they graduate, drop out, or flunk out (*feedout—fromput* followed by *output*); this entire set of trajectories constitutes that system's student *feedthrough*.

In set theory, a *relation* is the Cartesian Product of two or more sets of elements. Such a relation consists of a set of ordered pairs of elements, or more generally, *tuples*. Each *n*-tuple characterizes a pattern—that is, a conjoining of elements. For example, a 4-tuple characterizes the *feedthrough* of a particular student from *toput* at *time 1*, to *input* at *time 2*, to *fromput* at *time 3*, to *output* at *time 4*. One student might apply to a university music program (*toput*), be admitted as a music major (*input*), later change her major, completing a bachelor's degree in computer science (*fromput*), then get a good-paying job as a software engineer after graduation (*output*). Another 4-tuple is characterized by a different student who applies for a computer science major, but instead gets admitted to a general studies program, later leaves the university with no degree, and then is employed in a low-paying job.

When occurrences of students moving through the university are mapped into categories of classifications which represent 4-tuples, a *joint probability distribution* can be formed (from the Cartesian Product of *toput*, *input*, *fromput*, and *output* classifications which determine student *feedthrough* for the university). However, the *T* and *B* measures from information theory (Coombs, Dawes, & Tversky, 1970; Maccia & Maccia, 1966) do not provide specific predictions of temporal patterns (or trajectories); rather *T* and *B* coefficients are *measures of overall uncertainty* in the joint probability distributions of temporal occurrences. This is analogous to how an *F*-test in ANOVA indicates overall statistical significance, but does not tell us which contrasts are significant when there are more than two group means being compared.

Moreover, Frick (1983) subsequently proved mathematically that marginals (e.g., *toput*, *input*, *fromput*, *output*) of joint probability distributions cannot dependably predict cell values, that is, probabilities of conjoint occurrences of temporal events (e.g., *feedin*, *feedout*, *feedthrough*, *feedback*). He concludes:

> There is no unique solution to this set of equations [18–21, from the calculus of probability theory], since the determinate of the matrix of coefficients is zero.… The mathematical conclusion is that there is no way to uniquely determine the joint probability distribution given only the marginal probability distributions, except in a few special cases where the marginal probabilities are zeros and ones, or all equal. (p. 79)

Hence, the need for alternative methods was justified theoretically. APT was invented as such an alternative approach, which has been further developed into MAPSAT in recent years. Frick (1983, 1990) emphasized that the traditional approach taken to measurement in the LMA only uses marginal distributions, wherein variables are measured separately and then their relationships are estimated by statistical analysis (see Sect. 2.1).

2.3 Pros and Cons of MAPSAT Methods

The primary advantage of using MAPSAT methods is that *researchers can detect relations* (*temporal or structural patterns*) *that cannot be revealed by the linear models approach*. This is because the LMA assumes a *functional* relationship between two or more variables that are measured separately. MAPSAT methods do not assume *functional* relations—that is, algebraic equations which are mathematical functions in the set-theoretic sense. In set theory, the difference between a *relation* and a *function* is clearly defined (e.g., see Coombs et al., 1970, pp. 361–371).

The primary disadvantage of using MAPSAT methods is that *most researchers will first need to learn how to use them appropriately*. This is analogous to how one must learn about traditional measurement and statistics in order to use ANOVA, MANOVA, and linear/logistic regression methods. On the other hand, MAPSAT methods are much easier to learn and understand, since no complex mathematics, algebra, or statistics is required.

Use of MAPSAT methods requires a different approach to measurement of relations, since temporal and structural patterns are measured directly through observation of empirical phenomena. This requires development of a well-defined coding scheme that is related to research questions of interest. Then human observers must be trained to use the coding scheme. Subsequently, they must observe and code empirical phenomena to obtain the temporal or structural maps needed for addressing research questions. Human judgment is normally required in order to discriminate phenomena observed and to use well-defined classifications and their respective categories when creating temporal or structural maps. This requires quantitative and performative intelligence, and in particular *instantial* "knowing that" and *performative* "knowing how" (see Frick, 1997, pp. 111–115). If such discrimination and skill can be accomplished by computers and related technologies, then software could be written which can classify and categorize empirical phenomena to create such maps—if this is possible and can be done reliably.

MAPSAT does not directly inform a researcher which patterns are highly predictable or not. Such patterns may be anticipated from theoretical expectations or research questions, or they may be discovered serendipitously by visual examination of temporal maps. MAPSAT queries of temporal maps must be performed in order to get measures of temporal relations, such as conditional probabilities of patterns or proportion time.

MAPSAT pattern *results* can be used with quantitative research methods such as the LMA so that generalizations can be made from a sample to a population (Frick, 1990). We illustrate MAPSAT for several cases in this chapter. Space does not permit illustration of inferential statistics with MAPSAT here. See Frick et al. (2008) for descriptions of research studies using MAPSAT methods and statistical inference.

Finally, use of the LMA is appropriate when the goal of research is to discover or verify functional relationships—that is, characterized by algebraic equations. MAPSAT is appropriate for research whose goals are to discover or verify patterns that are ones that are not perfectly predictable (i.e., stochastic), in contrast to deterministic patterns where there is no uncertainty. See Frick (1983, 1990) for an in-depth discussion.

2.4 MAPSAT Methods

In MAPSAT, we *measure relations* directly. This is not a play on words, but a significant paradigm change in conceptualizing research problems and how we collect and analyze data: *map relations* instead of *measure variables*, and then *analyze relation maps* instead of *statistically associating variables*. We call this alternative approach MAPSAT: Map & Analyze Patterns & Structures Across Time.

MAPSAT yields results from analysis of occurrences of categorical relations (i.e., *n*-tuples from a Cartesian Product in set theory), not a statistical analysis of separate measures of variables, results from which might yield a correlation coefficient or regression equation for describing a relationship. In MAPSAT, there are two approaches that can be taken. In the *Analysis of Patterns in Time* (APT) approach, we map *temporal* relations. In the *Analysis of Patterns in Configuration* (APC) approach, we construct a map of structural relations, called *affect-relations*, in a system.

Dynamic Bayesian Network Analysis (DBNA) is similar to APT (cf. Jensen & Nielsen, 2007). However, APT methods differ from DBNA in that *Bayes Theorem is not assumed in APT* nor used in computing conditional probabilities; rather relative frequencies of temporal sequences or proportion of time determine APT conditional probabilities. There are other differences as well, particularly concerning assumptions about measurement itself. For an in-depth discussion of differences among APT, Bayesian reasoning, and the Linear Models Approach, see Frick (1983, 1990). For brief descriptions of examples of empirical research studies that use APT methods, see Frick et al. (2008).

2.5 Fundamentals of APT

In APT we create a *temporal map* as the *basic unit of measure*. So, instead of putting a single value of a variable in a cell of a spreadsheet, imagine that *each spreadsheet cell contains another spreadsheet*. What is a temporal map in APT? Table 19.2 illustrates a temporal map that might be created by an amateur meteorologist.

There are 18 joint temporal events (JTEs) in the temporal map in Table 19.2. Each joint event is coded at some point in time. Cells in column two contain information about the Unix Epoch Time (elapsed seconds since Jan. 1, 1970), as well as the duration of the joint event (in seconds). There are five classifications indicated by

Table 19.2 Temporal map from observation and coding of weather events, adapted from Frick (1990)

JTE	Epoch time: duration of JTE	Season	Air temperature (°F)	Barometric pressure (p.s.i.)	Precipitation	Cloud structure
1	1,417,436,508: dur.=1,470	{Fall	{33	{Above 30	{Null	{Cirrus
2	1,417,437,978: dur.=2,277	|	|	{Below 30	|	|
3	1,417,440,255: dur.=2,554	|	|	|	|	{Nimbus stratus
4	1,417,442,809: dur.=794	|	|	|	{Rain	|
5	1,417,443,603: dur.=1,095	|	{32	|	|	|
6	1,417,444,698: dur.=477	|	|	|	{Sleet	|
7	1,417,445,175: dur.=721	|	{31	|	|	|
8	1,417,445,896: dur.=1,026	|	|	|	{Snow	|
9	1,417,446,922: dur.=1,207	|	{32	|	|	|
10	1,417,448,129: dur.=410	|	{33	|	|	|
11	1,417,448,539: dur.=442	|	|	|	{Sleet	|
12	1,417,448,981: dur.=738	|	{34	|	|	|
13	1,417,449,719: dur.=2,647	|	|	|	{Rain	|
14	1,417,452,366: dur.=1,325	|	|	|	{Null	|
15	1,417,453,691: dur.=157	|	|	{Above 30	|	|
16	1,417,453,848: dur.=780	|	{35	|	|	|
17	1,417,454,628: dur.=1,464	|	|	|	|	{Null
18	1,417,456,092: dur.=1	|	{36	|	|	|

This entire temporal configuration of event occurrences would be inserted into *one cell* in a spreadsheet and would replace a single cell value as illustrated in Table 19.1

columns: season of year, air temperature in degrees Fahrenheit, barometric pressure, precipitation, and cloud structure.

Each singular temporal event (STE) is indicated in a cell. Every STE has associated with it the time it was coded, an event state (where a { indicates that there is a change in the classification value from what was coded earlier, and a | means that the previously coded event is continuing). For example, in JTE 4, precipitation changes to rain ({rain), while season continues to be fall, temperature continues to be 33°, barometric pressure continues to be below 30 pounds per square inch (p.s.i.), and cloud structure continues as nimbus-stratus. At JTE 6, precipitation changes to sleet, while the states of the other classifications continue.

Classifications consist of mutually exclusive and exhaustive event value designations. For example, if precipitation is rain, then it cannot be sleet or snow at that point in time when observing weather on Dec. 1, 2014, at a specific location. The null value means that there is nothing relevant to the classification that can be coded at that point in time. Event values can be categories (nominal), ranks (ordinal), whole numbers (interval), or decimal numbers (ratio).

2.6 Examples of Patterns and Associated Queries in APT

An APT query specifies a temporal pattern and returns results of matches found in the temporal map. This is what we mean by *measuring a relation* in APT. Results are reported below for both duration and frequency of pattern instances found in the temporal map illustrated in Table 19.2.

Pattern 1: APT Query for a 2-Phrase Sequential Pattern

WHILE the FIRST Joint Temporal Event is true (Phrase 1):
Season of Year is in state *starting or continuing*, value = *Fall*
Barometric Pressure is in state *starting or continuing*, value = *Below 30*
Cloud Structure is in state *starting or continuing*, value = *Nimbus Stratus*

- Duration when Phrase 1 is True = 13,436 s (out of 19,584 s total). Proportion of Time = 0.68607.
- Joint Event Frequency when Phrase 1 is True = 12 (out of 18 total joint temporal events). Proportion of JTEs = 0.66667.

THEN while the NEXT Joint Temporal Event is true (Phrase 2):
Season of Year is in state *starting or continuing*, value = *Fall*
Barometric Pressure is in state *starting or continuing*, value = *Below 30*
Precipitation is in state *starting or continuing*, value = *Rain*
Cloud Structure is in state *starting or continuing*, value = *Nimbus Stratus*

- Duration when Phrase 2 is True = 4,086 s (out of 19,584 s total), given all prior phrases are true. Proportion of Time = 0.20864.
- Joint Event Frequency when Phrase 2 is True = 3 (out of 18 total joint temporal events), given all prior phrases are true. Proportion of JTEs = 0.16667.
- Conditional joint event *duration* when Phrase 2 is true, given all prior phrases are true = 0.30411 (4,086 out of 13,436 s (time units)).
- Conditional joint event*frequency* when Phrase 2 is true, given all prior phrases are true = 0.25000 (3 out of 12 joint temporal events).

This is a 2-phrase APT query for Pattern 1. Each phrase specifies the conditions which must be true for that phrase to be true (a match) in the temporal map. Furthermore, the second phrase will *not* be considered a match in the map unless (a) it occurs *after* the first phrase becomes true and (b) all conditions in both the first *and* second phrases remain true in the map. Based on the observations coded in the map in Table 19.2, the proportion of time that precipitation was rain is 0.304, given that it was first true that the season was fall, the barometric pressure was below 30 p.s.i. and cloud structure was nimbus stratus. Another way of stating this is that the likelihood of rain occurring at some point in time was 0.304 under these prior conditions.

Pattern 2: APT Query for a 4-Phrase Sequential Pattern

WHILE the FIRST Joint Temporal Event is true (Phrase 1):
Cloud Structure is in state *starting or continuing*, value=*Nimbus Stratus*

- Duration when Phrase 1 is True=14,373 s (out of 19,584 s total). Proportion of Time=0.73392.
- Joint Event Frequency when Phrase 1 is True=14 (out of 18 total joint temporal events). Proportion of JTEs=0.77778.

THEN while the NEXT Joint Temporal Event is true (Phrase 2):
Barometric Pressure is in state *starting or continuing*, value=*Below 30*
Cloud Structure is in state *starting or continuing*, value=*Nimbus Stratus*

- Duration when Phrase 2 is True=12,111 s (out of 19,584 s total), given all prior phrases are true. Proportion of Time=0.61841.
- Joint Event Frequency when Phrase 2 is True=11 (out of 18 total joint temporal events), given all prior phrases are true. Proportion of JTEs=0.61111.
- Conditional joint event *duration* when Phrase 2 is true, given all prior phrases are true=0.84262 (12,111 out of 14,373 s) (time units).
- Conditional joint event *frequency* when Phrase 2 is true, given all prior phrases are true=0.78571 (11 out of 14 joint temporal events).

THEN while the NEXT Joint Temporal Event is true (Phrase 3):
Air Temperature is in state *starting or continuing*, value<=*32*
Barometric Pressure is in state *starting or continuing*, value=*Below 30*
Precipitation is in state *starting or continuing*, value=*Sleet*
Cloud Structure is in state *starting or continuing*, value=*Nimbus Stratus*

- Duration when Phrase 3 is True=1,889 s (out of 19,584 s total), given all prior phrases are true. Proportion of Time=0.09646.
- Joint Event Frequency when Phrase 3 is True=2 (out of 18 total joint temporal events), given all prior phrases are true. Proportion of JTEs=0.11111.
- Conditional joint event *duration* when Phrase 3 is true, given all prior phrases are true=0.15597 (1,889 out of 12,111 s (time units).
- Conditional joint event *frequency* when Phrase 3 is true, given all prior phrases are true=0.18182 (2 out of 11 joint temporal events).

THEN while the NEXT Joint Temporal Event is true (Phrase 4):
Air Temperature is in state *starting or continuing*, value<=*31*
Barometric Pressure is in state *starting or continuing*, value=*Below 30*
Precipitation is in state *starting or continuing*, value=*Snow*
Cloud Structure is in state *starting or continuing*, value=*Nimbus Stratus*

- Duration when Phrase 4 is True=1,095 s (out of 19,584 s total), given all prior phrases are true. Proportion of Time=0.05591.
- Joint Event Frequency when Phrase 4 is True=1 (out of 18 total joint temporal events), given all prior phrases are true. Proportion of JTEs=0.05556.

- Conditional joint event *duration* when Phrase 4 is true, given all prior phrases are true=0.57967 (1,095 out of 1,889 s (time units)).
- Conditional joint event *frequency* when Phrase 4 is true, given all prior phrases are true=0.50000 (1 out of 2 joint temporal events).

This 4-phrase query for Pattern 2 is more complex. First, cloud structure becomes nimbus stratus; then second, barometric pressure becomes less than 30 p.s.i.; then third, air temperature becomes less than or equal to 32° and precipitation becomes sleet; then fourth, air temperature becomes less than or equal to 31° and precipitation becomes snow. The likelihood of the fourth phrase being true is 0.58, given that the first three phrases become true in the order specified and remain true.

Space does not permit description of matching and counting algorithms in APT. Nonetheless, it should be clear that complex combinations of events and event sequences can be counted by querying temporal maps.

The results of these two queries could be put into a spreadsheet, as can be seen in Table 19.3, which shows the pattern probabilities for three different temporal maps (maps 2 and 3 not shown here). The *pattern specified* in the query becomes the *variable* and the results of the APT measure of the pattern become the *value* that could be put into a spreadsheet cell in SPSS or Excel. One can, for example, then compute means and standard deviations on APT query results for each pattern and perform other statistical analyses of these pattern measures. For example, the statistical correlation between measures of Pattern 1 and 2 from these three temporal maps is highly negative (–0.86, meaning the *higher* the probability of Pattern 1 [when nimbus stratus clouds and p.s.i. <30, then rain follows], the *lower* the probability of

Table 19.3 Example of a spreadsheet with APT query results for temporal patterns as the variables

Map	Pattern 1	Pattern 2
1	0.30	0.58
2	0.25	0.67
3	0.40	0.56
Mean	0.317	0.603
(*Standard deviation*)	0.076	0.059

The value in each cell is a measure of the probability of the relation (pattern)

Pattern 2 [when nimbus stratus clouds, then p.s.i. <30, then temp ≤32 °F and sleet, then temp ≤31 and snow follows]).

In summary, in APT we measure relations directly by identifying and matching patterns in temporal maps. Note that, in this chapter, we focus on APT, and while we show how APT can be used to map and analyze temporal relations in the Diffusion Simulation Game (DSG), MAPSAT methods can be used for many kinds of research problems (see Frick et al., 2008)

3 Diffusion of Innovations Theory and the Diffusion Simulation Game

To illustrate how APT is used for serious games analytics, we will next examine data from several play-learners who played the DSG, a simulation game that models aspects of DoI theory. In order to be successful in the game, play-learners must apply DoI theory in appropriate and timely ways.

3.1 Diffusion of Innovations Theory

While working on his doctoral dissertation on the diffusion of agricultural innovations, Everett Rogers became convinced that the diffusion of innovations followed a general pattern regardless of the type of innovation or the culture in which it was spreading (Rogers, 2003). He began developing a general model of diffusion and published the first edition of his book, *Diffusion of Innovations*, in 1962. Each subsequent decade he published an updated edition as he reviewed the latest research and theoretical developments and refined the model. At the time of publication of the fifth edition (2003), Rogers estimated that there were about 5,200 publications on diffusion, with roughly 120 new diffusion publications each year.

Rogers defines “diffusion” as a social process “in which an innovation is communicated through certain channels over time among members of a social system” (p. 5). The goal of communication with respect to an innovation is to reduce uncertainty by sharing information and subjective evaluations of the innovation. Rogers’ definition contains four main elements that are key to understanding the model, including:

1. The nature and attributes of the innovation
2. The communication channels through which information is disseminated
3. The time required for individuals to make a decision regarding the adoption of the innovation
4. The social system through which the innovation is diffused

A detailed description of DoI theory is beyond the scope of this chapter. However, knowing a little about a few key aspects of the model will aid in understanding the simulation game that is the focus of this chapter’s analysis.

A **communication channel** is “the means by which messages get from one individual to another” (Rogers, 2003, p. 18). *Mass media channels* enable a small number of people to spread their messages to a large audience. Mass media channels are generally effective in creating awareness about the existence of an innovation, especially among earlier adopters who tend to pay more attention to external sources of information. *Interpersonal channels* “involve a face-to-face exchange between two or more individuals” (p. 18). Interpersonal communication is less effective in creating awareness or interest in an innovation and more effective in persuading someone to try an innovation about which they are already aware, especially if the message is coming from someone who is “similar in socioeconomic status, education, or other important ways” (p. 18).

Based on decades of observation and research, Rogers developed a model of the **innovation-decision process**, which he defines as

> the process through which an individual (or other decision-making unit) passes from first knowledge of an innovation, to the formation of an attitude toward the innovation, to a decision to adopt or reject, to implementation and use of the new idea, and to confirmation of this decision. (p. 20)

Rogers describes five stages in this process. In the first edition of his book (Rogers, 1962), these stages were: awareness, interest, appraisal, trial, and adoption. By the fifth edition (Rogers, 2003) these stages had become: knowledge, persuasion, decision, implementation, and confirmation—and he contends that they usually occur in this specific sequence unless, for example, the decision stage precedes the persuasion stage because adoption was declared mandatory by an authority figure.

Rogers categorizes the individuals who form a **social system** according to their *innovativeness*, which he defines as “the degree to which an individual or other unit of adoption is relatively earlier in adopting new ideas than the other members of a system” (p. 22). The five categories range from *innovators*, who actively seek information about new ideas through relatively greater exposure to mass media and interpersonal networks that extend well beyond their local system, to *laggards*, who are the least connected to others in the system with many being near isolates, making them difficult to influence. *Early adopters* are of particular importance in the diffusion of an innovation because they have “the highest degree of opinion leadership in most systems” (p. 283), making them crucial in achieving a critical mass of adopters and influencing later adopters.

3.2 The Diffusion Simulation Game

The original DSG was conceived and created “in 1975–1976 at Indiana University by an Instructional Development Center team composed of professor Michael Molenda and six IST [Instructional Systems Technology] graduate students, led by Patricia Young and Dale Johnson” (M. H. Molenda, personal communication, May 9, 2011). The board game was to be used during a day-long workshop, and Molenda and Rice (1979) reported that it underwent extensive formative evaluation and refinement to ensure that the affective and cognitive objectives were achieved. Among these objectives were the ability to classify individuals by adopter type and communication role (e.g., opinion leader) based on described attributes, to identify the stages of the innovation-decision process, and to select the most effective diffusion activities based on the available information.

In the DSG, the player takes on the role of a change agent whose task is to influence the principal and teachers at a junior high school to adopt peer tutoring. The player may gather information about each staff member and also view diagrams of professional and interpersonal networks.

The player may also choose from a variety of diffusion activities, some of which target a single individual or up to five people. For example, the player may use the "Talk To" activity to have a face-to-face discussion with one staff member; the "Print" activity to distribute written materials to as many as five staff members; or the "Local Mass Media" activity to influence those who pay attention to the mass media. Each activity requires from 1 to 6 weeks to complete, and the player has 2 academic years (72 weeks) to persuade as many staff members as possible to move through the stages of the innovation-decision process and adopt peer tutoring.

The results of a player's choices are determined by an "algorithm board" (Molenda & Rice, 1979, p. 462) shown in Fig. 19.1. The circled numbers in Fig. 19.1 indicate which group of feedback cards should be accessed, one of which is randomly selected. Based on the chosen activity, the affected staff members, and in many cases previously chosen activities, the game monitor consults the algorithm board to determine the outcome. For example, if the "Talk To" activity is selected along with one of the opinion leaders (represented in the game by the letters F, H, and M), the game monitor is instructed to refer to the card set represented by the number 7. This particular card set contains 6 cards, 5 of which provide positive feedback and reward points, such as:

> He/she listens attentively to your ideas and shares them with his/her out-of-school compatriots. GAIN 2 POINTS FOR HIM/HER and ONE POINT FOR EACH OF HIS/HER SOCIAL CONTACTS.

The sixth card also provides positive feedback but does not reward points:

> A potentially useful contact; if he/she adopts, a number of others will be favorably disposed. Unfortunately, this is the week his/her family was moving into a new home…no time for serious talk. May be worth trying again later. NO POINTS.

The slight possibility of unfavorable results for what should be effective strategies is meant to model the stochastic nature of dealing with human beings in the real world. One of the affective goals of the game is to foster appreciation for the difficulty of diffusing an innovation.

In 2002, Frick led a development team in the creation of the DSG as an online simulation game (Frick, Kim, Ludwig, & Huang, 2003). Figure 19.2 shows the interface for this online version, which was developed using HTML, CSS, and XML for

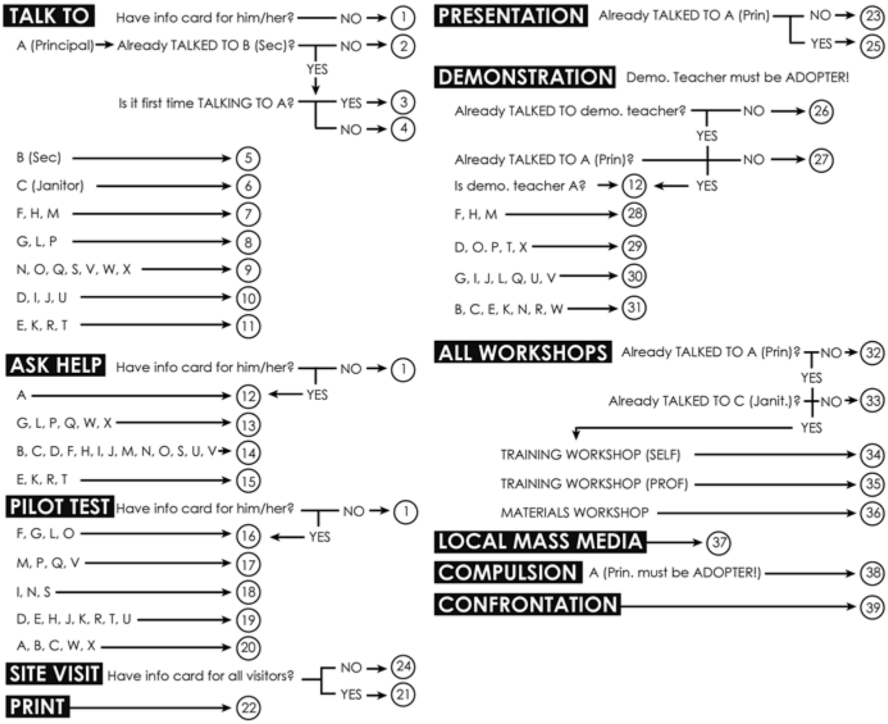
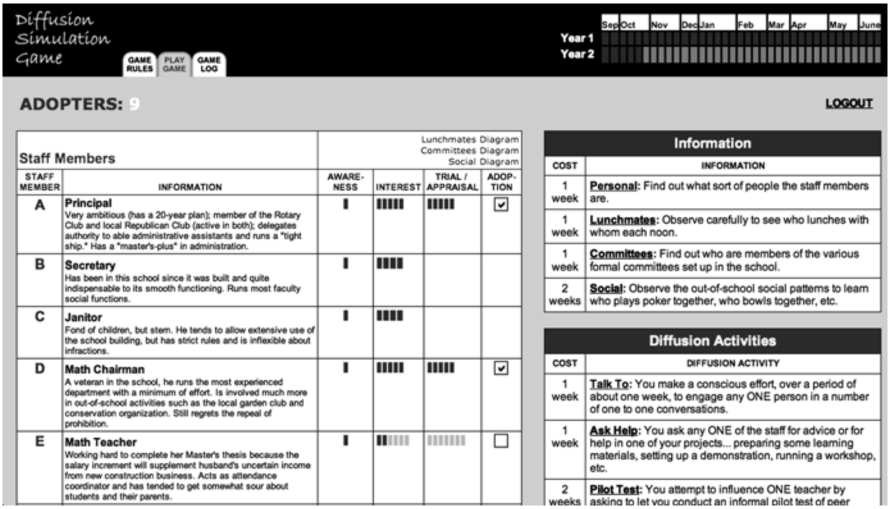

TALK TO
Have info card for him/her? NO 1
A (Principal) Already TALKED TO B (Sec)? NO 2
YES
Is it first time TALKING TO A? YES 3
NO 4
B (Sec) 5
C (Janitor) 6
F, H, M 7
G, L, P 8
N, O, Q, S, V, W, X 9
D, I, J, U 10
E, K, R, T 11
ASK HELP
Have info card for him/her? NO 1
A 12 YES
G, L, P, Q, W, X 13
B, C, D, F, H, I, J, M, N, O, S, U, V 14
E, K, R, T 15
PILOT TEST
Have info card for him/her? NO 1
F, G, L, O 16 YES
M, P, Q, V 17
I, N, S 18
D, E, H, J, K, R, T, U 19
A, B, C, W, X 20
SITE VISIT
Have info card for all visitors? NO 24
YES 21
PRINT 22
PRESENTATION
Already TALKED TO A (Prin) NO 23
YES 25
DEMONSTRATION
Demo. Teacher must be ADOPTER!
Already TALKED TO demo. teacher? NO 26
YES
Already TALKED TO A (Prin)? NO 27
Is demo. teacher A? 12 YES
F, H, M 28
D, O, P, T, X 29
G, I, J, L, Q, U, V 30
B, C, E, K, N, R, W 31
ALL WORKSHOPS
Already TALKED TO A (Prin)? NO 32
YES
Already TALKED TO C (Janit.)? NO 33
YES
TRAINING WORKSHOP (SELF) 34
TRAINING WORKSHOP (PROF) 35
MATERIALS WORKSHOP 36
LOCAL MASS MEDIA 37
COMPULSION
A (Prin. must be ADOPTER!) 38
CONFRONTATION 39

Fig. 19.1 Algorithm board in the original diffusion simulation game (Molenda & Rice, 1979)



	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	June
Year 1										
Year 2										

ADOPTERS: 9

Staff Members

Lunchmates Diagram
Committees Diagram
Social Diagram

STAFF MEMBER	INFORMATION	AWARE-NESS	INTEREST	TRIAL / APPRAISAL	ADOP-TION
A	**Principal** Very ambitious (has a 20-year plan); member of the Rotary Club and local Republican Club (active in both); delegates authority to able administrative assistants and runs a "tight ship." Has a "master's-plus" in administration.	▮	▮▮▮▮▮	▮▮▮▮▮	☑
B	**Secretary** Has been in this school since it was built and quite indispensable to its smooth functioning. Runs most faculty social functions.	▮	▮▮▮▮		
C	**Janitor** Fond of children, but stem. He tends to allow extensive use of the school building, but has strict rules and is inflexible about infractions.	▮	▮▮▮▮		
D	**Math Chairman** A veteran in the school, he runs the most experienced department with a minimum of effort. Is involved much more in out-of-school activities such as the local garden club and conservation organization. Still regrets the repeal of prohibition.	▮	▮▮▮▮▮	▮▮▮▮▮	☑
E	**Math Teacher** Working hard to complete her Master's thesis because the salary increment will supplement husband's uncertain income from new construction business. Acts as attendance coordinator and has tended to get somewhat sour about students and their parents.	▮	▮▮		☐

Information

COST	INFORMATION
1 week	**Personal**: Find out what sort of people the staff members are.
1 week	**Lunchmates**: Observe carefully to see who lunches with whom each noon.
1 week	**Committees**: Find out who are members of the various formal committees set up in the school.
2 weeks	**Social**: Observe the out-of-school social patterns to learn who plays poker together, who bowls together, etc.

Diffusion Activities

COST	DIFFUSION ACTIVITY
1 week	**Talk To**: You make a conscious effort, over a period of about one week, to engage any ONE person in a number of one to one conversations.
1 week	**Ask Help**: You ask any ONE of the staff for advice or for help in one of your projects... preparing some learning materials, setting up a demonstration, running a workshop, etc.
2 weeks	**Pilot Test**: You attempt to influence ONE teacher by asking to let you conduct an informal pilot test of peer

Fig. 19.2 Partial image of the online *Diffusion Simulation Game*

information display and storage, and PHP for interaction programming. The latest version of the game may be accessed at https://www.indiana.edu/~simed/diffusion/.

In Fig. 19.2, staff members (A-X) are listed on the left, with filled rectangles indicating each staff member's stage of adoption. Activities for getting information about staff members and diffusion activities are listed on the right. Elapsed time in weeks is shown on the top right. Vertical scrolling is typically required to see the entire game board in a Web browser.

Since 2006, when Frick released a public version with anonymous login, data from more than 30,000 game sessions have been collected (through April, 2014).

4 Application of *APT* to *DSG* Play-Learner Data

As with any designed learning experience, with serious games we must specify performance indicators of learning. Because the DSG uses DoI as its primary conceptual model, we began by identifying generalizations from *Diffusion of Innovations* (Rogers, 2003) that were applicable while playing the DSG. For example, Rogers says that mass media should be effective in spreading knowledge about an innovation, especially among innovators and early adopters. We then mapped these statements to actions that may be taken in the DSG, which involve combinations of activities, adopter types, and innovation-decision phases. Next we identified data associated with these actions and designed a database for data collection in which the columns are event classifications (e.g., activity selected, current stage in the innovation-decision process for each staff member) and the rows contain the relevant categories in each classification for each turn in a game.

We specified two general kinds of strategies. The first kind of strategy involved the selection of an activity available in the game at an appropriate time to influence staff members at particular stages of the innovation-decision process. Some activities, here referred to as *targeted* activities, require the selection of one or up to five staff members (targets). For example, the *Talk To* activity requires the selection of one staff member, while the *Site Visit* activity allows the selection of up to five staff members. The second kind of strategy involved the selection of particular staff members based on their attributes, which include adopter type, opinion leadership, and interpersonal relationships.

We specified nine strategies from DoI that should lead to success in the DSG, subsequently reviewed and confirmed by experts in DoI (Myers, 2012). Each of these strategies consisted of a pattern of joint occurrences of categories within the various classifications. To continue the example above, Strategy 3 says to use the Local Mass Media activity to gain points in the Awareness and Interest phases among earlier adopters. For details on the strategies, see Myers (2012, pp. 82–87).

In addition to the improvements to the DSG's computational model described above, we implemented a registration and login system to replace anonymous gameplay. This enabled us to associate multiple games with a single play-learner so that we could look for changes in patterns of strategy use over time. We also wrote

Table 19.4 Categories of game outcomes based on number of adoption points

Game outcome	Adoption points
Maximally successful	220
Highly successful	166–219
Moderately successful	146–165
Unsuccessful	0–145

a strategy scoring algorithm that analyzed the game state and assigned a score to each strategy based on the likelihood of its success in that turn. Strategy 3 would be assigned a high score if all or most of the earlier adopters needed points in the Awareness or Interest phases. Otherwise it would receive a low score and other strategies would have a higher probability of success.

Nearly 2 months after launching the revised DSG, we downloaded play-learner data for analysis. Of the 257 active play-learners, 240 gave us permission to use their data. We decided to examine only "finished" games, which we defined as achieving all 22 adopters or using all 72 available weeks. We found 109 play-learners had completed 1 or more games, while 27 had completed 2 or more games, and 14 had completed 3 or more games. From this sample, we selected three players to serve as illustrative examples with contrasting patterns here.

To simplify the APT queries, we recoded several variables into new APT classifications and categories. For example, the two measures of success in the game are the number of adopters achieved and the number of adoption points achieved. The number of points necessary to turn a particular staff member into an adopter depends largely on his or her adopter type, with innovators requiring as few as 5 points and laggards as many as 14 points. The points are distributed across the Awareness, Interest, and Trial phases that lead to Adoption. Obtaining all 22 adopters requires 220 points. When measuring success in the DSG, the number of points obtained is arguably a better metric than the number of adopters obtained. To understand this, imagine a game in which the player obtained 8 adopters, while the rest of the staff members were still in the Awareness or Interest stages. Compare this with a game in which the player obtained only 5 adopters, while the rest of the staff members had moved through Awareness and Interest and were in the Trial stage. Overall the latter player gained many more points toward adoption even though fewer adopters were obtained. We created a new APT classification named "Game Outcome" with the following categories based on final adoption points (see Table 19.4).

Table 19.5 shows the APT classifications used in this study, along with a brief description of each.

The three players selected for this analysis all showed some improvement over time in Game Outcome. Player 1 played 4 games; the first 3 were Unsuccessful, and the last was Moderately Successful. Player 2 played 11 games; the first 2 were Unsuccessful, and the last 3 were Highly Successful. Player 3 played 6 games; the first was Unsuccessful, the fifth was Maximally Successful, and the others were Moderately to Highly Successful.

Table 19.5 APT classifications for analysis of DSG play-learner data

Classification	Description
Unix epoch time	A unique timestamp for each turn
Player	The play-learner's identifier
Game	The game identifier. Each player has multiple games for analysis
Turn	The turn identifier for a game
Activity	The DSG activity chosen by the play-learner for this turn
Game outcome	A category based on the number of adoption points, as described in Table 19.4
Target opinion leader	"TRUE" if the person selected to engage in the turn's activity was an opinion leader. "FALSE" if the person selected was not an opinion leader. "NULL" if no person was selected
Target gatekeeper	"TRUE" if the person selected to engage in the turn's activity was a gatekeeper. "FALSE" if the person selected was not a gatekeeper. "NULL" if no person was selected
Target earlier adopter	"TRUE" if the person selected to engage in the turn's activity was an innovator or early adopter. "FALSE" if the person selected was not an innovator or early adopter. "NULL" if no person was selected
Target social connectedness	"TRUE" if the person selected to engage in the turn's activity had ten or more interpersonal connections with other staff members. "FALSE" if the person selected had fewer than ten connections. "NULL" if no person was selected
Target decision phase	The target's phase in the innovation-decision process at the start of the turn: "NULL," "Awareness," "Interest," "Trial," or "Adoption"
Target follower interest	Based on the percentage of the target's followers who are in the Interest phase: "High" > 65 %; "Medium" = 33–65 %; "Low" = 1–32 %; "None" = 0 %
Turn rank	As described earlier, a score for every optimal strategy was calculated for each turn. These scores were then assigned a rank from 1 (Best) to 10 (Worst, when no optimal strategy was used). The value for Turn Rank is the rank of the strategy used for the turn

We ran an APT query for every strategy to calculate the frequency of that strategy in each game. Strategy 1 specifies targeting earlier adopters and opinion leaders, and Strategy 8 specifies targeting people with a large number of interpersonal connections. However, these strategies must be considered in the context of the activity chosen, for if the activity is not appropriate (e.g., an activity like Print that raises awareness and interest used with targets who are already in the Trial phase), it will be less successful. Therefore, for strategies that include targeted activities, we also ran queries to see how frequently desirable targets were selected. We ran similar queries to calculate when those strategies were ranked high (in the top three ranks) and low (in the bottom three ranks). In general, we expected that greater use of high-ranking strategies would increase the probability of success in the game.

As an example, let's look at Strategy 3, which says to use Local Mass Media and Print activities to gain points in the Awareness and Interest phases among earlier adopters. Let's focus on Local Mass Media, which is not a targeted activity. The scoring algorithm for this activity counts the number of earlier adopters who need points in Awareness or Interest and divides that by the total number of earlier adopters. Therefore, this activity's strategy score will be highly ranked when many earlier adopters need points in Awareness or Interest. Use of this activity when it is highly ranked should increase the probability of a successful game outcome.

We have set up our data so that each game is a separate APT map. The APT query tool returns counts and proportions for each map. The first APT query looks at overall use of this strategy by counting the number of turns in which Local Mass Media is used in proportion to the total number of turns. Here is an example result from one play-learner's map:

> **Pattern 3:** Query Result for Player 1, Game 3
>
> WHILE the FIRST Joint Temporal Event is true (Phrase 1):
> **Diffusion Activity** is in state *starting or continuing*, value = *Local Mass Media*
>
> - Duration when Phrase 1 is True = 2 moves (out of 59 DSG moves total). Proportion of Time = 0.03390.
> - Joint Event Frequency when Phrase 1 is True = 2 (out of 74 total joint temporal events). Proportion of JTEs = 0.02703.

In this example, the play-learner used Local Mass Media in 2 out of 59 turns or 0.03390 (3.4 %) of the time. Using this query, we find Player 1 did not use Local Mass Media in the first two games (both Unsuccessful games). In the third game (also Unsuccessful), Player 1 used the activity in 2 out of 59 turns, a proportion of 0.03390. In the fourth and final game, the activity was used in 2 out of 68 turns, a proportion of 0.02941.

The next APT query further limits the turns to those that had high ranking strategy scores, defined as a Turn Rank value of "Less than or equal to 3." To continue with the previous example result:

Table 19.6 Use of local mass media activity by game outcome and strategy rank for turn

Player 1	*Un*	*Un*	*Un*	*Md*							
Overall	0.00	0.00	0.03	0.03							
High	0.00	0.00	0.02	0.03							
Low	0.00	0.00	0.02	0.00							
Player 2	*Un*	*Un*	*Md*	*Hi*	*Md*	*Hi*	*Md*	*Un*	*Hi*	*Hi*	*Hi*
Overall	0.00	0.05	0.03	0.07	0.05	0.06	0.04	0.05	0.02	0.10	0.10
High	0.00	0.00	0.00	0.02	0.02	0.02	0.02	0.00	0.00	0.05	0.05
Low	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.02	0.02
Player 3	*Un*	*Hi*	*Md*	*Hi*	*Mx*	*Hi*					
Overall	0.02	0.07	0.10	0.10	0.05	0.09					
High	0.02	0.02	0.03	0.03	0.03	0.03					
Low	0.00	0.05	0.05	0.08	0.03	0.06					

See Table 19.4 for definitions of unsuccessful, and moderately, highly, and maximally successful game outcomes

Pattern 4: Query Result for Player 1, Game 3

WHILE the FIRST Joint Temporal Event is true (Phrase 1):
Diffusion Activity is in state *starting or continuing*, value = *Local Mass Media*
Turn Rank is in state *starting or continuing*, value <= 3

- Duration when Phrase 1 is True = 1 moves (out of 59 DSG moves total). Proportion of Time = 0.02222.
- Joint Event Frequency when Phrase 1 is True = 1 (out of 74 total joint temporal events). Proportion of JTEs = 0.01351.

The final APT query (not shown for pattern 5) changes the Turn Rank value to "Greater than or equal to 6." Table 19.6 shows for all players and games (by game outcome) the proportions of Local Mass Media use overall (pattern 3), when its rank is high (pattern 4), and when its rank is low (pattern 5).

Player 1 seems to gain in his understanding of the strategy regarding the use of Local Mass Media with earlier adopters who need points in the Awareness and Interest phases. By his final game (Moderately Successful), he used the strategy in 3 % of his turns, always when it was highly ranked. Player 2 applied the strategy more sporadically; in her last two games she used it the most (10 % of turns), but it had a low ranking for 2 % of turns. Player 3 used the strategy relatively frequently, but the proportion of times when it was low ranking suggests that her timing was off and she needed to pay more attention to the innovation-decision phases of the earlier adopters.

For another example, we will focus on Player 3's use of Strategy 2, which says to use the Personal Information and Talk To activities to establish empathy and rapport in order to understand a client's needs, sociocultural values and beliefs, and previous exposure to related ideas. We will focus on the Talk To activity, which is especially useful with gatekeepers, people who control access to resources and can create obstacles to the diffusion of an innovation. Here is an example of a query result for one of Player 3's games:

Pattern 6: Query Result for Player 3, Game 3

WHILE the FIRST Joint Temporal Event is true (Phrase 1):
Diffusion Activity is in state *starting or continuing*, value = *Talk To*

Table 19.7 Use of talk to activity by game outcome and strategy rank for turn

Player 3	*UN*	*HI*	*MD*	*HI*	*MX*	*HI*
Overall	0.27	0.38	0.28	0.35	0.34	0.31
High rank	0.00	0.26	0.21	0.23	0.29	0.28
Low rank	0.16	0.02	0.00	0.00	0.00	0.00
W/Gatekeepers						
Overall	0.09	0.17	0.18	0.18	0.13	0.13
High rank	0.00	0.14	0.13	0.13	0.13	0.13
Low rank	0.02	0.02	0.00	0.00	0.00	0.00

Target Gatekeeper is in state *starting or continuing*, value *True*

- Duration when Phrase 1 is True = 7 moves (out of 43 DSG moves total). Proportion of Time = 0.16279.
- Joint Event Frequency when Phrase 1 is True = 7 (out of 82 total joint temporal events). Proportion of JTEs = 0.08537.

Now let's compare all of Player 3's games, including proportions of use when the strategy is ranked high (pattern 7) and low (pattern 8) for all targets, and then for targeted gatekeepers (patterns 9–11). See Table 19.7.

In her first (Unsuccessful) game, she used the Talk To activity less than in subsequent games, and when she used it, it was never one of the high ranking strategies. Furthermore, she targeted gatekeepers less than in subsequent games.

As we saw in the weather example above, APT is not limited to single-phrase queries of temporal maps. Indeed, its power lies in querying sequences of complex patterns that are not easily found in database tables or spreadsheets.

The DSG promotes the use of Strategy 2 (the use of the Personal Information and Talk To activities to establish empathy and rapport) by requiring the play-learner to use the Personal Information activity on his first turn to gather information about five people. Furthermore, attempts to use some other activities are stymied if the Personal Information and Talk To activities have not been used with certain people, especially gatekeepers. For example, if an attempt is made to talk to the principal before talking to the principal's secretary, the game provides this feedback:

> STOP! The secretary says the principal is too busy to see you. You're not going to have access to him without her "approval." Have a talk with her.

Savvy players quickly learn from their mistake. The results of an APT query that looks for instances in which the play-learner first uses the Talk To activity with the principal, then with the secretary, then with the principal again are shown below. Note that the secretary is a gatekeeper, but the principal is the only staff member who is both a gatekeeper and an opinion leader.

Pattern 12: Query Result for Player 3, Game 1

WHILE the FIRST Joint Temporal Event is true (Phrase 1):
Diffusion Activity is in state *starting or continuing*, value = *Talk To*
Target Opinion Leader is in state *starting or continuing*, value = *True*
Target Gatekeeper is in state *starting or continuing*, value = *True*

- Duration when Phrase 1 is True = 2 DSG moves (out of 43 DSG moves total). Proportion of Time = 0.04651.
- Joint Event Frequency when Phrase 1 is True = 2 (out of 86 total JTEs). Proportion of JTEs = 0.02326.

THEN while the NEXT Joint Temporal Event is true (Phrase 2):
Diffusion Activity is in state *starting or continuing*, value = *Talk To*
Target Opinion Leader is in state *starting or continuing*, value = *False*

Target Gatekeeper is in state *starting or continuing*, value = *True*

- Duration when Phrase 2 is True = 1 DSG moves (out of 43 DSG moves total), given all prior phrases are true. Proportion of Time = 0.02326.
- Joint Event Frequency when Phrase 2 is True = 1 (out of 86 total JTEs), given all prior phrases are true. Proportion of JTEs = 0.01163.

- Conditional joint event *duration* when Phrase 2 is true, given all prior phrases are true = 0.50000 (1 out of 2 DSG moves (time units)).
- Conditional joint event *frequency* when Phrase 2 is true, given all prior phrases are true = 0.50000 (1 out of 2 JTEs).

THEN while the NEXT Joint Temporal Event is true (Phrase 3):
Diffusion Activity is in state *starting or continuing*, value = *Talk To*
Target Opinion Leader is in state *starting or continuing*, value = *True*
Target Gatekeeper is in state *starting or continuing*, value = *True*

- Duration when Phrase 3 is True = 1 DSG moves (out of 43 DSG moves total), given all prior phrases are true. Proportion of Time = 0.02326.
- Joint Event Frequency when Phrase 3 is True = 1 (out of 86 total JTEs), given all prior phrases are true. Proportion of JTEs = 0.01163.

- Conditional joint event duration when Phrase 3 is true, given all prior phrases are true = 1.00000 (1 out of 1 DSG moves (time units)).
- Conditional joint event frequency when Phrase 3 is true, given all prior phrases are true = 1.00000 (1 out of 1 JTEs).

This 3-phrase query for pattern 12 found that Player 3 made the mistake of approaching the principal before talking to the secretary once during her first game only. The results for pattern 12 in her remaining maps showed that she never made this mistake again.

5 Using APT for Assessment

5.1 Formative Assessment During Gameplay

In the examples above, we analyzed data from a serious game to demonstrate how APT can be used to find evidence of a play-learner's understanding and application of the theory underlying a simulation game. This information could be used by an instructor (or by the play-learner herself) after gameplay to identify misconceptions or gaps in understanding.

The approach we used to compare patterns of gameplay data with optimal strategies could be applied during gameplay to provide an instructional overlay (Myers & Reigeluth, in press; Reigeluth & Schwartz, 1989) that delivers appropriate hints, coaching, or other forms of scaffolding during gameplay to improve learning and performance. This instructional support could be requested by the play-learner who is struggling to determine the best course of action, or it could be supplied at the start of a turn as a hint or at the end of a turn as an explanation or prompt for reflection.

In the DSG, for example, the game engine could calculate optimal strategy scores for the turn in progress, and a virtual mentor could provide appropriate generalizations from DoI theory to help the play-learner see the connection between the theory and the game. Similar to the examples above, the game engine could also use APT queries on a play-learner's previous game maps to identify persistent misconceptions, which might be addressed at the start of a game. For example, in Table 19.6 we saw that Player 3 was consistently using Local Mass Media when it was a low-ranked strategy, indicating that she did not understand its usefulness in raising awareness and interest among earlier adopters. At the start of her next game, the game engine could identify this problem and provide relevant generalizations from Rogers (2003):

> Generalization 5-13: *Mass media channels are relatively more important at the knowledge stage*, *and interpersonal channels are relatively more important at the persuasion stage in the innovation-decision process* (p. 205).
> Generalization 7-22: *Earlier adopters have greater exposure to mass media communication channels than do later adopters* (p. 291).

5.2 Using APT for Summative Assessment

Serious game analytics need not be limited to formative assessment. Summative assessment is normally considered to be an evaluation of an entity across a sample of cases or situations in order to make an inference about a population of cases (Reigeluth & Frick, 1999; Scriven, 1967; Worthen & Sanders, 1987). For example, we might want to determine the *effectiveness* of the DSG in terms of student learning achievement—that is, do students appropriately apply principles from DoI theory to play it successfully? Or, we might be interested in *efficiency* of learning via the DSG—that is, how quickly can students learn through playing the DSG repeatedly until they achieve success? Alternatively, we might be interested in comparing two different versions of the DSG, such as one with coaching and one without coaching, to determine which is more effective or more efficient.

APT can be used to make inferences from a sample to a population of cases. In other words, APT can be used to make generalizations about a class of cases, if appropriate sampling strategies are employed. That is, we first analyze patterns *within* each case, and then average probabilities of these patterns *across* cases in order to avoid aggregation aggravation. Probabilities of patterns resulting from APT queries are the measures of "variables" for each case (see Table 19.3). These measures can then be treated statistically in a normal manner to form means and standard deviations, and then subsequent analyses can be carried out (e.g., ANOVA, regression, factor, discriminant, cluster, Bayesian network, and other data mining approaches [e.g., see Jensen & Nielsen, 2007; Witten, Elbe, & Hall, 2011]). A caveat is that data must be collected as temporal maps in order to make APT queries about patterns. Such patterns normally cannot be inferred from the way data are typically collected with separate measures of variables, as Frick (1983) proved mathematically (see Sect. 2.2).

With respect to network analysis (NA) methods for summative assessment, MAPSAT Analysis of Patterns in *Configurations* (APC) could be used. APC is based on mathematical di-graph theory, as are most NA methods (e.g., Brandes & Erlebach, 2005). Properties of di-graphs can be measured with APC that are typically *not* done in NA such as wholeness, vulnerability, interdependence, passive dependence, and strongness. Space does not permit further elaboration here. See Thompson (2008).

6 Concluding Remarks

In this chapter, we have described Analysis of Patterns in Time and demonstrated its effectiveness for serious games analytics. Games have tremendous potential as immersive learning experiences that challenge play-learners to apply their knowledge and skills to solve authentic, difficult problems in a safe environment. Designers of serious games have vast amounts of empirical data available that can be used to assess the learning trajectory of a play-learner. APT can turn these data into actionable assessments that lead to personalized scaffolds targeting an individual's misconceptions and gaps in knowledge and skills. APT can provide *unobtrusive* assessments for analyzing play-learner interactions with serious games, in contrast with methods such as direct observations, video recordings, surveys, questionnaires, interviews, and traditional tests of learning achievement.

While APT can be used for formative assessment of individual cases, as illustrated in this chapter, it can also be used for summative assessment and for research whose goal is to make generalizations based on statistical inferences from a sample to a population. For example, APT can be a valuable research tool for investigating the *effectiveness* of simulations, games, and other forms of instruction by showing the relationship between what students experience and what they are learning. Myers and Frick (2015) are conducting such a study of the Diffusion Simulation Game to illustrate the use of APT for this purpose.

References

Brandes, U., & Erlebach, T. (2005). *Network analysis: Methodological foundations*. Berlin, Germany: Springer.

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, NJ: Prentice-Hall.

Frick, T. (1983). *Nonmetric temporal path analysis (NTPA): An alternative to the linear models approach for verification of stochastic educational relations.* Unpublished doctoral dissertation, Indiana University, Bloomington, IN. Retrieved from http://www.indiana.edu/~tedfrick/ntpa/

Frick, T. (1990). Analysis of patterns in time (APT): A method of recording and quantifying temporal relations in education. *American Educational Research Journal, 27*(1), 180–204.

Frick, T. (1997). Artificial tutoring systems: What computers can and can't know. *Journal of Educational Computing Research, 16*(2), 107–124.

Frick, T., Kim, K.-J., Ludwig, B., & Huang, R. (2003). *A Web simulation on educational change: Challenges and solutions for development*. Paper presented at the meeting of the Association for Educational Communication and Technology, Anaheim, CA. Retrieved from http://www.indiana.edu/~tedfrick/aect2003/frick_kim_ludwig_huang.pdf

Frick, T., Myers, R., Thompson, K., & York, S. (2008). *New ways to measure systemic change: Map & analyze patterns & structures across time (MAPSAT)*. Featured research paper presented at the annual conference of the Association for Educational Communications & Technology, Orlando, FL. Retrieved from https://www.indiana.edu/ ~tedfrick/MAPSATAECTOrlando2008.pdf

Jensen, F. V., & Nielsen, T. D. (2007). *Bayesian networks and decision graphs* (2nd ed.). New York, NY: Springer.

Maccia, E. S., & Maccia, G. (1966). *Development of educational theory derived from three educational theory models* (final report, project no. 5-0638). Washington, DC: U.S. Department of Health, Education, and Welfare.

Molenda, M., & Rice, J. M. (1979). Simulation review: The diffusion simulation game. *Simulation and Games, 10*(4), 459–467.

Myers, R. D. (2012). *Analyzing interaction patterns to verify a simulation/game model.* Unpublished doctoral dissertation, Indiana University, Bloomington, IN. Retrieved from http://webgrok.com/papers/RodneyMyers_DissertationFinal_Approved.pdf

Myers, R. D., & Frick, T. W. (2015). Measuring effectiveness of instructional games and simulations: Pattern analysis of play in the Diffusion Simulation Game. Bloomington, IN.

Myers, R. D., & Reigeluth, C. M. (in press). Designing games for learning. In C. M. Reigeluth, B. Beatty, & R. D. Myers (Eds.). *Instructional-design theories and models* (Vol. IV).

Reigeluth, C. M., & Frick, T. W. (1999). Formative research: A methodology for creating and improving design theories. In C. M. Reigeluth (Ed.), *Instructional-design theories and models: A new paradigm of instructional theory* (Vol. II, pp. 633–651). Mahwah, NJ: Lawrence Erlbaum.

Reigeluth, C. M., & Schwartz, E. (1989). An instructional theory for the design of computer-based simulations. *Journal of Computer-Based Instruction, 16*(1), 1–10.

Rogers, E. M. (1962). *Diffusion of innovations*. New York, NY: Free Press.

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York: Simon & Schuster.

Scriven, M. (1967). The methodology of evaluation. In R. Tyler, R. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation. AERA monograph series on curriculum evaluation* (pp. 39–83). Chicago, IL: Rand McNally.

Thompson, K. (2008). *ATIS graph theory*. Columbus, OH: System Predictive Technologies. Retrieved from http://www.indiana.edu/~aptfrick/reports/11ATISgraphtheory.pdf

Witten, I. H., Elbe, F., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Burlington, MA: Morgan Kaufmann.

Worthen, B., & Sanders, J. (1987). *Educational evaluation: Alternative approaches and practical guidelines*. New York, NY: Longman.

Author Index

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based Learning, DOI 10.1007/978-3-319-05834-4

Subject Index

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based Learning, DOI 10.1007/978-3-319-05834-4