Advances in Game-Based Learning

Dirk Ifenthaler
Yoon Jeon Kim *Editors*

Game-Based Assessment Revisited

Springer

Advances in Game-Based Learning

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

Dirk Ifenthaler • Yoon Jeon Kim
Editors

Game-Based Assessment Revisited

Springer

Editors
Dirk Ifenthaler
Learning, Design and Technology
University of Mannheim
Mannheim, Germany

Yoon Jeon Kim
Teaching Systems Lab
Massachusetts Institute of Technology
Cambridge, MA, USA

ISSN 2567-8086 ISSN 2567-8485 (electronic)
Advances in Game-Based Learning
ISBN 978-3-030-15568-1 ISBN 978-3-030-15569-8 (eBook)
https://doi.org/10.1007/978-3-030-15569-8

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

In 2012, Ifenthaler, Eseryel, and Ge published a first edited volume focusing on game-based assessment (GBA), covering the current state of research, methodology, assessment, and technology of game-based learning from international contributors. The 2012 volume remained the only collection in the field of assessment and game-based learning. After more than 5 years, advances in assessment, especially in the area of analytics, have been made. These advances shall be collected and critically reflected in this edited volume titled "Game-Based Assessment Revisited."

We organized the chapters included in this edited volume into three major parts: (I) *Foundations of Game-Based Assessment*, (II) *Emerging Methods and Practices*, and (III) *Best Practice Implementations.*

In Part I, the first chapter, titled "Game-Based Assessment: The Past Ten Years and Moving Forward," reports on previous research findings and current developments in game design, assessment practices, and analytics capabilities (*Yoon Jeon Kim, Dirk Ifenthaler*, Chap. 1). The next chapter, "Assessing Learning from, with, and in Games Revisited," presents six principles that may help researchers to engage in studies that involve the process of learning (*P.G. Schrader, Michael P. McCreery, Mark C. Carroll, Danielle L. Head, Jeffrey R. Laferriere*, Chap. 2). The following chapter, "Summative Game-Based Assessment," extends what has been developed and learned about formative game-based assessments into summative assessment practices (*Andreas Oranje, Bob Mislevy, Malcolm I. Bauer, G. Tanner Jackson*, Chap. 3). Next, "Stealth Assessment Embedded in Game-Based Learning to Measure Soft Skills: A Critical Review" discusses how to embed stealth assessment in game-based learning to empower learners from theoretical and practical perspectives (*Xinyue Ren*, Chap. 4). The final chapter of the first part, "Intrinsic Motivation in Game-Based Learning Environments," examines how researchers have implemented and assessed intrinsic motivation in game-based learning environments (*T. Fulya Eyupoglu, John L. Nietfeld*, Chap. 5).

In Part II, the opening chapter, "Examining Designed Experiences: A Walkthrough for Understanding Video Games as Performance Assessments," offers guidance for researchers to extract dynamic, emergent, and complex data from video game contexts and thus unlock the potential for games to function as performance assessments (*Michael P. McCreery, P. G. Schrader, S. Kathleen Krach, Jeffrey R. Laferriere, Catherine A. Bacos, Joseph P. Fiorentini*, Chap. 6). The next chapter, "Press Play! How Immersive Environments Support Problem-Solving Skills and Productive Failure," examines how student interactions during gameplay can be assessed in immersive environments without disrupting the flow of gameplay (*Benjamin Emihovich, Logan Arrington, Xinhao Xu*, Chap. 7). The following chapter, "New Perspectives on Game-Based Assessment with Process Data and Physiological Signals," highlights not only the potentials of process and physiological data but also the problems that can arise in this context (*Steve Nebel, Manuel Ninaus*, Chap. 8). Next, "A Provisional Framework for Multimodal Evaluation—Establishing Serious Games Quality Label for Use in Training and Talent Development" depicts an attempt in establishing a provisional framework of multimodal evaluation that can be used to generate quality labels for serious games, particularly in the training and talent development sector (*Wee Hoe Tan, Ivan Boo*, Chap. 9). The final chapter of the second part, "Scaffolding and Assessing Teachers' Examination of Games for Teaching and Learning," illustrates how formative and summative assessments were created using the GaNA framework to support participating preservice teachers in examining games as a form of curriculum and to allow the researcher to qualitatively and quantitatively capture the change in teachers' game literacy and the extent to which it was integrated with the teachers' design of game-based lesson plans (*Mamta Shah*, Chap. 10).

In Part III, the first chapter, "Assessing Game-Based Mathematics Learning in Action," focuses on extracting design and implementation heuristics related to game-based, learning-in-action assessment (*Fengfeng Ke, Biswas Parajuli, Danial Smith*, Chap. 11). The next chapter, "Bridging Two Worlds: Principled Game-Based Assessment in Industry for Playful Learning at Scale," offers an example of a working GBA practice in an industry context that implements evidence-centered learning design—integrated with the principles of Educational Data Mining to inform corresponding event-stream data design—for the production of data-driven educational games to support learning for students at scale (*V. Elizabeth Owen, Diana Hughes*, Chap. 12). The following chapter, "Effectiveness of Supply Chain Games in Problem-Based Learning Environment," aims to evaluate the game's effectiveness as a formative assessment tool in problem-based learning environment based on two main criteria: learning objective and game experience (*Linda William, Za'Aba Bin Abdul Rahim, Liping Wu, Robert de Souza*, Chap. 13). In another chapter in this part, "What Does Exploration Look Like? Painting a Picture of Learning Pathways Using Learning Analytics," three novel metrics that focus more on the learning process of students than on the outcomes are proposed (*José A. Ruipérez-Valiente, Louisa Rosenheck, Yoon Jeon Kim*, Chap. 14). Next, "Making a Game of Troublesome Threshold Concepts" shows the use of a gamified learning experience at the beginning of the learners' higher education journey to embed and assess technical threshold concepts (*Kayleen Wood*, Chap. 15). The concluding chapter, "Emerging Practices in Game-Based Assessment," argues for content-agnostic game engineering as a framework that helps provide multiple learning contents within a single game to achieve content-agnostic assessment (*Vipin Verma, Tyler Baron, Ajay Bansal, Ashish Amresh*, Chap. 16).

Without the assistance of experts in the field of game-based learning and assessment, the editors would have been unable to prepare this volume for publication. We wish to thank our board of reviewers for their tremendous help with both reviewing the chapters and linguistic editing.

Mannheim, Germany/Perth, WA, Australia — Dirk Ifenthaler
Cambridge, MA, USA — Yoon Jeon Kim

Contents

About the Editors

Dirk Ifenthaler (dirk@ifenthaler.info, www.ifenthaler.info) is professor and chair of Learning, Design and Technology at the University of Mannheim, Germany; adjunct professor at Deakin University, Australia; and UNESCO deputy-chair of Data Science in Higher Education Learning and Teaching 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 of Applied Teaching and Learning Research at the University of Potsdam, Germany. He 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 organizational learning. 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 USA. Professor Ifenthaler is the editor-in-chief of the Springer journal *Technology, Knowledge and Learning* (www.springer.com/10758). Dirk is the past president of the AECT (Association for Educational Communications and Technology) Design and Development Division; past chair of the AERA (American Educational Research Association) Special Interest Group Technology, Instruction, Cognition and Learning; and co-program chair for the international conference series on Cognition and Exploratory Learning in the Digital Age (CELDA).

Yoon Jeon Kim (yjk7@mit.edu) is the director of the Playful Journey Lab within MIT Open Learning. Her work has centered on the topic of innovative assessment and how technological advancement influences what we are measuring about student learning and how we are measuring it. Before joining MIT, she was involved in many game-and simulation-based assessment development and evaluation projects using evidence-centered design in combination with educational data mining techniques. Currently, she leads the Playful Assessment working group (https://playful.mit.edu/playfulassessment), where she and her colleagues explore playful, authentic, yet rigorous assessment approaches in both digital and nondigital learning environments. The core of her work is close collaboration with practitioners—empowering teachers to innovate around classroom assessment and use playful and authentic assessment tools that can truly impact student learning.

About the Authors

Za' Aba Bin Abdul Rahim (zaaba@nus.edu.sg) is lead (IT & Serious Games) with The Logistics Institute-Asia Pacific, National University of Singapore (NUS). He obtained his BSc in Information Technology and Business from Singapore University of Social Sciences (SUSS) and Sp. Dip. in Instructional Design for E-Learning from the National Institute of Education (NIE), Singapore.

Ashish Amresh (amresh@asu.edu) is the program chair for the Simulation Science, Games and Animation program at Embry Riddle Aeronautical University. In the past, he led the computer gaming curriculum initiatives at Arizona State University, where he founded the Computer Gaming Certificate, Camp Game, and UTBC summer programs. He is currently researching the development of software tools to improve game-based learning outcomes.

Logan Arrington (logana@westga.edu) is an assistant professor in the Department of Educational Technology and Foundations. His focus area is on the Instructional Technology program, where he teaches graduate-level courses on instructional design. Logan gained his Ph.D. in Instructional Systems and Learning Technologies from Florida State University. Logan's key research interests are failure-based instructional strategies and professional issues in the field of instructional design and technology. Logan is an active member in both the American Educational Research Association (AERA) and the Association for Educational Communications and Technology (AECT).

Catherine A. Bacos (bacosc@unlv.nevada.edu) is a Ph.D. student in the Interaction & Media Sciences program in the Teaching & Learning Department at the University of Nevada, Las Vegas. She received her M.Ed. in Curriculum & Instruction: Educational Technology at UNLV and her B.A. in English and American literature from New York University. Her research interests include examining social interactions in virtual and mixed reality environments and designing learning systems for children with autism.

Ajay Bansal (ajay.bansal@asu.edu) is assistant professor in the Software Engineering program in School of Computing, Informatics, and Decision Systems Engineering (CIDSE) at Arizona State University (ASU). He joined ASU in spring 2011 as lecturer. He worked as a visiting assistant professor at Georgetown University in Washington, D.C. He also worked in the industry for 4 years as a software engineer at SIEMENS India, Metallect Corp., and Tyler Technologies. His research interests are in semantics and the implementation of declarative programming languages (specifically logic programming and answer set programming), computational logic, knowledge representation, and automated reasoning. He has also been working in the area of web service description, discovery, and composition and was involved in the design and development of a web service description language called USDL (Universal Service-Semantics Description Language). He has won ten-year Test of Time award at the International Conference on Logic Programming 2016 and Fulton Schools of Engineering Top 5% Best Teacher Award in 2015 and 2018.

Tyler Baron (tjbaron@asu.edu) earned his Ph.D. in Simulation, Modeling, and Applied Cognitive Science in 2017 from Arizona State University in 2017 and has been a lecturer there since. His work focuses on content-agnostic game engineering (CAGE), which is finding ways to develop game mechanics that can be used to teach multiple distinct topics without any changes needed. His CAGE framework provides a basis for the rapid development of these types of game mechanics, allowing developers to quickly create multiple versions of educational games without having to rebuild each version from the ground up.

Malcolm I. Bauer (mibauer@ets.org) is a senior research scientist in the Cognitive and Technology Sciences Center at Educational Testing Service. His work focuses on research and development of innovative assessment and learning experiences and systems. In classrooms and in the lab, he has explored these ideas in the domains of middle school mathematics, statistics, English language learning, science and computer skills, and noncognitive skills that support academic performance and in systems that support computer-based formative and summative assessment, traditional paper-based assessment, and computer-based learning environments. His current work focuses on new methods of assessment design, game-based approaches to the learning and assessment of intercultural competency for teaching, integration of mathematics and science learning progressions, and cognitive and psychometric models to support these and other complex constructs.

Ivan Boo (ivan@seriousgamesasia.com) served as a founding member of Serious Games Association (Singapore) and director of Serious Games Asia. The main purpose of Serious Games Association (Singapore) is to create a bridge between subject matter experts from corporate organizations, healthcare, and education sectors and technologists for adopting, developing, and creating sustainable innovative technology-focused solutions. For Serious Games Asia, Ivan specializes in introducing gamification strategies for healthcare and corporate organizations to increase organizational productivity and work process efficiency. He partners with various technologists including programmers, engineers, data scientists, healthcare clinicians, and educators to develop a game or gamification-related projects.

Mark C. Carroll (mcarroll@unlv.nevada.edu) is a Ph.D. student and adjunct faculty at the University of Nevada, Las Vegas. He also serves as a graduate assistant for the Department of Teaching and Learning and collaborates with the faculty in the Interaction and Media Sciences program. Mark has 20 years of experience in the education sector in the southern Nevada area and 15 years of teaching experience in the secondary and postsecondary settings. He was the Task Force Chair for Technology in Fine Arts for 12 years with the Clark County School District and an instructor and instructional designer at the Alexander Dawson School.

Robert de Souza (rdesouza@nus.edu.sg) is the executive director and chief executive of the multiawarded The Logistics Institute – Asia Pacific based at the National University of Singapore. Prior to joining TLI – Asia Pacific, Dr. de Souza was executive vice president for V3 Systems in the Asia Pacific, corporate senior vice president and global chief knowledge officer at Viewlocity Inc., and vice chairman and CEO of SC21 Pte. Ltd. Dr. de Souza is a senior fellow at the National University of Singapore. He has published extensively and is a member of several editorial boards of international journals. Dr. de Souza is a chartered engineer and serves on multiple industry and academic boards. He has been awarded many honors for his work in urban logistics, risk management, and humanitarian logistics.

Benjamin Emihovich (bemihovi@umich.edu) is a current assistant professor of Educational Technology in the Education Department at the University of Michigan-Flint and is the program faculty coordinator for the online Educational Technology master's degree. He currently teaches undergraduate and graduate students in the areas of instructional design and technology as well as curriculum and instruction. His research area focuses on the following: game-based learning, assessments for learning in immersive environments, and emerging learning technologies. He is an active member in the following professional organizations: Association for Educational Communications and Technology (AECT) and American Educational Research Association (AERA).

T. Fulya Eyupoglu (fulya_eyupoglu@ncsu.edu) is a graduate research assistant and Ph.D. candidate, with a specialization in Digital Learning and Teaching, in the Department of Teacher Education and Learning Sciences at North Carolina State University. She received her bachelor's degree in Science Education and Master of Arts in Science and Mathematics Education from Middle East Technical University. Her international teaching career in science for 8 years directed her toward game-based learning. Fulya is especially interested in metacognition and motivation in game-based learning environments. She has served as a teaching assistant in the Instructional Design Course and Digital Game-Based Learning Course. Currently, she is enrolled in an NSF-funded game-based learning project as a researcher.

Joseph P. Fiorentini (fiorej2@unlv.nevada.edu) is a Ph.D. student in the Interaction & Media Sciences program in the Teaching & Learning Department at the University of Nevada, Las Vegas. He received his B.A. in English Writing from the University of Nevada and his M.S. in Business Leadership from Argosy University. His research interests include examining cognitive functions within virtual environments and emergent readers' attitudes toward reading through the use of video games.

Danielle L. Head (headd1@unlv.nevada.edu) is a Ph.D. student in the Interaction & Media Sciences program in the Teaching & Learning Department at the University of Nevada, Las Vegas. She received her M.Ed. in Curriculum & Instruction: Educational Technology and her B.A. in Secondary Education (Mathematics) at UNLV. She has been a graduate assistant at the University of Nevada, Las Vegas, since 2017. Her research interests include the design and validation of innovative psychometric assessments.

Diana Hughes (diana.hughes@aofl.com) is a game designer and creative leader specializing in serious games, especially educational games. She believes in making strong connections between gameplay and curriculum—by interacting with the game's mechanics, you're directly interacting with the material. Her projects have included research games for children on the autism spectrum, training games for the US Military, a graphics-free mobile game for blind players, educational games for ABCmouse.com, and Mastering Math, a game-based personalized learning system for children 4–8 years of age.

G. Tanner Jackson (gtjackson@ets.org) is a research director at Educational Testing Service and leads agile innovation teams that employ lean start-up practices to iterate and develop capabilities focusing on end users and marketplace applications. He is an accomplished cognitive scientist with prior funding from NSF, IES, and ONR and has over 100 published articles that focus on his work with intelligent tutoring systems, educational games, and innovative assessments. His research centers around adaptive and engaging educational environments that involve aspects of game design, natural language processing, and adaptive conversations. Additionally, he is interested in how users interact with complex systems and leveraging these environments to examine user interactions over time.

Fengfeng Ke (fke@fsu.edu) is an associate professor of the Educational Psychology and Learning Systems for Florida State University. Her current research focuses on game-based learning, inclusive design of computer-supported collaborative learning, and mixed-reality integrated immersive learning.

S. Kathleen Krach (skrach@fsu.edu), Ph.D., NCSP, is an assistant professor at Florida State University in the School Psychology Program. Her current research focuses on the use of game-based technology to develop assessments and interventions related to socio-emotional learning (specifically aggression) that are culturally sensitive and available across language groups. Dr. Krach is licensed as a psychologist in the states of Florida, Nevada, and Alabama and holds the credential of Nationally Certified School Psychologist.

Jeffrey R. Laferriere (laferrie@lvc.edu) is an associate professor and director of Secondary Education at Lebanon Valley College. Dr. Laferrier completed his doctoral work in Teacher Education at the University of Nevada, Las Vegas, in 2019. He specialized in using innovative approaches with technology and games to help preservice teachers define, identify, and intervene in bullying situations. In 2007, he was awarded a B.A. in Secondary Education: Mathematics from Arizona State University. He was awarded an M.A. in Education & Human Development: Educational Leadership & Administration from George Washington University in 2010. From 2015 until his graduation, he served as a graduate assistant at the University of Nevada, Las Vegas. Previously, he worked for 6 years as a high school math teacher and 2 years as a middle school assistant principal.

Liping Wu (wu_liping@rp.edu.sg) holds a master's degree in Industrial & Systems Engineering, National University of Singapore. She is currently the assistant program chair for the diploma in Supply Chain Management at Republic Polytechnic. She has more than 10 years of PBL facilitation experience, and she is an RP-certified specialist problem crafter. She works closely with her colleagues to review the SCM curriculum regularly to ensure that it is updated according to industry needs, with the adoption of a diversified range of teaching and learning strategies (e.g., lab lessons, supply chain games, etc.) to further enhance students' engagement and experience in their learning. As part of the curriculum development and updating, she constantly explores suitable supply chain games to be incorporated to enhance teaching and learning of RP SCM students.

Michael P. McCreery (michael.mccreery@unlv.edu) is an associate professor and codirector of the Interaction & Media Sciences Lab at the University of Nevada, Las Vegas. His research focuses on how principles of human–computer interaction can be applied to the creation of psychoeducational assessments and interventions.

Robert M. Mislevy (rmmislevy@ets.org) is Frederic M. Lord Chair in Measurement and Statistics at Educational Testing Service and emeritus professor at the University of Maryland. His research applies developments in technology, statistics, and cognitive science to practical problems in assessment. His work includes simulation-based assessment of network engineering with Cisco Systems and an evidence-centered assessment design framework. His recent publications include *Bayesian psychometric modeling, Bayesian networks in educational assessment, Psychometric considerations for game-based assessment,* and *Sociocognitive foundations of educational measurement.* He has received career contribution awards from AERA and NCME and four NCME Annual Awards for Technical Contributions to Measurement. He is a member of the National Academy of Education and past president of the Psychometric Society.

Steve Nebel (steve.nebel@phil.tu-chemnitz.de) is a researcher investigating educational video games or serious games. He graduated in the field of Media and Instructional Psychology and conducts his research at the chair of Psychology of Learning with Digital Media at the Technical University Chemnitz. Furthermore, he gives lectures in instructional psychology, educational videogame design, and further aspects of multimedia.

John L. Nietfeld (john_nietfeld@ncsu.edu) is a professor of Education Psychology in the Department of Teacher Education and Learning Sciences at North Carolina State University. He completed his Ph.D. in educational psychology from the University of Nebraska-Lincoln in 1999. Dr. Nietfeld's teaching and research interests lie in self-regulated learning (SRL) and metacognition. He has coauthored several journal articles centering on supporting self-regulation, improving metacognitive skills, investigations of metacognitive monitoring in classroom settings, and studies of motivation, including self-efficacy and the impact of extrinsic rewards. He has served on the editorial board for the *Journal of Educational Psychology, Journal of Experimental Education,* and *Metacognition and Learning*. His research has been supported by numerous internal and external funded projects. His current work focuses on how to create intelligent game-based learning environments that will facilitate and scaffold SRL skills for upper elementary school students in science.

Manuel Ninaus (m.ninaus@iwm-tuebingen.de) is a postdoctoral researcher at the Leibniz-Institut fuer Wissensmedien, Tuebingen, Germany, and obtained his Ph.D. in Neuropsychology from the University of Graz (2015). His research focuses on interactive learning technologies, such as game-based learning and neuro/biofeedback systems. He is an elected board member of the Serious Games Society and currently investigates the cognitive, emotional, and neurofunctional processes of game-based learning with a particular focus on numerical cognition.

Andreas Oranje (aoranje@ets.org) is general manager at Educational Testing Service, directing centers that conduct foundational and applied research on natural language processing, automated scoring of spoken and written work, multimodal dialogue systems, cognitive sciences, virtual performance assessments, and assistive technology. His past research focused on the psychometric design and statistical analysis of the National Assessment of Educational Progress, design of and evidence from game- and scenario-based assessments, and adaptive testing design for group score assessments. He led NAEP's psychometric research agenda from 2007 to 2012 and directed design, analysis, and reporting from 2013 to 2017. He was the 2019 program chair for AERA's Division D. His recent work focuses on applications of artificial intelligence to education, presenting at venues such as SXSWedu, NSVF, and ISTE.

V. Elizabeth Owen (v.elizabeth.owen@gmail.com) holds a Ph.D. in Digital Media (School of Education, UW-Madison) focused on game-based learning analytics. Currently director of Learning and Data Science at Age of Learning, she's committed to leveraging learning analytics to optimize adaptive, engaging learning systems. Previously a data scientist and learning designer with GlassLab Games, LRNG, and Metacog, Elizabeth's doctoral roots lie with Games+Learning+Society research and game development lab. Collaborators include EA, Zynga, and Popcap games, and Dr. Ryan Baker at UPenn in an ongoing Educational Data Mining. Prior to graduate school, Dr. Owen was a K-12 educator for a decade and a founding teacher at an L.A. charter school (LAAAE.org).

Biswas Parajuli (bparajuli@fsu.edu) is a doctoral student at the Department of Computer Science of Florida State University. His current research focuses on machine learning algorithms and applications.

Xinyue Ren (xr868414@ohio.edu) is a Ph.D. candidate in the Instructional Technology program at Ohio University. She earned her master's degree in education with specialization in critical studies in education. Her research foci include game-based learning (GBL), emerging technologies, online learning and teaching, and mobile learning. She is also an active member in the field of educational technology and has expertise in program evaluation, multimedia learning, and instructional design. She has participated in extensive studies and projects about the affordances of emerging technologies in enhancing learners' learning experiences, including using digital games to promote social justice education, social media to enhance training experience, and universal design for learning in online learning environments.

Louisa Rosenheck (louisa@mit.edu) is a designer and researcher of educational technologies in the MIT Education Arcade and Playful Journey Lab. She manages the design, content, and development of educational games and simulations and oversees the research on how games can be effectively used in both formal and informal educational settings. She also brings the playfulness of games into her assessment design work, creating digital and nondigital tools to help both students and teachers assess creative work and soft skills. In addition to these areas, she supports other designers and organizations in creating innovative educational technology and curriculum. Through capacity building partnerships on local and international ed tech programs, she aims to increase the deep learning in games and tools being produced for education.

José A. Ruipérez-Valiente (jruipere@mit.edu) completed his B.Eng. and M.Eng. in Telecommunications at Universidad Católica de San Antonio de Murcia (UCAM) and Universidad Carlos III of Madrid (UC3M), respectively, graduating in both cases at the top of his class. Afterward, he completed his M.Sc. and Ph.D. in Telematics at UC3M while conducting research in learning analytics and educational data mining at Institute IMDEA Networks. During this time, he completed two research stays of three months each, the first at MIT and the second at the University of Edinburgh. He has received more than a dozen academic and research awards and has published over 35 scientific publications in high impact venues.

Currently, he is a postdoctoral associate at the CMS/W Department at MIT, where he is part of the Teaching Systems Lab and Playful Journey Lab, and also collaborates with the Education Arcade in applying data science to large-scale free online courses and game-based environments to enhance human knowledge on how we learn.

P. G. Schrader (pg.schrader@unlv.edu) is a chair of the Department of Teaching and Learning, professor of Interaction and Media Sciences, and research scientist at the Center for Research, Evaluation, and Assessment at UNLV. Dr. Schrader completed his degree in Educational Psychology with a focus on Cognition and Instruction at the University of Connecticut. His work has broached the topic of learning in technology rich, immersive environments across disciplines, and he has examined the exchange of information within hypertext, multimedia, and massively multiplayer online games. Dr. Schrader's research has addressed a variety of contexts, methodologies, and content areas. Current projects include the examination of digital games and virtual environments from a human–computer interaction perspective. Dr. Schrader has published a wide range of studies that cover inter- and intratextuality, navigation, and learning. Dr. Schrader has received awards honoring his commitment to academics, the community, and higher education in general.

Mamta Shah (mamta@drexel.edu) is a postdoctoral scholar in the Games and Learning in Interactive Digital Environments (GLIDE) Lab in the School of Education at Drexel University, USA. Her research expertise resides at the intersection of learning technologies, learning sciences, teacher education, and educational psychology. Dr. Shah focuses on the theoretical and practical applications of teaching, learning, and assessing with interactive and interdisciplinary technologies, such as games, in different settings, including schools, workplaces, museums, and online environments. Her research has contributed to advancements in theory and practice on teaching and learning with technology by way of model development and testing, analysis of cognitive and pedagogic affordances of digital environments, design and integration of immersive and interdisciplinary environments, and teacher professional development.

Danial Smith (dsmith@fsu.edu) is a doctoral student at the Department of Scientific Computing of Florida State University. His current research focuses on digital games design and machine learning.

Wee Hoe Tan (whtan@fskik.upsi.edu.my) is an associate professor at Sultan Idris Education University in Malaysia and the vice chair of Serious Games Asia Society based in Singapore. Formally a 3D animator, he joined academia in mid-2004 to teach and develop multimedia courses in Kuala Lumpur University. In 2007, he pursued a Ph.D. at the University of Warwick, UK. Toward the end of his doctoral study, he began working as a consultant in the game industry. He was involved in the game coaching system design of the FPS Trainer project in Dundee and the game-based learning system design of the KNeoWORLD Project in San Francisco while promoting game-based learning in Southeast Asia through the Serious Games Asia Society. In 2017, Wee Hoe joined the Game Lab at the Singapore University of Technology and Design as a visiting faculty. He later won the Hubert Humphrey Fellowship award, making him a Humphrey Fellow in the College of Education, Pennsylvania State University, and a visiting research scientist at the Play2Prevent Lab in Yale University.

Vipin Verma (vverma9@asu.edu) is pursuing his Ph.D. in Human Systems Engineering from Arizona State University. His work focuses on integrating content-agnostic stealth assessment in educational video games. In the past, he worked for 5 years as a software engineer in various companies before beginning his Ph.D. in 2017.

Linda William (lwilliam@tp.edu.sg) is a senior lecturer in School of Informatics and IT. She obtained her Ph.D. (Information Systems) from School of Information Systems, Singapore Management University. Her research interests are in intelligent systems, decisions and data analytics, machine learning, metaheuristics algorithm, and serious game in education. She has published a book and several articles in international conferences and academic journals, such as the *Journal of the Operation Research Society* and *Journal of Urban Sciences*.

Kayleen Wood (kayleen.wood@scu.edu.au) is a lecturer in Taxation Law and Practice and Financial Accounting at Southern Cross University, Australia, where she is also a Ph.D. candidate. As a teaching and learning academic, her research centers on student engagement with and motivation for learning. She has brought together her qualifications in business and education, and a background in professional accounting, academic project management, and teacher education, for her Ph.D. thesis: a gamified learning experience for accounting education. As the recipient of a Vice Chancellor's Citation for Excellence in Student Engagement, her research is already being operationalized in curriculum design and teaching resources in the School of Business and Tourism, Southern Cross University.

Xinhao Xu (xuxinhao@missouri.edu) is an assistant professor in the School of Information Science and Learning Technologies at the University of Missouri-Columbia. He has been actively conducting research and teaching courses in technology-enhanced learning and instruction. His current research interests mainly focus on human–computer interactions for learning, immersive virtual learning environments, and game-based learning for STEM subjects. He is a member of major academic organizations such as Institute of Electrical and Electronics Engineers (IEEE), Association for Educational Communications and Technology (AECT), and American Educational Research Association (AERA).

Part I
Foundations of Game-Based Assessment

Chapter 1
Game-Based Assessment: The Past Ten Years and Moving Forward

Yoon Jeon Kim and Dirk Ifenthaler

1.1 Introduction

Educational assessment practice is challenging as there are a number of diverse concepts referring to the idea of assessment. Newton (2007) laments that the distinction between formative and summative assessment hindered the development of sound assessment practices on a broader level. Black (1998) defines three main types of assessment: (a) formative assessment to aid learning; (b) summative assessment for review, for transfer, and for certification; and (c) summative assessment for accountability to the public. Pellegrino, Chudowsky, and Glaser (2001) extend these definitions with three main purposes of assessment: (a) assessment to assist learning (formative assessment), (b) assessment of individual student achievement (summative assessment), and (c) assessment to evaluate programs (evaluative assessment). A common thread among the many definitions points to the concept of feedback for a variety of purposes, audiences, and methods of assessment (Ifenthaler, Greiff, & Gibson, 2018).

Digital game-based technologies are nudging the field of education to redefine what is meant by learning, instruction, and assessment. Proponents of game-based learning argue that students should be prepared to meet the demands of the twenty-first century by teaching them to be innovative, creative, and adaptable so that they can deal with the demands of learning in domains that are complex and ill-structured (Federation of American Scientists, 2005; Gee, 2003; Ifenthaler, Eseryel, & Ge,

Y. J. Kim
Teaching Systems Lab, Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: yjk7@mit.edu

D. Ifenthaler (✉)
Learning, Design and Technology, University of Mannheim, Mannheim, Germany

Curtin University, Perth, Australia
e-mail: dirk@ifenthaler.info

D. Ifenthaler, Y. J. Kim (eds.), *Game-Based Assessment Revisited*, Advances in Game-Based Learning, https://doi.org/10.1007/978-3-030-15569-8_1

2012; Prensky, 2001; Shaffer, 2006). On the other hand, opponents of games argue that games are just another technological fad, which emphasize superficial learning. In addition, opponents argue that games cause increased violence, aggression, inactivity, and obesity while decreasing prosocial behaviors (Walsh, 2002).

However, Ifenthaler et al. (2012) argue that the implementation of assessment features into game-based learning environments is only in its early stages because it adds a very time-consuming step to the design process. Also, the impact on learning and questions toward reliability and validity of technology-based assessment systems are still being questioned. Three distinguishing features of game-based assessment have been proposed and are widely accepted: (1) game scoring, (2) external, and (3) embedded assessment of game-based learning (Ifenthaler et al., 2012). Only recently, an additional feature has been introduced which enables adaptive gameplay and game environments, broadly defined as learning analytics (Ifenthaler, 2015) and specifically denoted as serious games analytics (Loh, Sheng, & Ifenthaler, 2015). Serious games analytics converts learner-generated information into actionable insights for real-time processing. Metrics for serious games analytics are similar to those of learning analytics including the learners' individual characteristics (e.g., socio-demographic information, interests, prior knowledge, skills, and competencies) and learner-generated game data (e.g., time spent, obstacles managed, goals or tasks completed, navigation patterns, social interaction, etc.) (Ge & Ifenthaler, 2017; Ifenthaler, 2015; Loh, Sheng, & Ifenthaler, 2015).

This chapter seeks to identify why research on game-based assessment is still in its infancy, what advances have been achieved over the past 10 years, and which challenges lie ahead for advancing assessment in game-based learning.

1.2 Game-Based Assessment and Assessment of Learning in Games: Why?

Games—both digital and nondigital—have become an important aspect of young people's life. According to a recent survey conducted in the United States, 72% of youth ages 13–17 play games daily or weekly (Lenhart, 2015). Gaming is also one of the most popular social activities, especially for boys, where 55% of them play games in-person or online with friends daily or weekly. While gaming gained more popularity in people's daily life, starting in early 2000, educational researchers began to investigate potential educational benefits of games for learning and what we can learn from well-designed games about learning and assessment (Gee, 2003).

So what are affordances of games for learning? First, people learn in action in games (Gee, 2008). That is, people interact with all aspects of the game and take intentional actions within the game. For its part, the game continuously responds to each action, and through this process, the player gradually creates meaning. Clearly, how people are believed to learn within video games contrasts to how people typically learn at school, which often entails memorization of decontextualized and abstract concepts and procedures (Shute, Ventura, Bauer, & Zapata-Rivera, 2009). Second, due to its interactive nature, learning by playing games can lead to conceptual understanding and problem-solving (Eseryel, Ge, Ifenthaler, & Law, 2011) in addition to domain-specific skills and practices (Bressler & Bodzin, 2016) that go beyond the basic content knowledge more commonly taught in the classroom. Steinkuehler and Duncan (2008) have found players in virtual worlds frequently engaging in social knowledge construction, systems-based reasoning, and other scientific habits of mind. This body of work shows that games in general have a lot of potential for contributing to a deep learning environment. In video games, players engage in active and critical thinking, they take on different identities, and they have opportunities to practice skills and find intrinsic rewards as they work on increasingly difficult challenges on their path to mastery (Eseryel, Law, Ifenthaler, Ge, & Miller, 2014; Gee, 2003).

Numerous studies have reported the benefits of games for learning as a vehicle to support student learning. In a meta-analysis study, Clark, Tanner-Smith, and Killingsworth (2016) reported that compared to nongame conditions, digital games had a moderate to strong effect in terms of overall learning outcomes including cognitive and interpersonal skills. Similarly, a literature review by Boyle et al. (2016) reports that games are beneficial for learning of various outcomes such as knowledge acquisition, affect, behavior change, perception, and cognition. Numerous studies also reported academic domain-specific benefits of games for learning including science and mathematics (Divjak & Tomić, 2011). To answer the question of what people are learning from playing games, researchers have been using a variety of methods including external measures, log data capturing in-game actions, and game-related actions beyond the game context (Ifenthaler et al., 2012; Loh et al., 2015).

1.3 Game-Based Assessment: Past 10 Years

Several meta-analyses have been published focusing on game-based learning. For example, Baptista and Oliveira (2019) highlight important variables in their literature search of more than 50 studies focusing on serious games including intention, attitude, enjoyment, and usefulness. A systematic review by Alonso-Fernández, Calvo-Morata, Freire, Martínez-Ortiz, and Fernández-Manjón (2019) focuses on the application of data science techniques on game learning data and suggests specific game learning analytics. Ke (2016) presents a systematic review on the integration of domain-specific learning in game mechanics and game world design. Another systematic review by Ravyse, Seugnet Blignaut, Leendertz, and Woolner (2017) identifies five central themes of serious games: backstory and production, realism, artificial intelligence and adaptivity, interaction, and feedback and debriefing. Accordingly, none of the abovementioned meta-analyses and systematic reviews have a clear focus on assessment of game-based learning.

Still, a line of research that emerged over the past 10 years was in relation to the question of how we can use games as an interactive and rich technology-enhanced environment to advance assessment technologies. That is, the primary goal of this line is to advance assessment using games (Ifenthaler et al., 2012). Earlier game-based assessment work has primarily focused on applying the evidence-centered design framework to develop assessment models with specific learning outcomes and skills in mind (Behrens, Mislevy, Dicerbo, & Levy, 2012). For example, Shute et al. (2009) describe an approach called stealth assessment—where in-game behavioral indicators (e.g., specific actions taken within a quest in Oblivion) are identified and make inferences about the player's underlying skills (e.g., creative problem-solving) without the flow of gameplay using logged data. Using this approach, one can use existing games to measure latent constructs, even if the game was not explicitly developed for the purpose of learning or assessment, as long as the game provides ample contexts (or situations) that elicit evidence for underlying skills and constructs (Loh et al., 2015). Similarly, using a popular game SimCity, GlassLab developed SimCityEDU to assess students' systems thinking (Dicerbo et al., 2015). These approaches have primarily used the evidence-centered design framework (Almond, Steinberg, & Mislevy, 2002) to align what people might learn from the game with what they do in games.

Eseryel, Ifenthaler, and Ge (2011) provide an integrated framework for assessing complex problem-solving in digital game-based learning in the context of a longitudinal design-based research study. In a longitudinal field study, they examined the impact of the massively multiplayer online game (MMOG) *Surviving in Space* on students' complex problem-solving skill acquisition, mathematics achievement, and students' motivation. Two different methodologies to assess student's progress of learning in complex problem-solving were applied. The first methodology utilized adapted protocol analysis (Ericsson & Simon, 1980, 1993) to analyze students' responses to the given problem scenario within the framework of the think-aloud methodology. The second methodology utilized HIMATT methodology (Eseryel, Ifenthaler, & Ge, 2013; Pirnay-Dummer, Ifenthaler, & Spector, 2010) to analyze students' annotated causal representations of the phenomena in question. The automated text-based analysis function of HIMATT enables the tracking of the association of concepts from text which contain 350 or more words directly, hence producing an adaptive assessment and feedback environment for game-based learning. For future game design, the algorithms produce quantitative measures and graphical representations which could be used for instant feedback within the game or for further analysis (Ifenthaler, 2014).

More recently, researchers have introduced learning analytics and data mining techniques to broaden what game-based assessment means (Loh et al., 2015). For example, Rowe et al. (2017) built "detectors" machine-learned algorithm using log data in the game to measure implicit understanding of physics, different strategies associated with productivity in the game, and computational thinking. While they did not use formal measure models (e.g., IRT or Bayes net), these detectors are implemented in the game engine to make real-time inferences of players. Similarly, Shadowspect developed at MIT Playful Journey Lab (Kim & Rosenheck, 2018) is another example of GBA that utilizes new advancements in learning analytics and educational data mining techniques in the process of game design and development for the purpose of assessment.

Hence, the application of serious games analytics opens up opportunities for the assessment of engagement within game-based learning environments (Eseryel et al., 2014). The availability of real-time information about the learners' actions and behaviors stemming from key decision points or game-specific events provides insights into the extent of the learners' engagement during gameplay. The analysis of single action or behavior and the investigation of more complex series of actions and behaviors can elicit patterns of engagement and therefore provide key insights into learning processes (Ge & Ifenthaler, 2017).

Ifenthaler and Gibson (2019) report how highly detailed data traces, captured by the Challenge platform, with many events per learning activity and when combined with new input devices and approaches bring the potential for measuring indicators of physical, emotional, and cognitive states of the learner. The data innovation of the platform is the ability to capture event-based records of the higher-frequency and higher-dimensional aspects of learning engagement, which is in turn useful for analysis of the effectiveness and impact on the physical, emotional, and cognitive layers of learning caused or influenced by the engagements. This forms a high-resolution analytics base on which research into digital learning and teaching as well as into how to achieve better outcomes in scalable digital learning experiences can be conducted (Gibson & Jackl, 2015).

1.4 Challenges and Future Work

While interests for game-based assessment peaked in 2009 when the GlassLab was launched to scale up this approach in the broad education system, many promises of game-based learning and assessment have not fully accomplished in the actual education system. Based on the reflection of the fields' achievements in the past 10 years and contributions to the current volume, challenges remain that the field of game-based assessment still faces as well as future work that researchers, game designers, and educators should address to transform how games are used in the education system.

While ECD has been the most predominant framework to design assessment in games, it is often unclear how different development processes leverage ECD to conceptualize game design around the competency of interest (Ke, Shute, Clark, & Erlebacher, 2019). For example, how can assessment models be formalized? How can formalized assessment models be translated to game design elements? When in the game design process, does this translation occur most effectively? How can competency models be transformed into interesting, engaging game mechanics? How can psychometric qualities be ensured without being too prescriptive?

Many established game-based assessment approaches focus on understanding the continuous progression of learning, thinking, reasoning, argumentation, and complex problem-solving during digital game-based learning. From a design perspective, it seems important that the game mechanisms address the underlying affective, behavioral, and cognitive dispositions which must be assessed carefully at various stages of the learning process, hence, while conceptualizing and designing games for learning (Bertling, Jackson, Oranje, & Owen, 2015; Eseryel et al., 2014; Ge & Ifenthaler, 2017).

Advanced data analytics methodologies and technological developments enable researchers, game designers, and educators to easily embed assessment and analysis techniques into game-based learning environments (Loh et al., 2015). Internal assessment and instant analysis including personalized feedback can be implemented in a new generation of educational games. However, it is up to educational research to provide theoretical foundations and empirical evidence on how these methodologies should be designed and implemented. We have just arrived in the age of educational data analytics. Hence, it is up to researchers, technologists, educators, and philosophers to make sense of these powerful technologies, thus better help learners to learn.

With the challenges brought on by game-based assessments including data analytics, the large amount of data now available for teachers is far too complex for conventional database software to store, manage, and process. Accordingly, analytics-driven game-based assessments underscore the need to develop assessment literacy in stakeholders of assessment (Ifenthaler et al., 2018; Stiggins, 1995). Game designers and educators applying data-driven game-based assessments require practical hands-on experience on the fundamental platforms and analysis tools for linked big game-based assessment data. Stakeholders need to be introduced to several data storage methods and how to distribute and process them, introduce possible ways of handling analytics algorithms on different platforms, and highlight visualization techniques for game-based assessment analytics (Gibson & Ifenthaler, 2017). Well-prepared stakeholders may demonstrate additional competencies such as understanding large-scale machine learning methods as foundations for human-computer interaction, artificial intelligence, and advanced network analysis (Ifenthaler et al., 2018).

The current research findings also indicate that design research and development are needed in automation and semi-automation (e.g., humans and machines working together) in assessment systems. Automation and semi-automation of assessments to provide feedback, observations, classifications, and scoring are increasingly being used to serve both formative and summative purposes in game-based learning.

Gibson, Ifenthaler, and Orlic (2016) proposed an open assessment resources approach that has the potential to increase trust in and use of open education resources (OER) in game-based learning and assessment by adding clarity about assessment purposes and targets in the open resources world. Open assessment resources (OAR) with generalized formative feedback are aligned with a specific educative purpose expressed by some user of a specific OER toward the utility and expectations for using that OER to achieve an educational outcome. Hence, OAR may be utilized by game designers to include valuable and competence-based assessments in game-based learning.

The application of analytics-driven game-based assessments opens up opportunities for the assessment of engagement and other motivational (or even broader: non-cognitive) constructs within game-based learning environments (Eseryel et al., 2014). The availability of real-time information about the learners' actions and behaviors stemming from key decision points or game-specific events provides insights into the extent of the learners' engagement during gameplay. The analysis of single action or behavior and the investigation of more complex series of actions and behaviors can elicit patterns of engagement and therefore provide key insights into ongoing learning processes within game-based learning environments.

To sum up, the complexity of designing adaptive assessment and feedback systems has been discussed widely over the past few years (e.g., Sadler, 2010; Shute, 2008). The current challenge is to make use of data—from learners, teachers, and game learning environments—for assessments. Hence, more research is needed to unveil diverse methods and processes related to how design teams, often including learning scientists, subject-matter experts, and game designers, can seamlessly integrate design thinking and the formalization of assessment models into meaningful assessment for game-based learning environments.

References

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four process architecture. *Journal of Technology, Learning, and Assessment, 1*(5), 3–63.

Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Applications of data science to game learning analytics data: A systematic literature review. *Computers & Education, 141*, 103612. https://doi.org/10.1016/j.compedu.2019.103612

Baptista, G., & Oliveira, T. (2019). Gamification and serious games: A literature meta-analysis and integrative model. *Computers in Human Behavior, 92*, 306–315.

Behrens, J., Mislevy, R., Dicerbo, K., & Levy, R. (2012). Evidence centered design for learning and assessment in the digital world. In M. Mayrath, J. Clarke-Midura, D. Robinson, & G. Schraw (Eds.), *Technology-based assessments for 21st century skills* (pp. 13–54). Charlotte, NC: Information Age Publishers.

Bertling, M., Jackson, G. T., Oranje, A., & Owen, V. E. (2015). Measuring argumentation skills with game-based assessments: Evidence for incremental validity and learning. In C. Conati, N. Heffernan, A. Mitrovic, & M. Verdejo (Eds.), *Artificial intelligence in education. AIED 2015* (Vol. 9112, pp. 545–549). Cham, Switzerland: Springer.

Black, P. J. (1998). *Testing: Friend or foe? The theory and practice of assessment and testing*. London, UK: Falmer Press.

Boyle, E. A., Hainey, T., Connolly, T. M., Grant, G., Earp, J., Ott, M., … Pereira, J. (2016). An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. *Computers & Education, 94*, 178–192. https://doi.org/10.1016/j.compedu.2015.11.003

Bressler, D. M., & Bodzin, A. M. (2016). A mixed methods assessment of students' flow experiences during a mobile augmented reality science game. *Journal of Computer Assisted Learning, 29*, 505–517. https://doi.org/10.1111/jcal.12008

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research, 86*(1), 79–122. https://doi.org/10.3102/0034654315582065

Dicerbo, K., Bertling, M., Stephenson, S., Jia, Y., Mislevy, R. J., Bauer, M., & Jackson, G. T. (2015). An application of exploratory data analysis in the development of game-based assessments. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics. Methodologies for performance measurement, assessment, and improvement* (pp. 319–342). New York, NY: Springer.

Divjak, B., & Tomić, D. (2011). The impact of game-based learning on the achievement of learning goals and motivation for learning mathematics - Literature review. *Journal of Information and Organizational Sciences, 35*(1), 15–30.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*, 215–251.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Eseryel, D., Ge, X., Ifenthaler, D., & Law, V. (2011). Dynamic modeling as cognitive regulation scaffold for complex problem solving skill acquisition in an educational massively multiplayer online game environment. *Journal of Educational Computing Research, 45*(3), 265–287.

Eseryel, D., Ifenthaler, D., & Ge, X. (2011). Alternative assessment strategies for complex problem solving in game-based learning environments. In D. Ifenthaler, P. I. Kinshuk, D. G. Sampson, & J. M. Spector (Eds.), *Multiple perspectives on problem solving and learning in the digital age* (pp. 159–178). New York, NY: Springer.

Eseryel, D., Ifenthaler, D., & Ge, X. (2013). Validation study of a method for assessing complex ill-structured problem solving by using causal representations. *Educational Technology Research and Development, 61*(3), 443–463. https://doi.org/10.1007/s11423-013-9297-2

Eseryel, D., Law, V., Ifenthaler, D., Ge, X., & Miller, R. B. (2014). An investigation of the interrelationships between motivation, engagement, and complex problem solving in game-based learning. *Journal of Educational Technology & Society, 17*(1), 42–53.

Federation of American Scientists. (2005). *Summit of educational games: Harnessing the power of video games for learning*. Washington, DC: Author.

Ge, X., & Ifenthaler, D. (2017). Designing engaging educational games and assessing engagement in game-based learning. In R. Zheng & M. K. Gardner (Eds.), *Handbook of research on serious games for educational applications* (pp. 255–272). Hershey, PA: IGI Global.

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

Gee, J. P. (2008). Learning and games. The ecology of games: Connecting youth, games, and learning. In K. Salen (Ed.), *The John D. and Catherine T. MacArthur Foundation series on digital media and learning* (pp. 21–40). Cambridge, MA: MIT.

Gibson, D. C., & Ifenthaler, D. (2017). Preparing the next generation of education researchers for big data in higher education. In B. K. Daniel (Ed.), *Big data and learning analytics: Current theory and practice in higher education* (pp. 29–42). New York, NY: Springer.

Gibson, D. C., Ifenthaler, D., & Orlic, D. (2016). Open assessment resources for deeper learning. In P. Blessinger & T. J. Bliss (Eds.), *Open education: International perspectives in higher education* (pp. 257–279). Cambridge, UK: Open Book Publishers.

Gibson, D. C., & Jackl, P. (2015). Theoretical considerations for game-based e-learning analytics. In T. Reiners & L. Wood (Eds.), *Gamification in education and business* (pp. 403–416). New York, NY: Springer.

Ifenthaler, D. (2014). AKOVIA: Automated knowledge visualization and assessment. *Technology, Knowledge and Learning, 19*(1–2), 241–248. https://doi.org/10.1007/s10758-014-9224-6

Ifenthaler, D. (2015). Learning analytics. In J. M. Spector (Ed.), *The SAGE encyclopedia of educational technology* (Vol. 2, pp. 447–451). Thousand Oaks, CA: Sage.

Ifenthaler, D., Eseryel, D., & Ge, X. (2012). Assessment for game-based learning. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning. Foundations, innovations, and perspectives* (pp. 3–10). New York, NY: Springer.

Ifenthaler, D., & Gibson, D. C. (2019). Opportunities of analytics in challenge-based learning. In A. Tlili & M. Chang (Eds.), *Data analytics approaches in educational games and gamification systems*. Cham, Switzerland: Springer.
Ifenthaler, D., Greiff, S., & Gibson, D. C. (2018). Making use of data for assessments: Harnessing analytics and data science. In J. Voogt, G. Knezek, R. Christensen, & K.-W. Lai (Eds.), *International handbook of IT in primary and secondary education* (2nd ed., pp. 649–663). New York, NY: Springer.
Ke, F. (2016). Designing and integrating purposeful learning in game play: A systematic review. *Educational Technology Research and Development, 64*(2), 219–244. https://doi.org/10.1007/s11423-015-9418-1
Ke, F., Shute, V. J., Clark, K. M., & Erlebacher, G. (2019). *Interdisciplinary design of game-based learning platforms*. Cham, Switzerland: Springer.
Kim, Y. J., & Rosenheck, L. (2018). *A playful assessment approach to research instrument development*. Paper presented at the Thirteenth International Conference of the Learning Sciences, London, UK.
Lenhart, A. (2015). *Teen, social media and technology overview 2015*. Washington, DC: Pew Research Center.
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* (pp. 3–29). New York, NY: Springer.
Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice, 14*(2), 149–170. https://doi.org/10.1080/09695940701478321
Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students knwo: The science and design of educational assessment*. Washington, DC: National Academy Press.
Pirnay-Dummer, P., Ifenthaler, D., & Spector, J. M. (2010). Highly integrated model assessment technology and tools. *Educational Technology Research and Development, 58*(1), 3–18. https://doi.org/10.1007/s11423-009-9119-8
Prensky, M. (2001). *Digital game-based learning*. New York, NY: McGraw-Hill.
Ravyse, W. S., Seugnet Blignaut, A., Leendertz, V., & Woolner, A. (2017). Success factors for serious games to enhance learning: A systematic review. *Virtual Reality, 21*(1), 31–58.
Rowe, E., Asbell-Clarke, J., Baker, R. S., Eagle, M., Hicks, A., Barnes, T., … Edwards, T. (2017). Assessing implicit science learning in digital games. *Computers in Human Behavior, 76*, 617–630. https://doi.org/10.1016/j.chb.2017.03.043
Sadler, D. R. (2010). Beyond feedback: developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education, 35*(5), 535–550.
Shaffer, D. W. (2006). *How computer games help children learn?* New York, NY: Palgrave Macmillan.
Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.
Shute, V. J., Ventura, M. I., 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. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). New York, NY: Routledge.
Steinkuehler, C., & Duncan, S. (2008). Scientific habits of mind in virtual worlds. *Journal of Science Education and Technology, 17*(6), 530–543.
Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan, 77*(3), 238–245.
Walsh, D. (2002). *Video game violence and public policy*. Retrieved from http://www.soc.iastate.edu/sapp/videogames2.pdf

Chapter 2
Assessing Learning *from*, *with*, and *in* Games Revisited: A Heuristic for Emerging Methods and Commercial Off-the-Shelf Games

P. G. Schrader, Michael P. McCreery, Mark C. Carroll, Danielle L. Head, and Jeffrey R. Laferriere

2.1 Introduction

Since the early and formal study of humans' interactions with media and technology (Reiser, 2001), tools and systems have evolved and are becoming more dynamic, emergent, and complex (Carroll & Campbell, 1999; Hilpert & Marchand, 2018; Schrader, 2008). Likewise, the field has expanded its views on humans and their interactions with technological systems like video games (Krach & McCreery, 2015; Schrader, McCreery, & Vallett, 2017). Considerable effort has been exerted into understanding how people learn from, with, and within game-based environments (Jonassen, Campbell, & Davidson, 1994; Salomon, Perkins, & Globerson, 1991; Schrader, 2008). Although there are numerous instances of arguments that extoll games' potential, examples of innovation, and quasi-studies, researchers have noted broad issues of quality, rigor, and generalizability when it comes to game studies (Clark, Tanner-Smith, & Killingsworth, 2014; Ke, 2009; Vogel, Vogel, Cannon-Bowers, Muse, & Wright, 2006; Wouters, Van Nimwegen, Van Oostendorp, & Van der Spek, 2013; Young et al., 2012).

There are numerous studies, reviews, and discussions involving video games in relation to their educative merits (Bediou et al., 2018; Clark et al., 2014; Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012; Girard, Ecalle, & Magnan, 2013; Ke, 2009; Mayer, 2015; Wouters et al., 2013; Young et al., 2012). Typically, games are viewed as a vehicle for authentic activity, learning, and performance (Barab,

P. G. Schrader (✉) · M. P. McCreery · M. C. Carroll · D. L. Head
University of Nevada, Las Vegas, Las Vegas, NV, USA
e-mail: pg.schrader@unlv.edu; michael.mccreery@unlv.edu; mcarroll@unlv.nevada.edu; headd1@unlv.nevada.edu

J. R. Laferriere
Lebanon Valley College, Annville, PA, USA
e-mail: laferrie@lvc.edu

D. Ifenthaler, Y. J. Kim (eds.), *Game-Based Assessment Revisited*, Advances in Game-Based Learning, https://doi.org/10.1007/978-3-030-15569-8_2

Gresalfi, & Ingram-Goble, 2010; Shaffer, Squire, Halverson, & Gee, 2005; Squire, 2006). There are far fewer definitions, or succinct descriptions, of video games and their characteristics (McGonigal, 2011; O'Brien, Lawless, & Schrader, 2010; Schrader & McCreery, 2012). On the one hand, this invites researchers to carefully and operationally describe the contexts they examine. On the other, some have argued that imprecise or vague operational definitions invite miscommunication and a general inability to broaden understanding of any scientific discipline (King, Young, Drivere-Richmond, & Schrader, 2001).

By most accepted perspectives, video games are complex systems. According to Hilpert and Marchand (2018), a complex system is one that is comprised of interactive elements, entities, processes, or agents. Further, complex systems are dynamic, and each of the elements, entities, processes, or agents interacts with each other in meaningful and potentially sophisticated ways. Finally, a complex system is one that emerges and evolves over time due to its many intricacies and dynamism. It is hard to envision a video game that fulfills the four essential traits (i.e., a goal, rules, a feedback system, and voluntary participation; McGonigal, 2011) but fails to exhibit complexity, dynamism, and emergence.

Efforts to improve methods within the field of video game research occur with some regularity (e.g., Baek, 2017; Ferdig, 2008; Ge & Ifenthaler, 2017; Ifenthaler, Eseryel, & Ge, 2012; Loh, Sheng, & Ifenthaler, 2015). These advances have served to improve the science, as well as expand opportunities for research into games as designed experiences and as assessments (DiCerbo & Behrens, 2012; Schrader, Deniz, & Keilty, 2016; Shute, Ke, & Wang, 2017; Ventura & Shute, 2013). Although it is vital to continue to examine and expand methods, some researchers concomitantly advocate for more intentional and meaningful operational definitions of games (O'Brien et al., 2010; Schrader & McCreery, 2012; Young et al., 2012). Specifically, this entails a careful account of the characteristics of the system from multiple perspectives (i.e., the human agents' and the designers' perspectives). In simple terms, the affordances that are available to the human agent (e.g., mouse-based movement or social emotes) and those that are designed into the system (e.g., video recording or XML interface design) have a direct influence on the approach to research. In the strictest sense, each affordance may translate into a variable for study (e.g., navigation or social interaction) or a means to extract data from the system (e.g., video logs or XML data dumps).

As a result, a complex system view of video games pushes researchers to account for process and emergent trends within video game systems. More traditional methods excel at comparisons and descriptions, but do not account for these elements. More importantly, research into games when viewed as a complex system invites new questions that are otherwise obvious to researchers. Given the continued evolution of games and the lingering need to reprise methods for these systems, this chapter is focused on describing a heuristic for classification and research informed by three key perspectives: (1) games are complex systems (Hipert & Marchand, 2018; Shalizi, 2006); (2) human-computer interaction is a viable framework to describe learning *with* and *in* these systems (Carroll & Campbell, 1999; Schrader et al., 2017); and (3) *process*-oriented data extracted from games can be informed from an analytics perspective (Bernacki & Walkington, 2018; Schrader et al., 2017).

This chapter is organized with an overall, conceptual review of the relevant literature. Next, six principles are presented that may help researchers engage in studies that involve the process of learning (see Table 2.1 for a brief overview of these principles). For each principle, questions to consider have been provided. These emerged from several years of research in the area and are intended to highlight some of the key elements, challenges, or pitfalls of each principle to think about when deciding on a system, research questions, framework, etc. The next few sections are intended to highlight the overall process through existing and ongoing work in three different game contexts (i.e., *Bully*, *The Deed*, and *League of Legends*). Each section includes (a) detailed descriptions of the games, (b) discussions of each relevant class of affordances (i.e., player, researcher, and developer), and (c) practical research examples, including the purpose, method, and strategies to analyze data. The chapter concludes with a discussion of how the heuristic and its principles guided each of the research examples to contextualize broader implications for learning, assessment, research, and design.

Table 2.1 Guiding principles for game-based research as contexts for process-oriented learning

Principles	Strategies	Considerations
Principle 1: framework	Identify the relevant research lens or framework, including appropriate variables, data extraction techniques, and questions	For this discussion, the purpose of a theoretical framework includes a perspective that empowers researchers to examine learning as a process
Principle 2: system	Examine the system for its characteristics and affordances, attending to the potential for interaction and data collection opportunities	Consider affordances in at least classes, the human agent or user and the developer/designer. Each has special significance and importance for the researcher
Principle 3: agency	Decode the theoretical framework as it pertains to variables associated with the human agent	Consider the states and traits and ways both may have an influence interaction within the system
Principle 4: methods	Aligned with the framework, develop a method, or methods, to extract data from the system	Consider the potential for the system to deliver data that address questions associated with process and interactions, including qualitative, quantitative, and mixed approaches
Principle 5: analyses	Examine data for patterns	Consider the advantage of analytic techniques that evaluate patterns over time, particularly associated with data that are dynamic and emergent
Principle 6: inferences	Draw inferences, establish models, and interpret findings	For this discussion, the heuristic is focused on describing ongoing processes and interactions, considering the human agency and the complexities of a system. Inferences should align with this perspective

2.2 Principle 1: Establish the Frame

Most researchers are trained to build programs of study and systems of understanding using internally consistent assumptions, theories, and findings. However, there is an ongoing need to examine the epistemological and ontological underpinnings of research. At a minimum, a clear and cogent theoretical lens provides the greatest opportunity for scientific communication, particularly for the purpose of exchanging those inferences as findings are interpreted for relevance by other scientists. More broadly, the theoretical framework serves every aspect of research, including developing and refining the question, identifying variables, deciding the methods appropriate for measuring the variables or extracting data from the environment, the ways in which patterns are detected, and the inferences drawn through the observation of those patterns. As such, a clear and cogent lens undergirds everything about a study, from its boundary conditions to the potential for integrating new findings into extant lines of work.

With respect to video games, there are ample perspectives, ideologies, and paradigms to theoretically and pragmatically frame research investigations. Some approaches typify these systems as interventions and tend to emphasize classical pre-post designs, variables linked to change or growth, and outcome-oriented analyses (Schrader et al., 2017). For example, Schenk, Lech, and Suchan (2017) examined the outcomes of video gameplay on probabilistic learning, a frame that leveraged a video game as an experience and context for an intervention-oriented study. This pre-post study gathered magnetic resonance imaging data, as well as post-experimental questionnaire data. Data analyses involved comparison approaches (i.e., ANCOVA) and suggested an important role of declarative knowledge and hippocampus involvement related to probabilistic learning. Alternatively, other approaches trend toward exploration, development, or optimization format that relies on some observable change, whether that is measured objectively through the observation of variables or some other means (e.g., design-based research or design-based learning; Design Based Research Collective, 2003; Qian & Clark, 2016). In these cases, the focus of the research is to establish a set of best practices or optimized set of tools that are informed by learning goals and objectives (i.e., an improved game). For example, Ke (2009) used the process of video game authoring to enhance mathematical thinking in a design-based learning approach. Compared to the frame of using a game as an intervention, Ke employed a game-design experience to contextualize computer coding and mathematics content in an engaging activity.

The wide range and applicability of designs, methods, and theoretical frameworks serves to strengthen research associated with games. There are boundless questions and orientations and a commensurate array of approaches to address questions from those perspectives. However, it has already been noted that most approaches focus on outcomes or characteristics. Few approaches are equipped to delve deeply into the process of learning from, with, or within these interactive and complex systems (Jonassen et al., 1994; Salomon et al., 1991; Schrader, 2008;

Schrader et al., 2017). Games provide numerous opportunities to assess performance and study learning as a process (McCreery, Krach, Schrader, & Boone, 2012; McCreery, Schrader, Krach, & Boone, 2013; McCreery, Vallett, & Clark, 2015; Tettegah, McCreery, & Blumberg, 2015). Because this perspective contrasts sharply when compared to other, more typical, approaches, the importance of establishing a theoretical frame, particularly one that aligns with a notion that games are emergent, dynamic, and complex systems, cannot be overstated. It is assumed that the researcher has a few specific questions or hypotheses in mind, but there are a few additional questions to consider when establishing the framework.

- Given the theoretical lens or framework, what implications are there for the types of questions that this perspective is equipped to explain (e.g., change and statistical null-hypotheses)? Consider time-intensive questions, rather than change dependent questions.
- What are the appropriate and/or unique implications for the variables in the study? How does the framework influence the definition and operationalization of the pertinent variables?
- Are there inconsistent or incompatible perspectives, epistemologies, or ontologies relative to the theoretical framework that contextualizes the research?

2.3 Principle 2: Identify Attributes of System

In traditional research, it is necessary to identify the methods for inquiry and the variables under investigation. With respect to evolving research with games, it is similarly crucial to identify the characteristics of the system involved in a study. Further, researchers should examine agency from three unique perspectives: the players, the developers, and the researchers. It is useful to remember that the affordances experienced by a player and the affordances designed into a system by a developer do not necessarily overlap. Collectively, these two sets of affordances also influence the capability for research. Said another way, what is relevant and important from a player perspective may not be what was intentionally designed in a system and neither set of affordances may be all that relevant to a researcher.

By the mid-twentieth century, psychologists expanded the notion of perception, action, and the importance of acknowledging the mutuality in the seemingly disparate roles of agent and environment in perceptually rich systems (Gibson, 1977, 1986; Greeno, 1994; Mace, 1977). Throughout this work, Gibson (1977, 1986) established a notion of affordance, which pertains to characteristics of an environment to provide opportunities for action. Although Gibson described natural environments, the concept of affordance has been applied to various constructed and designed environments (Gaver, 1991). When applied to technological contexts, an affordance holds meaning from at least two perspectives: (a) the human agent involved in acting and perceiving within the system and (b) the developers and designers who created the system. These two classes of affordances combine in unique ways that have special relevance for researchers.

At a minimum, understanding these two types of affordances provides cues about which types of variables are measurable. For example, 3D massively multiplayer games involve interactions among human agents. This implies a variety of interpersonal interactions, as well as spatial relationships. As a result, researchers might capture chat data, spatial navigation data, or some record of in-game behaviors (McCreery et al., 2012, 2015). More broadly, this understanding hints at strategies to exploit the technology for data capture. This could include system logs, video recording, or some form of biometric data capture (McCreery et al., 2013: Schrader & Lawless, 2007; Schrader et al., 2017). In either case, a crucial to investigating and assessing learning within games is to deeply and meaningfully understand these affordances so the implications for research are apparent. Some useful questions are outlined below:

- After carefully examining the game, which affordances are important to the research?
- Which affordances, if any, can be leveraged in ways that facilitate defining variables or extracting data relative to those variables?
- How can the design characteristics facilitate data collection and the methods to examine the question?
- In what ways does the system exhibit emergence and dynamism?

2.4 Principle 3: Consider the Human Agents

The potential for environments to afford action is one of the core assumptions of most branches of psychology. In the literature, there are numerous and well-established constructs that have been linked to learning; variables associated with human performance are many and varied (e.g., self-efficacy, situated interest, cognitive load, affect, prior knowledge and experience, presence). As a result, it is judicious to incorporate pertinent constructs when studying learning within systems like games. However, some theories may not be equipped to reconcile the influence of user variables when compared to variables inherent to the context. Alternatively, the field of human-computer interaction (HCI) involves the study of human motivation, action, and experience as it pertains to the agents' interactions with technology (Carroll & Campbell, 1999). From this perspective, learning and behavior are studied in direct relation to the capabilities of users in conjunction with the elements of design (i.e., hardware, software, content, and context).

One of the primary foci of HCI research is to generate evidence that informs design, particularly of the hardware and software involved in these environments. This includes the designed, digital elements that users experience, as well as the physical interfaces and controls players use to express their intent. Like most research, HCI is performed through rigorous examination of outcomes and performance. Unlike most research, HCI adopts a perspective that users' experiences are those that are informed by the mutuality between individual characteristics (i.e., states and traits) and the relevant attributes of the context (i.e., user affordances and designed affordances).

From this point of view, constructs that account for the human agent's performance should be part of the overall research design. The literature has examined numerous variables in relation to human performance, including prior knowledge, self-efficacy, expertise, cognitive load, affect, personality, and situational interest (Alexander, 1992, 2003; Alexander & Dochy, 1995; Bandura, 1997; McCreery, Krach, & Nolen, 2014; Sweller, 1988). Additionally, successful gameplay also relates to numerous physical components (e.g., sequential and repetitive key presses and controller movements). As a result, there are numerous biomedical characteristics to consider, including stress, galvanic skin response, heart rate, and reaction time (Mirza-Babaei, Long, Foley, & McAllister, 2011; Mirza-Babaei, Nacke, Gregory, Collins, & Fitzpatrick, 2013).

Collectively, these variables contribute to the successes and failures to execute users' intentions within a game system. Although some systems may not be optimized for maximum player performance (i.e., the affordances are limited, difficult to detect, or not aligned with players' goals), users' characteristics also have a direct influence on performance. Ultimately, researchers are advised to consider the following questions when examining the users' characteristics as they pertain to the questions under investigation and the context being studied:

- Which, if any, individual differences have a higher than average likelihood to influence the process and outcomes?
- Which factors associated with the individuals (i.e., states and/or traits) are pertinent to the question and the model under investigation?
- In what ways do the user's characteristics and experiences interact with the system's affordances?
- How do these interactions relate to, and have implications for, the questions, hypotheses, etc.?

2.5 Principle 4: Identify Methods to Capture Data

Collectively, the attributes of the system and user (i.e., their states and traits) combine into a research context that is complex. Although a few theories directly address complexity associated with learning and training in technological contexts (e.g., van Merrienboer & Kirschner, 2018; van Merrienboer & Sluijsmans, 2009), a *complex systems* view of learning and technology expands these perspectives considerably (Hilpert & Marchand, 2018; Marchand & Hilpert, 2017, 2018). This is particularly true in terms of the implications associated with methods to capture data and address questions associated with a complex systems approach to games and assessment.

Complex systems are collections of elements, characteristics, and components that interact in ways that generate intricate and interrelated behavioral patterns (Hilpert & Marchand, 2018). Fundamentally, these behavioral patterns exhibit three key attributes: complexity, dynamism, and emergence. Essentially, interactions in complex systems are influenced by multiple integrating components (complexity), are continually shaped by those interactions (dynamism), and evolve over the duration of the experience (emergence).

When a complex systems approach is applied to digital games research and assessment, the implications for characterizing the environment become straightforward. Players interact with highly sophisticated systems, in which the users' states and traits in conjunction with the affordances of the system all have an impact on behavior (complexity). More importantly, that play is tuned by soft-failure, trial and error, feedback systems, rules, and sometimes other players (dynamism) (Laughlin & Marchuk, 2005; McGonigal, 2011; Squire, 2006; Vallett, 2016). Lastly, the dynamics of players' behavior occurs throughout the gameplay experience (emergence).

Although the complex systems perspective has clear implications for defining and characterizing video games, the implications for methods to extract data from these environments are less obvious. Video game research is often difficult due to the tremendous breath of available games, each type of which is identified by different mechanics, interfaces, formats, etc. and a lack of empirical research (Young et al., 2012). As such, identifying variables and best research practices is a serious challenge. Alternatively, a complex systems approach from the lens of human-computer interaction shapes methods in two specific ways: (1) data must address complexity, dynamism, and emergence, and (2) research designs must account for the attributes of the system (i.e., its affordances) and the user (i.e., states and traits).

Considering this, there are a few questions to consider as designs are constructed.

- Is the system capable of directly generating objective data (e.g., log data, database extraction)?
- Does gameplay have a clear set of initial conditions (e.g., equivalent maps, starting positions, levels), or does gameplay exist in a more fluid state (e.g., persistent worlds like World of Warcraft)?
- Based on the understanding of the system, what opportunities to constrain experiences or manipulate variables exist?
- What existing technological tools are available to facilitate data extraction from the system? Does data extraction and/or coding rely more heavily on research labor?
- Is it possible to sequence data collection strategies (e.g., time series) to account for emergence and process-oriented perspectives?

2.6 Principle 5: Identify Patterns Among the Data

The fundamental purpose of any analysis is to elucidate patterns among the data, regardless of the type or nature of the data being examined. Based on a frame that espouses learning as a process, pattern detection techniques must have some capacity to account for the dynamic and complex nature of the data as they occur over periods of time. For many methods, this type of change is difficult to identify and characterize. When applied to emergent data, most analytic techniques, particularly quantitative techniques, rely heavily on comparisons between two points in time (e.g., t-test, ANOVA, MANCOVA). For complex systems, this low-dimensional, data-independent approach is not enough to measure or explain patterns in a context that contains many parts, whose behaviors vary significantly and are dependent on the other elements in that system (Shalizi, 2006). Even repeated measures techniques, which incorporate various algebraic trends across multiple points in time (e.g., linear, loglinear, parabolic), do so in ways that examine differences rather than emergent trends. Qualitative analyses are somewhat more adept at deconstructing changes over time, but they may not necessarily align with the nature of data that are extracted from complex systems.

Fortunately, there are a few data analytics approaches that researchers have developed for data from complex systems (Shalizi, 2006). These techniques range from qualitative comparisons using heat maps to machine-learning/artificial intelligence-oriented logistic regression analyses. Each approach has a distinct benefit and entails different methods and criteria.

Although this chapter is not intended to provide a primer on the techniques that are useful in process-oriented data analysis, some approaches that are relevant for video games research are listed in Table 2.2. Regardless of the technique selected, researchers are encouraged to consider the potential output of these techniques and how the findings address the original question within the context of pragmatism.

2.7 Principle 6: Draw Inferences, Establish Models, and Interpret Findings

Generally, inferences associated with low-dimensional, independent data systems (e.g., t-test, ANOVA, MANOVA) are straightforward; significance testing indicates whether or not one should reject a null hypothesis. Alternative approaches have similar, well-established inferential potential. With respect to qualitative methods, inductive techniques are codified to yield findings that address pertinent questions. Generally, researchers seek data and patterns that provide some evidence of a magnitude or quality of change. This is very different for data that involve even minimal degrees of emergence. Rather than show change between two points, these methods evaluate inversion points, points of change, or opportunities to describe shifts in patterns.

Table 2.2 Analytic techniques for pattern detection in game-based research

Technique	Description	Research application	Challenges
Think aloud	Verbal reporting of experiences during the process of play	Expose and externalize decision-making and thinking processes as the experience unfolds	Unprompted and potentially unnatural. May have negative impact on gameplay
Heat map	Concatenate emergence into a single, visual representation	Provide some indication of behavior over time, represented as a visual map. Qualitative contrasts of maps by type, group, etc. (e.g., experts vs. novices) are possible	Subjective inference. Limited to pre-defined variables
Path analysis	Seeking latent structure over time, path analysis provides a model of magnitude and significance for the hypothesized causal connections among nodes	Delineating events in games as nodes and contrasting classes of paths (e.g., successful outcomes vs. unsuccessful ones) may yield insight into	Paths are decoupled from time, limiting inferences associated with emergence
Neural network	Data analytics method that captures input/output relationships to estimate future values of those inputs and outputs	Neural networks can be used to detect patterns and make outcome predictions using time series data, like those available in games	Cumbersome in terms of variable and algorithm definition. Best with original, source data, which are not typically available in proprietary games
Bayesian/ Markov network	Methods to detect probabilistic relationships and statistical dependencies among "events"	A network generated by an expert can be used as another form of performance inference. Bayesian networks can be used to detect latent variables and structure. Markov networks can be used to detect cyclic dependencies	Reliant on prior data to build initial probability models, which have questionable quality or value. Can become unwieldy with systems that include large numbers of variables
Logistic regression	Minimally, logistic regression is a technique to explain the relationship between one dependent binary variable and one or more independent variables	Automated methods of logistic regression can iterate the process and incorporate massive fields of data to build detailed models of time-based performance. Those regression models can then be applied to real-time behavior to predict performance	Applied in this way, logistic regression is laborious or heavily dependent on a learning AI. The just-in-time overcorrection analyses are useful to determine likelihood of outcomes, but less adept at addressing research questions

Some techniques that have been applied to complex systems research involve the process of decoding a central measure of information within that system (Hilpert & Marchand, 2018; Marchand & Hilpert, 2017, 2018; Shalizi, 2006). For example, a classroom exhibits the characteristics of a complex, dynamic system. Because complexity is characterized by multiple, interacting parts, each of which has its own history and sub-set of influences, it would be difficult to ascertain the beginning point in time that a classroom began, particularly as a research context (Marchand & Hilpert, 2017). Alternatively, it may be possible to discern a trend function or general model of behaviors at the time. This measure, albeit incredibly dense in terms of information, provides a point in time from when patterns emerge. In this way, this initial measure serves as a microgenetic function that highlights how the system unfolds and provides extensible model of patterns and trends.

Hilpert and Marchand (2018) described complex systems as self-organizing systems. As such, systems like these tend to progress toward and exhibit stability over time. As a result, identifying a trend function is useful in terms of giving a broad characteristic associated with the system, hence the term *microgenetic function*. The trend function also serves to shape many inferences about complex systems. Specifically, researchers endeavor to identify points of inversion, phase shift, or instability relative to the trend function. Evidence of instability suggests something noteworthy occurred. Accordingly, the practice of seeking points where the system is unstable provides researchers with opportunities to document crucial events (e.g., learning, new approaches to behavior, coachable moments). Those events provide the specific opportunities for inference.

In some games, a centralized measure or trend function may not be necessary. Many games provide an opportunity to establish a limited set of initial conditions (e.g., starting level, initial resources, map selection) and impose researchable constraints on the investigations (e.g., time limit, avatar selection, role, class selection). This simplifies the development of a model but the approach to inference building is still relevant. Specifically, game researchers who are interested in process-oriented data and time-intensive questions should also seek phase inversions and shifts. In games, a change of phase could be due to the introduction of error, exploitation of an opponent's mistake, or the user becoming more attuned to the system. This might be a point where a victorious strategy was employed or when one player exploited the mistake of another.

Whatever the case and whichever analytic technique is employed, inferences and models in games research in which learning is process-oriented are approached somewhat differently than traditional methods. Fortunately, there are a few general ways of thinking that can be employed to facilitate inference generation, model building, and interpretation of findings. Some examples are:

- What do the patterns or trends over time indicate, imply, or suggest?
- Are there expert or formative trends against which the observed trends may be compared?

- Are there instances when the observed behavior departs from the anticipated trend? What are the circumstances of those departures and what do they indicate?
- Are there differences in the system or human agent that might account for changes to the models?

2.8 Heuristic Applied

One way to appreciate these principles is to consider their application to different research questions, variables, and contexts. As a result, we examine three different contexts and provide an overview of three ongoing research projects to exemplify a games-based assessment heuristic. The games include *Bully*, *The Deed*, and *League of Legends*. In each case, the framework has been established as one that involves learning as a process (i.e., Principle 1), and the hypotheses were generated prior to selecting a system for research. In most cases, selecting the system involves a cycle of experience with the game system to ascertain fit and alignment with the research questions. The following sections are organized in a way that highlights characteristics of the systems as examples of the six principles noted above.

2.8.1 Bully (2006)

2.8.1.1 Principle 2: Bully as a System for Research

Upon initial screening, *Bully* was identified as a game that involved sensitive topics about bullying. Specifically, *Bully* is a commercially available game, developed by Rockstar Games (2006). The players are expected to go to class, to participate in recreation, and to fight. *Bully* contains themes of social exclusion, body shaming, gendered stereotypes, economic divisions, and discrimination and power dynamics among children, which makes it an ideal candidate for students who are learning about bullying or school aggression.

The setting of the game *Bully* takes place in a fictional rural New England town and at Bullworth Academy. The story follows the main character Jimmy (the single player-controlled protagonist) who quickly discovers that Bullworth Academy is full of bullies and sets out to bring peace to the school. The game allows players to explore the school and the surrounding town as they work through story missions in a somewhat linear way. These experiences are designed and intended to help victims of bullying.

The story is divided into six chapters. Each chapter has a new set of bullies that the main character must overcome while helping a fellow student who is the victim of bullying. Within each chapter, there are a handful of missions that progress the story (e.g., gathering objects, helping other students/teachers) and to ultimately overcome the antagonists in that chapter (e.g., jock, greasers, etc.). Ultimately, *Bully* demonstrated an appropriate plot and role-taking/perspective-taking design that aligned with promoting social awareness and pre-service teacher training associated with bullying.

2.8.1.2 Principle 3: Agency of Bully

A deeper examination of *Bully* revealed that the affordances highlighted in Principle 3 allowed valuable experiences for the purpose of an intervention, as well as appropriate, though not ideal, methods to capture data.

Player Affordances The game contains numerous visual affordances that allow players to seek information, including an on-screen mini-map with way-point indicators, a task (mission) list to remind them what they are working on/toward, and prompts to select specific controls to interact with objects (e.g., trash can, locker, soda machine, or people). The game includes other visual prompts (e.g., visual cues yellow floor marking, or text reminders) for missions and attending class (i.e., clock warning). The game also includes a typical set of player inputs and controls, including pause and save functions.

Developer Affordances In addition to the functions that are available to the player, the developers have included numerous functions that are not strictly necessary for gameplay. Specifically, developers included early game walkthroughs of controls, fights, breaking in lockers, buying soda "health," building weapon inventory (cherry bomb), and gaining charisma effected by attending classes. Collectively, the developers collected plot elements to create "realistic" situations that could and often do happen. The graphics are sufficiently detailed to allow for identification of various, although sometimes exaggerated, stereotyped bullies and victims.

Researcher Affordances Collectively, the player and developer affordances allow researchers to observe the consequences decision-making through gameplay capture. For example, a cursory analysis would indicate whether or not the player exhibits reactive or proactive aggressive to other students (i.e., response to acts of aggression or humiliates students to solve some missions). Because the game actions are recordable, researchers can tally interactions with various NPCs, while the player explores the world, interacts with objects, and selects weapons. Researchers can also examine trait differences among players (e.g., gender, race).

2.8.1.3 Principles 4, 5, and 6: Practical Research Example

Purpose In an ongoing study involving pre-service teachers, *Bully* was examined in relation to promoting social awareness and interpersonal understanding, particularly those that result from unbalanced power dynamics in schools. This example followed a typical intervention design and participants engaged with *Bully* in two sessions, with opportunities to reflect on their experiences after each session. One notable addition to conventional intervention research is that process-oriented data were also collected for the purpose of indicating which events, actions, and interactions among user states and traits were a meaningful component of the experiences.

Method To evaluate the value of the intervention holistically, pre-service teachers (PSTs) were randomly assigned to one of two groups (treatment and control). Typical demographic variables, self-efficacy associated with detecting bullying in school settings, and an aggression inventory were collected at this time. All participants received pre-existing professional development with bullying and activities to train PSTs to detect bullying in schools. In addition, participants in the treatment group were directed to play the game during two different sessions. PSTs assigned to the control group did not engage in gameplay but received comparable training. All gameplay for the treatment group was recorded.

To evaluate process-oriented questions, there was a reflection component after each of the two sessions. Participants were shown portions of their gameplay and asked to respond to specific prompts about bullying and the events that they witnessed and to reflect about the authenticity of the events that transpired during the game. They were also asked to deconstruct their thinking during those instances. These responses were incorporated into a codebook that included researcher observations of gameplay-associated observable behaviors categorized as reactive or proactive aggression.

At the end of the second play-through, PSTs were also given analogous transfer task, in which PSTs completed a bullying observation/intervention sheet that documents what teachers should be doing in the situation that took place during the game.

Analysis Comparative analyses focused on the potential for *Bully* to provide authentic, supplemental experiences for training PSTs in detecting bullying in school settings. Specifically, self-efficacy measures were contrasted between treatment and control. Process-oriented analyses were intended to identify moments of value, events that provided teachable moments, and provide some indication of aggressive tendencies associated with in-game behavior. Specifically, planned analyses include a path analysis to highlight the relationship among aggressive tendencies in the game (i.e., proactive and reactive aggression behaviors), the aggression inventory, and performance on the analogous task. PST responses are intended to provide additional insight into decisions and relative importance of gameplay events.

2.8.2 *The Deed (2015)*

2.8.2.1 Principle 2: The Deed as a System for Research

The Deed is a commercially available game developed by Pilgrim Adventures and GrabtheGames Studios (2015). *The Deed* incorporates a choose-your-own adventure style of play with a macabre narrative-driven murder mystery game. A single player game, *The Deed* challenges the player to make a series of choices through non-player character interactions and collects and plants evidence in a way that supports the main character's (i.e., player's) innocence and frames a non-player character as guilty.

Using a keyboard or mouse, the player navigates through their avatar's childhood home, can speak to several family members and staff, search through different rooms, and is given the choice of picking up objects to aid them in committing the perfect murder. The game is partitioned into four acts: (1) exploration of the house (prefaced by an introduction), (2) dinner, (3) a time to plant evidence and commit the titular deed, and (4) an interview with the Inspector who comes to investigate the murder.

2.8.2.2 Principle 3: Agency of The Deed

Player Affordances By comparison to other systems, *The Deed* is somewhat simplistic in its design. Players use a keyboard and mouse to interact with the game. Specifically, players move and navigate the game, inspect and obtain objects throughout the house (e.g., weapons/evidence), and place them in their inventory. Similarly, players can use the mouse and keyboard to plant evidence if they choose to do so. Most importantly, player interactions with narrative elements greatly influence the final success or failure in the game.

Developer Affordances The Deed involves a robust collection of affordances that enhance the opportunities for action but are not necessarily required for gameplay. For example, developers added an option to watch or skip the introduction, to customize settings, and to save/reset/exit the game. Additionally, the nature of NPC interactions is constrained by the developers in a way that shapes the game but do not always meaningfully change the player's ability solve the game. For example, a player can only speak to each NPC once; their dialogue choices affect the character's reaction to the player and events later in the game. Developers also added distractors to the environment; objects generally contain irrelevant information. However, in some instances, objects reveal historical plot elements and some contain concealed evidence or weapons. Developers also shaped play by limiting the number of weapons and pieces of evidence; only two items can be picked up and they cannot be exchanged (note: a warning message is displayed before the action is completed).

Each act progresses in a pre-specified way, with key instances and moments determining the progress in the game and advancement to subsequent acts. Depending on the amount of suspicion the player has successfully diverted, one of three endings will occur: a prison sentence for murder, getting away with the murder but not receiving the family inheritance, or getting away with murder and receiving the family inheritance. Because there are many possible paths that lead to these outcomes, *The Deed* is a game that affords replayability and multiple endings.

Researcher Affordances Both the player and developer affordances combine in ways that benefit research. For example, *The Deed* is a controlled environment with a finite number of choices and actions to take. As a result, the number of possible confounds is greatly reduced. More importantly, it is possible for researchers to fully map and catalog all the possible behaviors in the game. Data could be extracted from back end sources or direct observation (i.e., video recording). Many of the behaviors are binary in nature (i.e., they either did or did not happen), allowing coding techniques and intercoder agreement to be extremely accurate. In addition to a finite set of actions, *The Deed* also offers the same experience for all subjects. This has the advantage of allowing researchers the choice to study multiple instances of play while studying or varying other aspects of the experience.

2.8.2.3 Principles 4, 5, and 6: Practical Research Example

Purpose Like *Bully*, *The Deed* offers users the opportunity to behave in ways that exhibit proactive and reactive aggression. Unlike the expansive storyline of *Bully*, *The Deed* is programmed to include a finite number of actions. Researchers can catalogue the entire play-space of *The Deed* and code every single player decision in an objective manner. Because it is self-contained and affords aggressive actions, *The Deed* is currently being examined as a performance assessment of aggression.

Method Following a traditional single-group design, participants are asked to complete a battery of assessments and correlates of aggression prior to engaging with the game. Players are then given a brief tutorial on playing the game and asked to play to conclusion two times. All gameplay is recorded using screen-capture software.

Analysis Analysis begins with the development of a coding catalogue, which includes all possible behaviors and choices in the game. This catalogue is used to code player's decisions and actions. In this case, response choices during dialogue interactions are assigned a variety of markers associated with aggression. These markers are observed over time and theorized to function as indicators of the participant's aggressive tendencies.

2.8.3 League of Legends (2009)

2.8.3.1 Principle 2: LOL as a System for Research

League of Legends (LOL; Riot Games, 2009) is a commercially available competitive online game. Its most common mode involves two teams of five players attempting to control a small map and defeat the other team's defensive structures and central base. In LOL, players control an avatar, known in the game as a champion. At the time of this writing, there are more than 140 unique champions from which players are free to choose. Each champion is defined by unique set of abilities that interact with other players, elements of the game environment, and NPCs in distinct ways. Champions' abilities are designed to fit within at least one roles: ranged damage dealing champions, support champions (e.g., healing or damage mitigation), champions that can function independently, and champions that are mobile and can create opportunities for advantage. Most game modes prohibit the same champion from being selected by more than one player.

Although each game evolves differently, each game is a predictable experience. Players are free to choose roles, but typically do so in somewhat predictable ways (e.g., support or offense). There are three primary routes (i.e., lanes) to the enemy base, a top lane, middle lane, and bottom lane. Additionally, there are other areas of the map that are patrolled by champions in a jungle area. The map remains the same and all players start with the same resources, champion experience, and defensive structures.

In many ways, LOL is analogous to chess in the sense that each team shares the exact opportunities for success and that stages of play (i.e., early game, mid game, end game) emerge over time. Further, success is largely determined by players' skill, knowledge of the game, and ability to exploit opponent's errors. LOL also exists within a constrained space, defined by limited paths of movement, number of players, and champion abilities. Unlike chess, success in LOL is also reliant on dynamic and complex elements, like the actions of teammates and interactions among champion abilities. The fundamental mechanics of the game are relatively simple, but the interactions among parts are challenging for developers to balance and for players to master.

2.8.3.2 Principle 3: Agency of LOL

Player Affordances In LOL, players select from more than 140 unique champions and adopt 1 of 5 distinct positions based on zones of the map (i.e., top, middle, bottom attack, bottom support, or jungle). Actions within the game are heavily dependent upon repetitive clicks, keystrokes, and mouse movement. The game system provides users with access to a minimal, up-to-minute performance of everyone in the match and instantaneous feedback relative to actions. LOL is characterized by a constant march of weak NPCs (i.e., minions). When a player's champion executes the final strike on a minion, the champion receives a currency that is used to upgrade items and increase the champion's power. Champions also acquire experience and levels in the game, gaining power in more predictable ways. Another significant element in the game is the ability for players to discern other players action using invisible wards. Wards are a vital aspect of the game and provide visual information about the enemy's location, capabilities, strategies, etc. This information is otherwise unavailable due to a "fog of war" (i.e., intentionally obscured map information).

Developer Affordances LOL developers have incorporated numerous additional affordances into the game that are not strictly relevant for play. For example, the business model of LOL involves microtransactions. Players may pay to change the cosmetic elements of their champions or to gain access to additional champions. Additionally, developers included mechanisms to record gameplay, review games, and receive information about matches (online). In some types of play, developers included options to compete and keep a ranking of success. This system is highly detailed and serves to match players with opponents of comparable ability.

Researcher Affordances Collectively, player and developer affordances in LOL combine to provide some notable options for researchers. The ranking system is a verified and validated method to differentiate players based on skill and performance. As such, it is a viable means to infer expertise for purposes of comparisons, model building, and problem-solving. Specifically, experts' games are recorded and publicly available. It is feasible to review these matches and develop expert models and heuristics of play. Additionally, the system involves tremendous physical interaction at a rapid pace; most professional players can execute as many as 10 actions (Lejacq, 2013). As a result, researchers could examine several variables to understand human-computer interaction (e.g., user activity, visual spatial acclimation, design usability, input and output systems, feedback, information systems).

2.8.3.3 Principles 4, 5, and 6: Practical Research Example

Purpose LOL is distinct from the other contexts described above in terms of its level of complexity. In LOL, there are ten different players that control unique champions, each of which exhibits a different configuration of abilities. Each player also has unique skills, knowledge, states, and traits. Although the game space is predictable, the interactions of these constituent parts exceeds the level of complexity seen in *Bully* or *The Deed* by a wide margin. As a result, a multi-phase study is being conducted to (1) examine the play space from the perspective of experts, (2) determine the behaviors that are most closely aligned with success (i.e., winning matches) and develop an instrument to evaluate those behaviors in game, and (3) leverage prediction techniques to examine phase changes within the LOL competitive and amateur play.

Method For the first two phases, a mixed-method approach was implemented to extract data from experts via an online forum system (i.e., Reddit). Players were asked multiple rounds of questions that pertained to events, actions, and behaviors in game that result in winning. These responses were qualitatively analyzed for themes. Those themes served as the stems in a survey instrument (phase 2). This survey was distributed widely, validated, and psychometrically evaluated. More importantly, the items correspond to behaviors that are observable in game. In phase 3, research plans involve using the survey instrument to develop a behavioral assessment matrix. That matrix will be used to examine players' activities in game. Following a single-group comparison, expert and novice players will be asked to complete several games.

Analysis Analysis of phases 1 and 2 included qualitative content analysis and psychometric item validation techniques (i.e., exploratory and confirmatory factor analysis, reliability analysis, expert panels via distance). Planned analysis of phase 3 includes path analytic techniques to contrast group trends. Additionally, the data will be examined for their suitability and viability in terms of Bayesian/Markov techniques to examine phase shifts and inversions.

2.9 Conclusion

The studies discussed here vary in the degree to which each emphasized process-oriented data and their value when expanding researchers' ability to examine questions in context. Further, this work is intended to exemplify the potential benefits of an HCI and complex systems perspective for all studies, including those that implement traditional pre-post designs. Initially, the approach described by Schrader and McCreery (2012) (i.e., looking at learning *from*, *with*, and *in* games) was an attempt to consider interactions within systems and the potential to examine the nuances of learning as a function of those dynamics (Schrader, 2008; Schrader & McCreery, 2012). More recently, there has been increased interest in learning as a process, one that is mutually influenced by the nature of the system and the characteristics of the user (i.e., from an HCI perspective; Schrader et al., 2017). An HCI approach is focused more intently when games are viewed as complex systems. Additionally, this perspective expands the relevance of agency in terms of data collection and user experience.

Overall, the heuristic described here is applicable to examine a variety of questions that examine learning *from*, *with*, and *in* games. Specifically, the ability to determine the influence of games on learning (e.g., as with an intervention involving *Bully*) is greatly improved when accounting for process-oriented data. In this example, knowing which elements from the experience generate the most salient emotional connection and response is a crucial part of understanding the game's impact. Alternatively, learning *with* a system is slightly different. Still, it is possible to leverage the process-oriented data to expound on the experiences. In *The Deed*, process data revealed a variety of interactive elements, particularly those related to aggression. In this case, users' interaction *with* the system has been shown to serve as an analog to standard paper and pencil assessments. Lastly, research associated with learning *in* a system is the clearest example of the heuristic in practice. Games like *LOL* are highly complex and dynamic; their conditions constantly update as a result of countless actions and interactions within the game. In this example, process-oriented data are the only way to observe the transition or inflection points that relate to performance. Collectively, the opportunities for research and the variety of classes of research questions are vast.

Although there are important and valuable benefits to this approach, there are also some limitations to and situations into which the process does not translate well. For example, this approach assumes that researchers have accepted a view that learning is a process. This approach is not ideally suited for studies that seek differences among limited points in time. Further, the research questions in these examples had been defined prior to the selection of the system. This allowed for a very intentional screening process that limits the potential for questions to be invited by emerging tools. Sometimes, researchers experience a new design and are inspired to ask different and original questions. Because this process is intentional and begins with a research question and framework, this emergent inquiry is somewhat mitigated. Further, systems are not equally suited for data collection. Although there are a variety of strategies that exist to extract process-oriented data from users' experiences with games, some of these approaches may be incredibly laborious. The economics of research in terms of effort to impact should certainly be considered.

Ultimately, games are emergent systems. Data from game experiences is poised to address more nuanced questions that relate to growth, change, and maturation. Games exhibit a very different set of characteristics and implications for research questions, variables, and methods. A complex systems approach from an HCI lens contrasts sharply from research involving games as interventions. The examples described here highlight the value of examining the system throughout the experience of play. Specifically, data are linked to time-intensive variables and questions, agents in the system are accepted as multifaceted and interactive, and learning is process-oriented. On their own, traditional approaches may not be suitable or robust enough to address complexity, emergence, and dynamism as they pertain to data collection, analyses, and inference generation. Fortunately, expanding the methods of assessing learning from, with, and in games provides at least a few additional opportunities to expand the nature of questions, data collection strategies, and types of inferences available to researchers. Ultimately, improving methods as a field will enhance the collective ability to understand games-based assessment from all perspectives.

References

Alexander, P. A. (1992). Domain knowledge: Evolving themes and emerging concerns. *Educational Psychologist, 27*(1), 33–51.

Alexander, P. A. (2003). The development of expertise: The journey from acclimation to proficiency. *Educational Researcher, 32*(8), 10–14.

Alexander, P. A., & Dochy, F. J. R. C. (1995). Conceptions of knowledge and beliefs: A comparison across varying cultural and educational communities. *American Educational Research Journal, 32*(2), 413–442.

Baek, Y. (2017). *Game-based learning: Theory, strategies and performance outcomes*. Hauppauge, NY: Nova Science Publishers.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: Freeman.

Barab, S. A., Gresalfi, M., & Ingram-Goble, A. (2010). Transformational play: Using games to position person, continent, and context. *Educational Researcher, 39*(7), 525–536. https://doi.org/10.3102/0013189X10386593

Bediou, B., Adams, D. M., Mayer, R. E., Tipton, E., Green, C. S., & Bavelier, D. (2018). Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. *Psychological Bulletin, 144*(1), 77–110.

Bernacki, M. L., & Walkington, C. (2018). The role of situational interest in personalized learning. *Journal of Educational Psychology, 110*(6), 864–881. https://doi.org/10.1037/edu0000250

Carroll, J. M., & Campbell, R. L. (1999). Artifacts as psychological theories: The case of human-computer interaction. *Behaviour & Information Technology, 8*(4), 247–256.

Clark, D., Tanner-Smith, E., & Killingsworth, S. (2014). *Digital games, design and learning: A systematic review and meta-analysis (executive summary)*. Menlo Park, CA: SRI International.

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*(2012), 661–686.

Design Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5–8.

DiCerbo, K. E., & Behrens, J. T. (2012). Implications of the digital ocean on current and future assessment. In R. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 273–306). Charlotte, NC: Information Age Publishing.

Ferdig, R. E. (2008). *Handbook of research on effective electronic gaming in education* (Vol. I–III). Hershey, PA: Information Science Reference.

Gaver, W. H. (1991). Technology affordances. In *Proceedings of CHI'91* (pp. 79–84). New York, NY: ACM.

Ge, X., & Ifenthaler, D. (2017). Designing engaging educational games and assessing engagement in game-based learning. In R. Zheng & M. K. Gardner (Eds.), *Handbook of research on serious games for educational applications* (pp. 255–272). Hershey, PA: IGI Global.

Gibson, J. J. (1977). The theory of affordances. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing: Toward an ecological psychology* (pp. 67–82). Hillsdale, NJ: Lawrence Erlbaum.

Gibson, J. J. (1986). *The ecological approach to visual perception*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Grabthegames Studios. (2015). *The deed*. Grabthegames Studios.

Greeno, J. G. (1994). Gibson's affordances. *The Psychological Review, 101*(2), 336–342.

Hilpert, J. C., & Marchand, G. C. (2018). Complex systems research in educational psychology: Aligning theory and method. *Educational Psychologist, 53*(3), 185–202. https://doi.org/10.1080/00461520.2018.1469411

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

Jonassen, D. H., Campbell, J. P., & Davidson, M. E. (1994). Learning with media: Restructuring the debate. *Educational Technology Research and Development, 42*(2), 31–39.

Ke, F. (2009). A qualitative meta-analysis of computer games as learning tools. In R. Ferdig (Ed.), *Handbook of research on effective electronic gaming in education* (Vol. I, pp. 1–32). Hershey, PA: Information Science Reference.

King, F. B., Young, M. F., Drivere-Richmond, K., & Schrader, P. G. (2001). Defining distance learning and distance education. *AACE Journal, 9*(1), 1–14.

Krach, S. K., & McCreery, M. P. (2015). Technology and positive behavioral interventions and support: Evaluation, selection, and implementation of computer-based socioemotional training. In S. Y. Tettegah & D. Espelage (Eds.), *Emotions and technology, communication of feelings for, with and through digital media: Emotions, technology, and behavior*. Waltham, MA: Elsevier.

Laughlin, D., & Marchuk, N. D. (2005). *A guide to educational computer games for NASA*. Washington, DC: National Aeronautics and Space Administration's Office of the Chief Education Officer, Technology and Products Office.

Lejacq, Y. (2013, October 24). How fast is fast? Some pro gamers make 10 moves per second. *ABC News*. Retrieved from https://www.nbcnews.com/technolog/how-fast-fast-some-pro-gamers-make-10-moves-second-8C11422946

Loh, C. S., Sheng, Y., & Ifenthaler, D. (Eds.). (2015). *Serious games analytics. Methodologies for performance measurement, assessment, and improvement*. New York, NY: Springer.

Mace, W. M. (1977). James J. Gibson's strategy for perceiving: Ask not what's inside your head, but what your head's inside of. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing* (pp. 43–65). Hillsdale, NJ: LEA.

Marchand, G. C., & Hilpert, J. C. (2017). Complex dynamic systems approaches to educational research. *The Journal of Experimental Education, 85*(2), 173–174. https://doi.org/10.1080/00220973.2016.1247555

Marchand, G. C., & Hilpert, J. C. (2018). Design considerations for education scholars interested in complex systems research. *Complicity, 15*(1), 1–14.

Mayer, R. E. (2015). On the need for research evidence to guide the design of computer games for learning. *Educational Psychologist, 50*(4), 349–353. https://doi.org/10.1080/00461520.2015.1133307

McCreery, M. P., Krach, S. K., & Nolen, A. L. (2014). The protagonist and their avatar: Learner characteristics in a culture of simulation. *The International Journal of Gaming and Computer-Mediated Simulations, 6*(2), 30–37.

McCreery, M. P., Krach, S. K., Schrader, P. G., & Boone, R. (2012). Defining the virtual self: Personality, behavior and the psychology of embodiment. *Computers in Human Behavior, 28*(3), 976–983.

McCreery, M. P., Schrader, P. G., Krach, S. K., & Boone, R. (2013). A sense of self: The role of presence in virtual environments. *Computers in Human Behavior, 29*(4), 1635–1640.

McCreery, M. P., Vallett, D., & Clark, C. (2015). Social interaction in a virtual environment: Examining sociospatial interactivity and social presence using behavioral analytics. *Computers in Human Behavior, 51*, 203–206.

McGonigal, J. (2011). *Reality is broken: Why games make us better and how they can change the world*. New York, NY: The Penguin Press.

Mirza-Babaei, P., Long, S., Foley, E., & McAllister, G. (2011). Understanding the contribution of biometrics to games user research. In *Proceedings of DiGRA 2011 Conference: Think Design Play* (Vol. 6, pp. 1–13). Retrieved from http://www.digra.org/digital-library/publications/understanding-the-contribution-of-biometrics-to-games-user-research/

Mirza-Babaei, P., Nacke, L. E., Gregory, J., Collins, N., & Fitzpatrick, G. (2013). How does it play better? Exploring user testing and biometric storyboards in games user research. In *Conference on Human Factors in Computing Systems - Proceedings* (pp. 1499–1508). https://doi.org/10.1145/2470654.2466200

O'Brien, D. A., Lawless, K. A., & Schrader, P. G. (2010). A taxonomy of educational games: Genres and applications. In Y. K. Baek (Ed.), *Gaming for classroom-based learning: Digital role playing as a motivator of study* (pp. 1–23). Hershey, PA: Information Science Reference.

Qian, M., & Clark, K. R. (2016). Game-based learning and 21st century skills: A review of recent research. *Computers in Human Behavior, 63*, 50–58.

Reiser, R. A. (2001). The history of instructional design and technology: Part 1: A history of instructional media. *Educational Technology Research and Development, 49*(1), 53–64. https://doi.org/10.1007/BF02504506.

Riot Games. (2009). *The league of legends*. Los Angeles, CA: Riot Games.
Rockstar Games. (2006). *Bully*. New York, NY: Rockstar Games.
Salomon, G., Perkins, D. N., & Globerson, T. (1991). Partners in cognition: Extending human intelligence with intelligent technologies. *Educational Researcher, 20*, 2–9.
Schenk, S., Lech, R. K., & Suchan, B. (2017). Games people play: How video games improve probabilistic learning. *Behavioural Brain Research, 335*, 208–214. https://doi.org/10.1016/j.bbr.2017.08.027
Schrader, P. G. (2008). Learning in technology: Reconceptualizing immersive environments. *AACE Journal, 16*(4), 457–475.
Schrader, P. G., Deniz, H., & Keilty, J. (2016). Breaking SPORE: Building instructional value in science education using a commercial, off-the shelf game. *Journal of Learning and Teaching in Digital Age, 1*(1), 32–46.
Schrader, P. G., & Lawless, K. A. (2007). Dribble files: Methodologies to evaluate learning and performance in complex environments. *Performance Improvement, 46*(1), 40–48.
Schrader, P. G., & McCreery, M. P. (2012). Are all games the same? Examining three frameworks for assessing learning from, with, and in games. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 11–28). New York, NY: Springer.
Schrader, P. G., McCreery, M. P., & Vallett, D. B. (2017). Performance in situ. In Y. Baek (Ed.), *Game-based learning: Theory, strategies and performance outcomes* (pp. 421–437). Hauppauge, NY: Nova Science Publishers.
Shaffer, D. W., Squire, K. R., Halverson, R., & Gee, J. P. (2005). Video games and the future of learning. *The Phi Delta Kappan, 87*(2), 104–111. https://doi.org/10.1177/003172170508700205
Shalizi, C. R. (2006). Methods and techniques of complex systems science: An overview. In T. S. Deisboeck & J. Y. Kresh (Eds.), *Complex systems science in biomedicine* (pp. 33–114). New York, NY: Springer.
Shute, V. J., Ke, F., & Wang, L. (2017). Assessment and adaptation in games. In P. Wouters & H. van Oostendorp (Eds.), *Instructional techniques to facilitate learning and motivation of serious games*. Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-39298-1_4
Squire, K. D. (2006). From content to context: Videogames as designed experience. *Educational Researcher, 35*(8), 19–29. https://doi.org/10.3102/0013189X035008019
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257–285.
Tettegah, S. Y., McCreery, M. P., & Blumberg, F. C. (2015). Toward a framework for learning and digital games research. *Educational Psychologist, 50*(4), 253–257.
Vallett, D. (2016). Soft failure in educational and learning games. In R. Lamb & D. McMahon (Eds.), *Educational and learning games: New research*. New York, NY: Nova Science Publishers.
van Merrienboer, J. J. G., & Kirschner, P. A. (2018). *Ten steps to complex learning: A systematic approach to four-component instructional design* (3rd ed.). New York, NY: Routledge.
van Merrienboer, J. J. G., & Sluijsmans, D. A. (2009). Toward a synthesis of cognitive load theory, four-component instructional design, and self-directed learning. *Educational Psychology Review, 29*, 5566.
Ventura, M., & Shute, V. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior, 29*(6), 2568–2572. https://doi.org/10.1016/j.chb.2013.06.033
Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research, 34*(3), 229–243.
Wouters, P., Van Nimwegen, C., Van Oostendorp, H., & Van der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology, 105*(2), 249–265. https://doi.org/10.1037/a0031311
Young, M. F., Slota, S., Cutter, A. B., Jalette, G., Mullin, G., Lai, B., … Yukhymenko, M. (2012). Our princess is in another castle: A review of trends in serious gaming for education. *Review of Educational Research, 82*(1), 61–89.

Chapter 3
Summative Game-Based Assessment

Andreas Oranje, Bob Mislevy, Malcolm I. Bauer, and G. Tanner Jackson

3.1 Introduction

The interplay between *games*, *assessment*, and *learning* has so far been considered primarily within a formative assessment setting, exploring how design frameworks from these fields are connected and disconnected (e.g., Mislevy et al., 2016). The focus on formative applications is natural because of the close connection between the learning that happens in games when trying to master a new set of challenges that the game designer has put in front of the player and the learning that happens in education when trying to master a new subject, problem, or topic. However, there are other motivations to connect games, learning, and assessment that span beyond formative assessment. These include the notion of situating assessments in highly complex and interactive environments that may elicit the recruitment of skills and knowledge that cannot be gauged with the same level of fidelity achievable in traditional assessments. This is often referred to as "hard-to-measure" skills (Stecher & Hamilton, 2014). A second motivation is the notion that these interactive environments are inherently motivating due to the application of sophisticated game mechanics and, therefore, bring out the best performance. We will expand on this later on extensively. Regardless, there is ample reason to consider game-based assessments (GBAs) for assessment purposes that are summative in nature (e.g., interim, benchmark, end of course, accountability).

The purpose of this chapter is to extend what has been developed and learned about formative game-based assessments (GBAs) into summative assessment practices. We are trying to further understand a joint design space for games and assessments, develop a common design framework, and relate that to specific use cases, by attempting to address the following key questions:

A. Oranje (✉) · B. Mislevy · M. I. Bauer · G. T. Jackson
Educational Testing Service, Princeton, NJ, USA
e-mail: aoranje@ets.org; rmmislevy@ets.org; mibauer@ets.org; gtjackson@ets.org

D. Ifenthaler, Y. J. Kim (eds.), *Game-Based Assessment Revisited*, Advances in Game-Based Learning, https://doi.org/10.1007/978-3-030-15569-8_3

1. What are core game design principles, why are they important, and in what way do they align with the goals and principles of summative assessments, particularly those that use simulation- and scenario-based tasks? Are there goals and principles that should be adopted? Particularly, how would that alter the types of claims we make and the psychometric and statistical models we would develop and apply? In other words, what are the intersections and compatible constraints among the learning, game, and assessment design spaces?
2. Which, if any, game design principles are less compatible with different types of summative assessments and why? From an engineering optimization perspective, what are the key trade-offs? How can we develop a common framework within this joint space?
3. What practices should summative assessments borrow from game development environments and how would we modify them? This question centers primarily on development processes (from ideation, to release, to data monitoring, and to product updating). How would these borrowed practices alter assessment processes and practices? Can we borrow machinery across the fields to address the trade-offs and build effective GBAs?

These questions are relatively broad and cannot be fully answered based on what we know so far. However, we can make substantial progress toward providing a compelling, alternative frame for current and future assessments. We do so in this chapter by first discussing some use cases of summative GBAs to put considerations and discussions in a practical context. Next, we provide foundations, perspectives, and motivations in an attempt to share our underlying value and belief system. Subsequently, we develop links and design trade-offs between games, GBAs, and summative assessments. The final part presents design considerations for summative GBAs. We intend that this chapter provides a basis for further exploration and use of design processes to support the development of game-based summative assessments.

3.2 Use Cases

Summative assessments denote a class that contains a wide range of assessment types. Answers to the key questions above will be very different for different types, and therefore, we need to be more specific about what kinds of use cases we anticipate. We describe three particular cases that we believe span the majority of summative assessments currently in existence or under development: drop-from-the-sky, accreditation, and individual. Following descriptions, we provide in Table 3.1 some key characteristics that are relevant in the discussion about GBA.

Drop-from-the-sky assessments, or more formally called “group-score assessments” such as the National Assessment of Educational Progress (NAEP), the Trends in International Mathematics and Science Studies (TIMSS), and the Program for International Student Assessment (PISA), are comparative assessments of

Table 3.1 Overview of most common summative assessment use cases

	Drop-from-the-sky	Accreditation	Individual achievement	Individual certification
Stakes for participants	Low	Low	High	High
Stakes for other stakeholders	Medium	High	Medium	High
Administration time[a]	Short to medium	Medium to long	Long	Long
Opportunities for GBA	Increase engagement, assess hard-to-measure constructs	Increase engagement, assess hard-to-measure constructs	Increase predictive validity, fairness	Lower cost and higher standardization on practical components
Context	High generalizability, low context	Low to medium context	Low	Mostly highly contextualized

[a]Given the range of assessment designs, administration times were categorized as “short” for 60 min or less, “long” for 3 h or more, and “medium” for everything in between

educational outcomes across educational systems. They are not tied to a particular curriculum, hence the term drop-from-the-sky, and are low stakes for both participants and most potential stakeholders up to a certain level (e.g., teachers, principals, school districts, but not states or countries). In fact, no results are divulged at individual levels (e.g., participant, school). The assessments are generally short in terms of test administration time, which can be a challenge for GBAs that might require some time to set a necessary context for a particular activity (see last row in Table 3.1) or to learn control mechanics. On the other hand, these assessments tend to support active research agendas and study and implement innovations in assessment frequently. Engagement can be a factor in these assessments, given their low-stakes nature for test takers, particularly at higher grades, and therefore, GBAs may be able to increase engagement. Whereas most of these assessments target relatively traditional academic subjects, there have been some ventures into different types of skills (e.g., collaborative and adaptive problem-solving in PISA).

Closely related to drop-from-the-sky assessments are exams used in the accreditation of institutions, such as the Major Field Test (MFT). These exams are somewhat more targeted to a specific performance range (e.g., undergraduate seniors) but are largely low stakes for participants. However, they are high stakes for education systems (e.g., universities) in order to obtain accreditation and, commensurately, be eligible for funding, including their students being able to obtain federal student loans. Some systems use these assessments also for program improvement. The low-stakes nature of this type of assessment for participants in combination with the demographics may translate to lack of motivation to participate. For these assessments, focusing on twenty-first century skills (e.g., design thinking, collaboration) might be attractive as a way to gauge student learning effectiveness across majors in order to assess institutional effectiveness in addition to departmental effectiveness.

In addition, the typical unconstrained environments of GBAs could lend themselves well to demonstrating and measuring these more complex types of skills.

The majority of summative assessments are high-stakes individual-level assessments used to gain entrance and admission, qualify for certain jobs or benefits (e.g., advanced college placement), obtain certification, and so on. Examples are TOEFL, bar and medical certification exams, computer and network certification exams, and the SAT and ACT. They are predictive in nature, meaning that these assessments not only make claims about what a test taker currently knows but also serve as an indicator of future performance (e.g., will likely succeed in college or graduate school, can perform certain tasks such as residential plumbing/HVAC). Motivation is hardly an issue in these assessments given that they are usually governed by a warrant (e.g., obtaining a license, entering a selective university). They might benefit from more interactive, authentic, but virtual assessment components, for example, to predict whether someone can indeed do a particular operation instead of relying on expensive and time-consuming practical exams. In such cases, the contextualized environments that GBAs provide could be an advantage, rather than an impediment to generalizability. We distinguish between individual achievement assessments and certification assessments in Table 3.1 and subsequent discussions.

3.3 Foundations

The foundations for summative game-based assessment contain psychometric considerations as well as theoretical and practical considerations for game design.

Mislevy et al. (2014) provide psychometric considerations for GBAs and establish three psychometric paradigms while discussing the various challenges and opportunities that arise for assessment design, psychometric modeling, and data analysis. These paradigms can be summarized as follows: (1) assessment entirely outside the game environment; (2) assessment within the game environment, based on a priori designed work products that are explicitly expected in the game; and (3) game play as assessment based on a priori defined (classes of) actions. This provides an assessment dimension to the distinction of "small g" and "big G" placing video games (small g) in a larger ecosystem of learning (big G) (Gee, 2007). Subsequently, Mislevy et al. explore how and where the Game Design Framework (GDF) developed at Electronic Arts Entertainment is compatible with the evidence-centered design (ECD, Mislevy, Steinberg, & Almond, 2003) framework and develops a merger of the two, coined evidence-centered *game* design (EC*g*D). The key idea is to use agile-based development cycles common in (entertainment) software development, including green/yellow light pitches as go/no-go decision points, concept, preproduction, production, postproduction (all macro), and sprints (micro), to cycle through the various ECD layers concurrently but with different intensities devoted to each of the layers across subsequent cycles. Finally, analysis approaches as well as psychometric opportunities are discussed. Paradoxically, through log files, at the same time very sparse (e.g., diffuse, ambiguous, undefined data points, partial depending on what the test taker did) and very rich information (e.g., trace data, click streams, continuous tracking) is harvested. Opportunities present themselves in how such data can be used to develop different types of adaptation and responsiveness, improve engagement, and present meaningful and actionable reports. Those opportunities are different for different use cases of assessments, and an inventory of applications is provided in the paper.

A widely cited scientific study of game design is *Rules of Play* (Salen & Zimmerman, 2003), which takes an observational approach defining and describing elements of games (digital and analog) as schemata and perspectives. Salen et al.'s foundation is to describe games through rules, play, and culture, each providing different ways to uncover the many complex layers of games. Under rules, schemata focus on the definition of games, meaning, how rules shape games, and how formal and informal rules play different roles at different times. Play focuses on the experience of players, player types, how games emerge as experiences, and concepts such as engagement and flow. Culture, finally, focuses on the idea of games being situated in society, and many links can be drawn with the socio-cognitive perspective for assessment, for example, in how meaning is established and value is assigned. Possibly, one of the most important contributions is the search for a definition of play. Whereas this search is not resolved nor is it clear how it could be resolved, a broad yet useful approximation that we adopt here is that *play is the freedom to move in a constrained space*. This definition goes well beyond our focus on electronic games with some video component but is useful in that it directly makes clear the inherent tension between control by a game player and control by a game designer. What constraints does a game designer place in a space to create a space players want to interact with and how much control does the player get to manipulate and move about that space? This is somewhat parallel to the tension between a student's freedom to approach, answer, and solve problems in unique and creative ways and an assessment designer's desire to verify with confidence the presence or absence of a particular skill or element of knowledge. We revisit this tension later on.

In contrast, a highly practical perspective is provided in *The Art of Game Design* (Schell, 2008), using 100 lenses through which to view games by providing questions a game designer can ask to improve the odds of creating a successful game. The foundations of game design consist of the elemental tetrad: mechanics, story, aesthetics, and technology. These four components have to work together in unison to create compelling, engaging games. Yet, the experience is in the player's mind, and games at best can only provide the conditions for an optimal experience. It is therefore not surprising that the core mechanics include one skill that resides inside the player. According to Schell, the core mechanics are space, objects (with attributes and states), actions, rules, skill, and chance. Space defines the "magic circle" of play, not surprisingly also being a key component in Salen and Zimmerman (2003). As a mechanic, it is an entirely abstract, mathematical construct of what places exist within a game and how those places are defined, including whether they are continuous or not, how many dimensions they have, and whether the bounds of the spaces are connected or not. Objects are placed inside the space and can be manipulated. They can range from characters to tokens, and each has a particular state (e.g., a token can have various values, assume a particular color, and be either located behind a rock waiting to be found or in a character's inventory) and attributes (e.g., can be picked up, can be moved, can be given to someone, can be used to buy something). Schell considers anything that can be seen an object. We would argue that much of what can be seen is part of the aesthetics and might not provide directly for interaction, a key attribute of objects, even though they influence the tone and emotional environment, likely conditioning the meaning (and evidentiary value) of interactions. Actions are the things a player can do with objects (e.g., move forward, jump, grab, start a dialogue with a character). They are the verbs associated with the game play. Rules are probably the most fundamental component in that all the other mechanics, with the exception possibly of "skill," being a player mechanic, can also be defined in terms of rules. Most importantly, rules attribute most if not all meaning to the game situated within the space, objects, and actions provided and define the goals of the game. Skill defines the abilities a player is expected to bring to the game and, likely (and increasingly), hone by playing the game. These can be physical, mental, or social in nature. Finally, chance is considered a separate game mechanic as a lot of play in many games emerges from the notion that there is not a certain outcome to every action. For example, it can drive suspense, surprise, challenge, and overcoming adversity, which in turn can improve engagement and contribute to the creation of compelling experiences.

Across these game design theories, the following ideas are particularly important when considering applications of these features to summative assessment:

- Games are emergent systems, in which compelling and engaging experiences emerge from a number of factors, one of them being the game itself.
- Games require meaningful interaction, in which meaning is provided through several powerful layers: the mechanics of the game itself, the psychology of the player, the immediate social environment of the player(s), and the cultural setting of the game and the player.
- Games are governed by rules, and seemingly minor changes in rules can have tremendous impact on the game experience. Successful games tend to maximize player impact with minimal input while ensuring meaningful interaction.

3.4 Perspectives

There are several perspectives associated with GBA that provide context to some of the arguments made in favor and against GBA. It is critical to realize that these arguments apply differentially to the various use cases and that the same design space opportunity can be highly desirable in one use case but an anathema in another use case. Indeed, occupying an undifferentiated position for or against GBA does little justice to the complexities of and opportunities in the design space. Successful GBA development is premised on a far more discriminating view.

Mislevy, Behrens, DiCerbo, and Levy (2012) state that games and assessment principles are compatible because they both build on the same principles of learning. We would like to be more concrete and argue that games and assessment are both grounded in psychology and that there are several direct links. Games and assessment both serve for players/test takers to show skill, knowledge, and abilities, show this relative to others, reach clearly stated goals, and receive rewards for good performance. Where they differ is that games focus on learning in the service of engagement and enjoyment (Koster, 2014), whereas assessments focus on cognition and learning itself (Bennett, 2010). Following Deci and Ryan's (2000) *Theory of Self-Determination*, the three pillars of motivation they identify are achievement, control, and relatedness. Games and assessments fulfill all three of these, albeit at times in different ways. Achievement seems ubiquitous across games and assessment through leveling up, the enjoyment of figuring something out, and obtaining satisfactory scores. A player's need for control is fulfilled in games in the sense that a lot of fulfillment might stem from gaining control over (seemingly) chance elements or hidden logic in the game, so that a virtual enemy can be defeated or a puzzle solved. A test taker's need for control is fulfilled in assessment in the sense that there is an intended direct relationship between achievement and the outcome of the assessment and that achievement can be controlled to a large extent through training. Finally, whereas the social aspects of games are well known and create relatedness in many ways, assessments provide relatedness in at least two possible ways: the shared experience of taking the assessment itself and, depending on the outcome, belonging to a particular class of achievers.

A related perspective that is often mentioned in game design theory, game-based learning, or game-based assessment is the idea of *flow* (Csíkszentmihályi, 1975, 1990), a state of heightened engagement and concentration during which learning productivity and assessment validity increase (Schmit & Ryan, 1992). Flow is a central concept in game design (Salen & Zimmerman, 2003; Schell, 2008), and creating experiences that induce it is probably the most critical pursuit for game designers. It could, however, be equally important for assessment, assuming that the goal is to obtain a read on the highest ability of a test taker. A core mechanic for inducing flow is to ensure that the level of challenge and the skill of the player are in balance. If the challenge is too great, anxiety or frustration is likely induced. If the skill of the player is too great, boredom may follow. Therefore, as a player becomes more able through playing a game, games make players "level up" in order to maintain this critical balance. Summative assessments have similar mechanics in computerized adaptive testing (CAT; Weiss & Kingsbury, 1984). Whereas the skill of the test takers in CAT is generally considered static at a particular time and the motivation for adaptation is usually related to measurement precision, a better experience is arguably created for the test taker by adapting the test difficulty to the test takers' ability.

A complementary perspective, related to the notion of games being emergent systems, is the adoption of a socio-cognitive or situative perspective (Greeno, 1998) centering on the (environmental) experience of the test taker (or learner) in combination with the mental frame within which those experiences are placed. This is a departure from more traditional construct-centered approaches to test development, in which constructs for the most part are considered as knowledge and skill representations. This is not to say that reliability, validity, and fairness are not fundamental virtues. The key is that we need to think differently about how we characterize and evaluate those virtues and how we establish evidence. Foremost, it means that there can be a considerable inferential distance between the presented assessment materials and what the test taker makes of that. In some cases, the final product has specific features that would unlikely be obtained without carrying out a specific process. However, in many other cases, we have the opportunity to strengthen our inferential argument by looking at both process (Kerr & Chung, 2012) and product. Most importantly, we need to be precise about the relationship of the claims that we make and the content coverage and generalizability (Dunbar, Koretz, & Hoover, 1991) on which we base those claims (Frederiksen & Collins, 1989; Messick, 1994). It is for that exact reason that game-based assessments so far have focused on formative applications that attempt to offer learning and inform learning in low-stakes environments that tolerate a greater deal of uncertainty.

A final, sociological perspective or notion is about generational shifts over time. For example, while TV is still a major source of screen-based consumption for young children (Rideout, 2014), game consumption is catching up quickly at the cost of nondigital activities (The Entertainment Software Association, 2013). This means that the value that society places on game experiences is different than it was mere decades ago. As a result, an evaluation of the merits (and demerits) of summative GBAs is likely cast within a specific generational context. Similarly, summative assessments for college admission were a particularly important part of promoting an equitable, merit-based system for access to higher education. Currently, it seems, society and employers are more concerned about a lack of twenty-first century skills such as global awareness, creativity and innovation, information literacy, and cross-cultural competence (e.g., Scott, 2017). These skills are not typically associated with admission testing and seem to require a more complex assessment environment to assess appropriately. We return to this later on.

3.5 Motivation for Game-Based Assessment

So far, we have discussed how games and assessment may correspond, provided an exceptionally brief introduction to game design theory, and offered some perspectives about GBA more generally. What we have not made clear is what game-based assessments add over existing assessment paradigms. There are certain claims about the virtues of more naturalistic or authentic assessments, but there have been far fewer attempts to substantiate such claims empirically (e.g., Ercikan & Pellegrino, 2017). For example, there are no studies that show that such assessments predict student learning outcomes. Yet, we have several reasoned motivations that are worth describing, with the understanding that there is an unfulfilled gap to provide empirical evidence about implied efficacy.

There are abundant indicators that there is a rapidly increasing interest in and adoption of cognitively based assessments by national and multistate assessment systems (e.g., PARCC, Smarter Balanced, NAEP). By interest in cognitively based assessments, we mean an interest in assessing and reporting on the cognitive processes involved in solving and reasoning about a problem (Leighton & Gierl, 2007). For example, what specific cognitive strategies did the test taker use and what cognitive errors were made as that test taker developed an answer to a question? How can that inform subsequent learning goals? Collecting this type of evidence is not a new interest or desire: there has always been interest in diagnostic, even formative, assessments that would be able to specifically pinpoint gaps in performance and preferably reveal how those gaps could be closed. Summative, high-stakes assessments are generally not very well set up to garner such information because their goal is to obtain highly reliable information about many things, and in fact, relatively little testing time is available given the level and breadth of information that is expected. Yet, significant investment in the development and validation of cognitive theory provides a basis for making more detailed claims about student performance and in particular about intermediate reasoning as evidenced by steps taken within a problem-solving process as indicators of partial understanding. Add to that advances in technology, it is possible to view games as a way to create safe, adaptive, and engaging (because of learning, achievement) learning and assessment environments to manipulate otherwise time-, space-, or cost-prohibitive objects (e.g., Shaffer, 2006; U.S. Army, 2012). In the following, we focus on specific game-based characteristics that could be particularly advantageous to apply to specific summative assessments.

3.5.1 Interaction

Alongside the development of cognitive and learning theory, substantial development of technology-based assessments has occurred that provides the opportunity to have interactive conversations with many test takers and collect data reflective of that interaction (e.g., Ramanarayanan, Evanini, & Tsuprun, 2019). The notion is that those interactive conversations provide a space to obtain more detailed, reliable information in relatively little time. Adaptive tests are an example of such interactions that operate at scale, tend to focus on reliability with respect to one or a few dimensions, and usually entail relatively discrete evidence components (i.e., items or tasks). Games have a long and successful history of developing, understanding, and capitalizing on meaningful interactions. Take for example a game such as The Sims (Electronic Arts, 2018), which provides the user with an extensive environment to build complex characters, maintain social relationships, and pursue life goals, all of which interact with each other in intricate ways. As such GBA is emerging not just as a promise (Gee, 2007; Klopfer, Osterweil, & Salen, 2009) but as a reasoned approach to learning and assessment (Mislevy et al., 2013). The argument presented here is that the extraordinary level of interaction typical in game environments can be leveraged for assessment to make inferences at a wider range of grain sizes than current assessments and also to reduce the time required to obtain reliable information about a wide range of knowledge, skills, and achievements (KSAs).

3.5.2 Hard-to-Measure Skills

Another area in which games tend to excel is in creating immersive environments and modeling commensurate complex relationships between game objects, spaces, and actions, represented in higher order rules. From the hyper-realistic war simulator Battlefield (Electronic Arts) to the addicting fantasy immersion in World of Warcraft (Blizzard Entertainment), such games present complex environments in which players need to choose from a wide range of possible actions, need to perform many actions simultaneously, and need to respond to highly dynamic circumstances and settings. Translating this to assessments, these environments and associated interactivity present the opportunity to measure skills that are highly complex. The notion is that those skills are typically difficult to measure effectively with shorter, discrete items, particularly when there are space, time, or cost constraints (e.g., atomic reactions, evolution, solar systems) in place, or, at least, would require a sizable number of items to collect evidence about all the nuances. Naturally, these environments are highly contextualized, which presents a potential barrier for generalizability. On the other hand, one could argue that these skills are only meaningful in particular contexts and not as generalizable as more basic or foundational skills.

Another class of hard-to-measure skills is what are commonly referred to as twenty-first century skills or soft skills such as creativity, collaboration, critical thinking, communication, information technology literacy, flexibility, adaptability, cross-cultural competence, initiative, leadership, and productivity (e.g., Partnership for 21st Century Skills, 2011; Stecher & Hamilton, 2014). All these skills are complex in nature, could be perceived to manifest as either skills or traits, and can be faked in basic self-report-based measures. GBAs could provide environments in which test takers can show these skills rather than report on them. Two obvious game genres in this context are role-playing and multiplayer games.

In sum, well-designed GBAs can provide rich contexts in which the use of complex skills is required to successfully navigate and solve problems. In addition, the level of telemetry that can be harnessed from the interactions that take place in said contexts can provide rich evidence to make inferences about these complex skills. It is important to note that when we argue that there is a particular fit of the assessment of more complex skills to GBAs, this does not imply anything about the length of such GBAs. That is to say, more targeted micro-games that focus on a particular set of interactions (e.g., social mechanics) can still be highly immersive and complex in the nature of those particular interactions.

3.5.3 Engagement

As an entertainment industry, honing maximum engagement is the most critical component of a commercially successful video game. Due to the emergent nature of games and the complex interaction between players and game rules, there are no fixed formulae. However, there are themes (e.g., battling an adversary, getting an early yet temporary view and experience of success, pursuing a larger-than-life quest) and mechanics (e.g., compelling environments, goals with levels, interaction with rewards) that are known to engage players (Koster, 2014) for a sustained amount of time and place players ideally between boredom (low challenge, high skill) and anxiety (high challenge, low skill). Each of these pursues basic motivators of meaningful achievement, control, and relatedness/belonging as described earlier. Relating this back to the assessment context, there are two important considerations: learning and student experience. As noted above, one of the primary motivators is achievement, which is obtained by learning new knowledge, skills, or abilities. As it turns out, the learning that is so integral to assessment (Bennett, 2010) is also fundamental to gainful game play (Mislevy, Behrens, DiCerbo, & Levy, 2012). In other words, games are especially well set up as environments in which KSAs are learned and tested, including when the learning itself (e.g., speed of learning) is the object of interest. In terms of student experience, most assessments, particularly summative, are not known for their engaging nature. After all, ensuring standardized testing conditions to yield reliable, fair, and valid results is paramount above all else. The argument we make here is that engagement is centrally important to validity and fairness. Different test formats resonate more or less with different test takers, and a particular mode induces maximum performance in only some students. The modes that currently dominate assessment practices (e.g., multiple choice, short essays) have, to a large extent, been based on logistical considerations (e.g., printing, physical shipping, machine scanning, human scoring) that no longer apply. We would go as far as *believing* that engagement is an obligation for assessments (although it might take a while before people pay with a currency other than their data and privacy to take assessments for their enjoyment). Yet, it is important to note that the type of engagement in assessment might not be the same engagement in commercial games and might not be very sensible to compare the two. Most importantly, assessments purport a larger goal beyond the assessment itself, to obtain some type of information and make some type of decision based on that, which transcends beyond the assessment itself. Commercial games generally do not. This larger goal at the very least means that there is some extrinsic motivation for participating in the assessment. Discussing intrinsic and extrinsic motivation and the transfer between the two is beyond the scope of this chapter, but the basic notion holds that the type of engagement of a GBA or a commercial game is likely not comparable.

Many GBAs have been developed in the past decade (e.g., Barab, Gresalfi, & Ingram-Goble, 2010; Metcalf, Kamarainen, Tutwiler, Grotzer, & Dede, 2011; Shute, Ventura, & Kim, 2013) across a wide array of academic topics and more general skills such as social and emotional learning and inquiry. Significant empirical data have been collected for many of these, and deeper understanding about the virtues and challenges is surfacing. For example, organizing and carefully designing telemetry up front is critical. That being said, a lot is still to be learned about context, transfer and generalizability, assessment reliability, and validity. In fact, it is important to note that not everything is better off in a GBA environment. For example, it is challenging to argue that basic skills' assessments are best conducted in a game environment in which reliability and generalizability have to be negotiated. Yet, assessment as we know it currently is, itself, a proxy in many ways, which we have come to accept as evidence. In the following, we discuss ways that the goals and methods of games and summative assessments can be at odds and what that implies for summative GBA.

3.6 Design Trade-Offs

As noted so far, there is substantial overlap across the activities developed, principles followed, and approaches taken in the design of games, learning, and (educational) assessments. This is particularly true for formative assessments (Bauer et al., 2017). There are also trade-offs that need to be dealt with up front when considering GBAs, particularly for summative assessments. We are highlighting the following design choices or trade-offs as we see them as particularly fundamental: competing goals, audiences served, and development practices. As will soon become clear, some of the trade-offs between games and summative assessments are parallel to those between formative and summative assessment. Note also that the design trade-offs differ vastly across use cases.

For many games, an important goal is to provide meaningful, engaging experiences. There are many ways to create such experiences and many more to disrupt and eliminate those. Possible ingredients for meaningful, engaging experiences are to provide freedom to explore, freedom to fail without serious consequence, choice, surprise, success, rich interaction, compelling narratives and themes, and variability, to name just a few. For assessments, an important goal is to establish, with reasonable certainty, that a player can (or cannot) do something or knows something, where that something is often generalized to a higher level of abstraction (e.g., grade-level mathematics ability rather than mastery of single-digit additions) and intended to be predictive. The most effective and scalable way to increase certainty is to ask the player to show the skill or knowledge of interest multiple times in a standardized fashion (i.e., replicability) and across various contexts. In other words, the focus on parallelism and repeated measures of a priori determined generalized constructs in assessment entails mechanics that, currently, do not work as well with the need for free discovery, choice, and surprise within meaningful contexts that is often central to enjoyable games. The notion of "currently" is to emphasize that this is likely a temporary state in the sense that there is interest in assessing (complex) constructs in environments that are more similar to the noisy environments for which performance is predicted. Some games quite skillfully build those environments and, therefore, could play a role in developing assessment.

Games provide immediate feedback related to directly observable goals because there is no desire to make claims beyond the boundaries of the game. Summative assessments pursue quite the opposite, making no or few claims about what the test taker just did in the test but instead making claims about what the test taker is able to do or will be able to do more broadly. In fact, the notions of reliability and comparability suggest that we would like to be able to make comparable claims even when different test takers encounter completely different sets of tasks. Usually fairly opaque reporting scales are employed to convey those claims at a higher level of abstraction. Subsequently, the two fields tend to have very different notions of immediacy. With notable exceptions, games primarily focus on goals that are short term (a couple of seconds to a couple of hours) and tend to fulfill instant gratification needs related to the activity. Summative assessments fulfill a variety of gratification terms, from the middle long term (e.g., college entrance examination) to long term (e.g., certification for a particular profession, job application). That being said, game studios are actively maintaining user relationships beyond a single purchase (e.g., platforms such as Steam and various console-based virtual stores), and it is quite likely that prediction will become an important part of game producers' repertoires in order to develop even more customized, engaging experiences and, as such, sell more games.

Another related area in which games and summative assessments show some tension around goals is the notion that successful games provide a way for all players to be successful in some form. Without a sense of accomplishment, games may quickly end up on the shelf. Games use a variety of techniques to accomplish this, including allowing players to select different levels of difficulty, replay the game as often as they want, save the game at various points, be introduced to new challenges slowly and with ample opportunity to practice and acquire the required skills first through inconsequential failure, level up many times or redo previous levels, use cheat codes in order to bypass particular challenging sections of the game, and play the game in ways the designers had not even imagined. In high-stakes summative assessments that are used for selection (e.g., admission, hiring) where there are more people interested than there are spaces available, a contest occurs (e.g., Holland, 1994; cited in Dorans, 2012). Unless in the service of a more accurate measurement of a skill or ability (Attali & Powers, 2008), there is no intention to provide all test takers with a way to be successful in some form. In particular, the assessment has no need to maximize the amount of time test takers engage with the test once minimum reliability and validity criteria have been met. While there is a thriving market around test preparation, the assessments themselves do not offer practice during the testing, and cheating (i.e., playing the assessment in ways that were not intended by the designers) is one of the most important threats to the credibility, fairness, and validity of individual assessments. One could argue that adaptive testing provides some type of leveling, but the test taker does not have much choice. Obviously, the goals of drop-from-the-sky or formative assessments are much better aligned with games in the sense that the intention of those assessments is that every learner must be successful by driving education improvement policies or by providing learning opportunities themselves, as well as indications of how the learner can improve and master all required KSAs. The growth that particularly formative assessments aspire to becomes a confounding factor for summative assessments that aim to provide a singular comparison of what students know and can do.

Related to the idea of competing goals is the notion that games can be deemed successful when the engagement and entertainment value of the game can be validated, pending many other factors such as cost, availability, platform, and so on. For (summative) assessments, the validation of the assessment claims is what constitutes value and success, pending many other factors such as costs, availability, score reporting quality and timeliness, and so on. In other words, the validation needs are distinct because the goals are.

In summarizing the goals, opportunities, and challenges that the game, learning, and assessment fields hold, several design trade-offs in the design of GBAs surface. Some of the most salient are the following:

- The surprise and variability important for games versus the need for repetition and standardization in assessments.
- The need to obtain specific, controlled evidence about a person's knowledge or skill in assessments versus the desire to explore freely and provide choice (to follow a particular path) in games, thereby possibly never providing evidence about some of the skills.
  - Note that there are many games that have the opposite problem. Games such as Tetris or Pong are very narrow and do provide very reliable evidence about a particular skill, albeit a skill that is narrow and not very applicable outside of those games.
- The attraction of a compelling narrative for a targeted segment of the population in games versus the goal of cross-contextual understanding and the desire to make fair, generalizable predictions for a very large segment (e.g., all public K-12 students in a state) in assessment.
- The focus on immediate rewards and gratification in games versus the longer-term goals associated with assessment.
- The goal to select, classify, and rank in assessments versus the opportunity to try (and fail) until the player succeeds in a game.

The key challenge is to find instances of games and assessment in which the goals of the two fields can be met simultaneously and up to some acceptable level. While none of the goals may be met maximally at the same time, we are looking to (1) reach levels for each one of multiple criteria that are good enough for our needs and (2) end up with a joint solution that is optimal across those criteria. For example, in GlassLab's *SimCityEDU: Pollution Challenge!*, the ability to explore was preserved, though in a somewhat restricted sense by removing a number of the capabilities (e.g., no civic buildings were activated), and the assessment goals were met by developing multiple related challenges and requiring specific actions to be applied multiple times in order to be successful in the game (e.g., replacing coal power plants with cleaner alternatives).

The audiences that are served in a game or an assessment are quite different. Typically, games serve the player(s) and, possibly, parents of the player who want their offspring to be entertained in a safe and age-appropriate fashion. Playing the game is not a requirement imposed by anyone (except the player) in order to obtain something else. For all practical purposes, the stakes are low to nonexistent, and failure has little consequence. In fact, failure is an important part of improving the skills necessary to play the game and, ultimately, succeed. The main barriers to playing a game are not having access to the necessary equipment, lack of time, and the possible presence of physical or cognitive barriers in using the equipment (e.g., disabilities).

Summative individual assessments ultimately serve the test takers to reach some goal and remove a barrier placed by someone else (e.g., admission, certification). However, unlike in games, there is a host of other audiences involved besides the test taker. Examples are an institution that needs to make good admission and scholarship decisions, an institution up for accreditation, a principal evaluating teachers, a teacher looking to tailor instruction, researchers and policy makers evaluating the effectiveness of educational policies, policy makers making funding decisions, and so on. By listing out these audiences, it becomes immediately clear that assessments are used for decisions about people (and the institutions they belong to), which demands a high degree of fairness.

A possible tension arises between games and assessments around fairness. Assessments gain acceptance when they provide a fair opportunity to everyone to show proficiency. Therefore, assessment designers go through a thorough process to make sure that the outcomes of a test are identical for those with the same proficiency, regardless of anything else (e.g., gender, age, race/ethnicity, student disability). In contrast, most games are built for specific groups of players (i.e., markets) that are known to buy and play games. For example, action games such as Battlefield 4 are primarily male-oriented (adolescent and up), and games such as Disney Princess Royal Ball are primarily geared toward preteen girls. In other words, the techniques that help games be very targeted to a specific audience and bring commercial success would, at the same time, disqualify an assessment. Furthermore, even within game genres that are nonlinear in terms of the sequence of activities (e.g., an open area to build on or search in), some students may not receive the same opportunity to show their skills as others because they may not have encountered similar problems to solve. The implication is that game genres and mechanics have to be chosen carefully and scrutinized against a lens of fairness.

By the same token, game-based assessments could provide alternative modes of assessments for students who may not do as well in more standardized testing environments. In that sense, these assessments could actually increase fairness. This is likely highly dependent on the type of assessment, the construct of interest, and the context of the assessment and needs to be studied experimentally to substantiate a claim of fairness. In particular, it would be critical to ensure that the key features of the targeted skills be challenged in comparable ways even though other aspects of the assessment would differ for different test takers based on their background knowledge and interests—a "conditional" sense of fairness (Mislevy et al., 2013). To do so requires some kind of cognitive model of the targeted skills, such as a learning progression or theory of domain performance. This idea arises in connection with the following considerations as well.

With several touch points in the humanities, the design of games and assessments otherwise primarily draws upon psychology (cognition, learning, motivation) and logic (computer sciences, reasoning). They also share a grounding in empiricism. That is to say, what is known, familiar, or held true in both fields, at least until proven otherwise, is based on empirical research that has tested competing hypotheses in (samples of) populations. For the discussion here, we consider as a hypothesis any idea or practice about what works or does not work for assessments or games. For example, an assessment or learning hypothesis could be a learning progression, stating that there are some number (e.g., five) of levels associated with a construct (e.g., systems thinking, argumentation, proportional reasoning) and that those levels are progressively more complex. (Ideally the progression would specify features of situations students encounter and the kinds of things they can do at each level well enough to guide game/task developers but with enough generality that different narratives, goals, or mechanics could be employed.) The top level describes the most sophisticated facility with the construct at hand. A game hypothesis could be a game mechanic (e.g., the way a character is controlled, how probability is assigned to an outcome, how many points are earned at a particular junction), stating that the way the mechanic is defined and executed creates effective and engaging game play. The tuning of the mechanic represents a perfect or near-perfect interaction with all the other mechanics to yield optimal play. Note that neither of these hypotheses may in fact hold, but they are defined in a way that they can be tested.

The approach that the two fields take, however, is vastly different in terms of how well hypotheses are developed before they are tested, how often hypotheses are tested, and how quickly alternative hypotheses are developed. This taken together results in decidedly different development speeds. For assessments, the tendency and, certainly, the tradition have been to develop hypotheses relatively fully before any empirical testing occurs. This is in part due to some maturity of the field including a rich literature that can be drawn upon to build hypotheses. The empirical testing itself generally takes the form of relatively large samples (e.g., several hundred up to many thousands) to support statistical modeling of the resultant data. The data collection at those levels is quite expensive, and therefore, each major study may only entail a single data collection based on carefully constructed instrumentation (e.g., a test, an interview protocol). In contrast, the field of game design relies far more on tacit knowledge from past experiences, and many hypotheses are quickly developed, tested in small samples (e.g., one or two subjects in a play-testing session), and based on relatively low-grade instrumentation (e.g., wireframes, prototypes, single-slice minimally functional prototype) before the investment is made to create production-grade materials (e.g., art, coding). Once the end-production process has started, the core mechanics are for all practical purposes in lockdown mode.

A possible tension between assessment and game design might arise around the different development speeds. Generally, game design will move much faster early on while assessment design is still developing hypotheses and instrumentation. Without careful coordination, the assessment designers might bring critical requirements to the table when the game designers have already placed the mechanics in lockdown, at which point the assessment goals of the GBA can no longer be fully realized (e.g., Klopfer et al., 2009, p. 19 and 31). Alternatively, at a fairly late state in the development of the GBA, critical requirements surface that result in expensive game redesign work and significant delivery delays. While we presented this as a possible tension, it is really a design trade-off in the sense that the game design approach allows for a lot of early, empirically tested, development iterations, which could be very effective in a field of assessment that is relatively underdeveloped. Regardless, and on a much more practical level, it is important to start the assessment design very early on, in a highly iterative fashion between game and assessment designers, and to build in significant time during which the game is in the concept and preproduction stages. This will be much longer than what is considered normal in game development for those stages and possibly shorter than what is the norm in assessment development. Even more important is to develop reusable concepts, methods, elements, mechanics, and processes jointly that can be used to develop future GBAs. In other words, the methods that both fields currently rely on, such as design patterns or automated scoring engines in assessment and engagement or feedback mechanics in games, have been established over many iterations. For GBA, this would not be any different, including carrying out many data-driven feedback loops to sharpen evidence elicitation and developing classes of mechanics that can respond to modifying situations, affordances, and work products.

3.7 Designing Summative Game-Based Assessments

Now that we have discussed in what ways games and summative assessments would or could fit together and indicated design trade-offs, we lay out considerations for designing and analyzing summative GBAs. This section builds on the ideas that summative GBAs are used for traditionally hard-to-measure constructs that require a considerable amount of interaction and that more basic skills are better assessed in more traditional formats. This section also relies on the notion that there are many types of summative assessment and that each use case has different implications for GBA design.

3.7.1 Gamification, Edification, and Serious Games

Gamification, or the application of game principles to a nongame context including educational assessments, continues to be the object of substantial controversy (e.g., John, 2014). Arguably, a lot of that controversy might be definitional, but the main tension that often surfaces is when an existing learning or assessment activity is enhanced by applying some game mechanics. In the most basic and problematic form, a game is offered as an extrinsic reward for successfully completing a non-game activity (e.g., some rote learning work) and is not much different from a parent making game console time contingent on completing homework assignments or achieving particular grade levels. To state the obvious, we do not consider this game-based assessment, and the balance is entirely tipped in favor of assessment (and possibly learning). The opposite would be edufication, in which you try to (retro) fit learning and assessment experiences into a game in a way that is not central to the game. More subtle applications use reward systems (e.g., points, stars, tokens) that are more directly connected to the activity itself, provide immediate feedback on success and failure, and set clear intermediate and general goals. However, it is not clear why this would be considered gamification rather than aspects of good learning or assessment design. Nothing about those enhancements creates emergent, highly responsive, or complex rule-based interactions. There is a newer breed of summative assessments adopting scenarios and, within those, simulations to tap into more complex practices and skills-based (rather than knowledge) constructs (e.g., NAGB, 2014). These environments do adopt a number of core game mechanics in a meaningful way and are emergent in the sense that the test taker can be (temporarily and mentally) transported into an interesting space where achievements are to be accomplished. We consider scenario-based assessments, sometimes also referred to as serious games, as a reasonable approach for development and research toward summative game-based assessments. When assessing generalizable skills, such an approach to GBA makes most sense in lower-stakes contexts such as drop-from-the-sky- and accreditation-type assessments where reliability at the individual level is traded for construct validity of more complex constructs. This approach would also make sense in certification-type individual assessments where demonstrating mastery of a particular practice is critical from a safety perspective (e.g., electrician, heart surgeon) while a standardized and cost-effective examination is desired. That being said, the development of such scenario-based tasks is still in its infancy, and the following design considerations are pertinent in the pursuit of more engaging, meaningful, and effective interactions to measure complex constructs:

- Current instantiations are highly constrained and linear in order to achieve relatively certain measurement of relatively narrow concepts. Expanding the (complexity, richness of the) space test takers are able to explore will be critical in order to better achieve emergence through choice and compelling representation. This will, in turn, require more detailed competency models to support that level of complexity, without losing measurement control, likely supported by various loops that bring the player back or forward to yet untested competencies. As noted before, this does not necessarily need to imply longer interactions. A Cisco Networking Academy (CNA) simulation-based game or troubleshooting problem might take a half hour or more in a learning context, but summative tests and certification exams obtain more reliable evidence with several shorter simulation

tasks of only several minutes each to address the low generalizability phenomenon common to performance assessments (Dunbar et al., 1991). For broader use in individual achievement tests, developing very short interactions will be critical.
- Current instantiations are relatively limited in their responsiveness to test takers' actions, other than recording those actions. There are three considerations for responsiveness worth pursuing:
  - Increase responsiveness to create a more realistic and richer environment that can more effectively scaffold and provide a level of feedback that invites further exploration and discovery.
  - Increase responsiveness to create a more realistic environment that can better represent real-world stressors in a practical component.
  - Increase adaptability to optimally balance between boredom and anxiety given a student's abilities represented as a complex constellation of constructs.

  All of these would, to a large extent, rely on more detailed competency models similar to the first set of considerations, in addition to fairly sophisticated psychometric models.
- Current instantiations develop narratives and characters that are relatively shallow in order to minimize the time required to introduce the scenario and to limit potential construct-irrelevant factors. The downside is that such scenarios remain at a level of abstraction that does not necessarily serve all aspects of construct validity. A key consideration would be to develop richer narratives through symbolism, graphical representations, and possibly development of environments that can serve as a space for multiple assessments, similar to how expansion packs and sequels operate. As mentioned earlier, context dependence relative to the need to make generalized claims may or may not be a significant issue as the context might be foundational to those generalizations. (For example, a detailed inquiry game in the context of volcanoes can provide strong evidence about a student's ability to carry out inquiry with volcanoes but necessarily less about inquiry skills more generally construed.)

3.7.2 Principled Design

As we argued throughout and signified by introducing ECgD early on, principled design is a critical component for effective GBA development. Without a principled design process, one is likely to end up with a great game that has poor assessment properties or vice versa. The implications are that trying to retrofit an assessment in an existing game or game elements in an existing assessment is not only difficult but is likely to fail because neither was designed or optimized for this purpose. Besides the general recommendation to work within ECgD, the following considerations are important to highlight:

- Perhaps the most critical aspect is to make sure the assessment goals and constructs are represented meaningfully in game objects, rules, and spaces and vice versa. This requires a balanced design approach that considers both fields simultaneously. In a summative context, those objects, rules, and spaces need to be amenable to shorter interactions in order to be able to present a wider range of contexts and improve generalizability.
- As a derivative, any design changes have to be considered and negotiated with consideration for the entire system, which can be challenging when vastly different sets of expertise, and therefore resources, are required. On the other hand, if done well, the measurement and game qualities coincide fully. A unified design framework that includes design objects such as macro- and micro-designs, described earlier, that bring together game and assessment design elements that show the degree to which they satisfy game and assessment design criteria needed for the solution helps this happen. This is true for any GBA, summative or formative.
- Finally, it appears that an effective approach for marrying game and assessment design is to develop four guiding documents (or otherwise information repositories), in addition to possibly more technical specifications that are specific to the technology being used (e.g., Unity3D, HTML5).

  - Macro-design—what are the game themes, objects, and narratives and what is the student model? This document is the basis for pitching the GBA and working out a high-level concept. For summative GBAs, the themes, objects, and narratives have to be relatively short, be highly recognizable, and use easily understood mechanics detailed in the following document.
  - Micro-design—what are the specific mechanics, game flow, story line, levels, evidence model(s), and task models? This document is the basis for developing the GBA.
  - Telemetry—what elements are captured from the activity and what does each element mean? How does each element contribute to a generalizable assessment argument and what elements can help identify and account for context effects? This document is the basis for building an evidence collection engine including the database and data flows into and out of the system.
  - Evidence identification and accumulation—how are each of the elements scored and what does the inferential (statistical) model look like? Does the model account appropriately for context and generalizability relative to the purpose of the type of summative assessment, where certification assessments might require more context and achievement assessments relatively little? This document is the basis for building the assessment engine including scoring (taking elements from the telemetry document), adaptability (directly related to elements from the micro-design document), and reporting. This document is also called the "four-process document" referring to the process architecture (Almond, Steinberg, & Mislevy, 2002) covering task selection, evidence identification, evidence accumulation, and reporting.

3.7.3 Reusability

We discussed previously the notion that reusability of processes, mechanics, methods, and concepts is critical toward developing an integrated, jointly designed field that is effective across the goals it set out to achieve. The more designers can reuse, the less time spent on reinventing, the more time spent on improvement, and the quicker capabilities accumulate and mature in terms of quality, versatility, and validity. An important dimension of reusability is the level at which this is pursued. For example, reusability can be pursued at a very high level by reusing learning progressions or reusing general analysis tools. As a result, for every reuse, a lot of specifications have to be established. On the other hand, very specific interactions can be reused, such as particular character movements, functioning of interactive features such as levers and buttons, or specific scoring mechanics. As a result, many separate reused components have to be assembled and tested for interaction for every reuse.

The aforementioned four-process architecture for assessment systems (Almond et al., 2002) is one way to efficiently organize reusability. This architecture specifies the four main processes involved in implementing assessment as well as the various components, actors, and directions of interaction that are associated with it. Together, we call that the assessment engine. The basic idea is that each of these processes and components is reusable as well as replaceable but that their scope is sufficiently defined that this can be done relatively independently. That is, the aim of the architecture is to find an appropriate middle ground between fine-grained and general reusability. At some point, type of technology (soft- and hardware) as well as basic structure rules (e.g., data formats, how components communicate) need to be decided on as well, and while this might affect reusability at some level, the point of the architecture is to specify the messages that need to occur between processes, rather than a specific technology.

In practice, underlying reusable configurations of the work product, the approach to evidence identification, and psychometric model fragments can connect game design and psychometric modeling at the level of an "evidentiary skeleton" that game designers can clothe with various surface features. At the same time, they can be assured that to a first approximation, usable and pertinent evidence will be obtained (Mislevy, Steinberg, Breyer, Johnson, & Almond, 2002) and that assessment goals can be met. As noted before, in individual assessments, security is a prime concern. In GBAs, this concern is increased because memorable and identifiable narratives, characters, and mechanics make it easier to transfer details of the assessment to future test takers, the fear being that the assessment becomes a memory test. Reusability of underlying mechanics is critical in order to manage larger pools of assessment assets that can be strategically deployed to manage exposure.

3.7.4 Analysis

A desire of principled design is the notion that opportunities to collect specific evidence are designed into the activity up front and that an analysis model can be established up front. This is certainly the expectation and practice for established assessments. It is also practical and efficient when the constructs of interest are relatively well understood, their measurement is well understood, and a well-controlled measurement that only yields modest amounts of highly targeted evidence can be organized. Naturally, when an assessment attempts to measure constructs that are less removed from the complexities of real worlds, achieving a similar level of understanding and control takes a lot more effort. Subsequently, an iterative process is established between learning the intricacies of the student model and applying the student model in a confirmatory fashion. In addition, the vast freedom of movement that is provided in resultant relatively non-restricted spaces allows for the recording of many actions. Yet, whether all those actions are pertinent evidence is not necessarily clear, and valuation of that evidence becomes a primary challenge—a challenge that was self-inflicted in an attempt to better understand the world around us. Because the action space is less constrained and what players will do cannot be fully predicted, game designers play-test prototypes and "game slices" early and often and feed what they learn back into the design.

Returning to principled design, a summative assessment does still require a relatively high level of confidence in evidence, and therefore, a primary consideration for the development of summative game-based assessments has to be a bootstrap strategy, in which better controlled measurements are used to triangulate into less controlled environments. Mislevy et al.'s (2014) first two paradigms for assessment would describe a starting point as evidence characterized by the third paradigm developed, eventually letting go of any external evidence sources. Following an earlier argument made about game-based assessments being particularly apt to measure more complex skills and practices, at some point, there would logically not be any triangulating evidence from more traditional assessments, although there could still be some coalescence around foundational skills. At that point, validity evidence could be gleaned from parallel and/or multiple measurements in which the question of context dependence becomes really interesting. More specifically, to what extent the context is a unique component of the construct rather than one of many possible contexts is an important consideration in terms of the ability of the assessment developer to create parallel measurements.

Some tension might arise between principled design and approaches that are sometimes employed in analyzing the seemingly troves of data representing many test taker behaviors or actions under the umbrella term educational data mining (Baker, 2010). In particular, fishing expeditions in this digital ocean (DiCerbo & Behrens, 2012) are prone to capitalization on chance and not able to withstand cross-validation. On the other end of the spectrum, statistically powerful models exist that are highly restrictive in terms of the expected characteristics of the data in terms of distributional requirements and the relationships among different variables.

The understanding of summative GBAs or even just scenario-based assessments is not at a level that such models are useful, and it is well possible that an altogether different paradigm is required. The ideal is an interplay between design and discovery, toward the end of a rich but comprehensible performance space that produces construct-relevant evidence (Mislevy, Behrens, DiCerbo, & Levy, 2012). Although there is much we have yet to learn about GBA, one clear lesson is that creating a rich environment, collecting massive data from whatever actions are produced, and hoping that psychometricians will somehow be able to score it is not an effective way to develop them.

We would argue that in the end, it is all about the level of claims that is desired. In a summative GBA environment, there are likely some claims that are close to observed behaviors (e.g., the number of times a quest was completed) and some that are at higher inferential levels (e.g., the level of functioning of the test taker on a proportional reasoning learning progression) supported by multiple layers of evidence. As such, the statistical framework should be able to address multiple levels of inference either in a single comprehensive model or in multiple parallel models. Naturally, a single model is preferable in order to make consistent connections between the various inferences. Whatever statistical models are used at a lower level of inference (e.g., dendrograms and cluster models, classification and regression trees, Markov decision process modeling) or a high level of inference (e.g., item response theory models, generalized diagnostic models, structural equation models), an overarching probabilistic framework should be employed that connects the various (latent) constructs of interest. Fairly direct approaches that provide a natural language to both describe and statistically model evidence models are Bayes' Nets or directed acyclic graphs (Pearl, 2000). More sophisticated models, such as generalized mixed models (Moustaki & Knott, 2000), make use of one or more link functions to link observable variables to latent variables. Under both these types of overarching framework models, a wide range of specific inferential models can be employed. Note that this is also consistent with the idea of reusability and the four-process architecture introducing and replacing modules relatively independently but with adequate understanding of the full assessment system.

One component that we have only briefly discussed above under the four-process architecture for assessment systems is scoring. The trade-off that surfaces is about what type of inferences is made and under what process (e.g., evidence identification, evidence accumulation). In scoring, some form of raw telemetry is converted to a variable that represents some meaning about correctness, incorrectness, and sometimes partial correctness. This is accomplished through logical rules that are being applied (e.g., completing all three missions and taking less than 15 turns is "correct," completing only one mission or taking more than 30 turns is "incorrect," everything else is "partially correct") or through some inference that is either executed by human raters based on some rubric or left to an automated scoring engine that is trained and, in some cases, based on machine learning techniques. However, even if those inferences are not explicitly and/or statistically modeled (such as might be the case with human raters), there are still inferential models underlying these variables, and the term "observable variable" should not be used lightly if some interpretation or scoring has already been applied. More scoring up front puts more burden on the interpretation of the raw data, while less scoring places that burden on the sophistication of the statistical modeling. Access to the lower-level data is critical to bootstrap our way to improvements in assessment development and scoring.

3.7.5 Empirical Approach

Finally, the confluence of traditionally separate types of expertise and the relatively primal development level of this field means that each development attempt will require significant research alongside it, even if reusability is employed to the fullest extent. This, in turn, will require that much data be collected. The fields of assessment and game design are quite complementary in that sense, and combining the two should provide a reasonable basis for learning. In assessment, empirical research is highly methodical and controlled, focusing on field tests with large samples in precisely targeted populations commensurate the claims that are intended to be made and before the assessment is used operationally. Given the stakes, an operational assessment is not something to tinker with, but the investment in any single item could be relatively small. Therefore, development is to some extent a matter of the creation of many candidate items, several iterations of expert review, the selection of items based on a relatively large data collection from a pilot study or field test, and assembly of those items into operational test forms. In game design, early on a lot of continuous testing with small samples is conducted (i.e., "play-testing"), testing out many design choices and working with dozens of versions of very early vertical slices (i.e., a playable version that addresses most or all core mechanics once) of the game-to-be (i.e., "preproduction"). The investment in the final game mechanics, including art and coding all the interactions (i.e., "production"), is relatively large and costly to reverse. Large-scale testing does happen in a beta phase, when the game is largely out and distributed to relatively large groups. It is perfectly acceptable in software development to still make a substantial number of tuning changes to a beta version before general access/release, although the core narrative and mechanics are cast at that point.

Given that the investment in a scenario-based assessment is more akin to a game than one or several assessment items in terms of investment and given that much has to be learned during development, employing extensive play-testing and small-scale qualitative data collections (e.g., cognitive labs, think aloud protocols, eye tracking) is critical. Subsequently, given the summative nature of the claims that are intended to be made (including scores that have to hold for some period of time), a significant field test or pilot assessment before operational use in the appropriate target population will also be required. While such data collections are difficult and expensive, getting the GBA wrong is likely more expensive than in the game situation. Beta testing is a very convenient way of large-scale testing: you test the game automatically in your target audience, and you receive feedback directly from your future customers, while you can market the game itself. Unfortunately, this kind of testing will not work in cases in which security is at stake or, as mentioned above, comparable scores have to hold for some period of time and, therefore, is not appropriate for summative GBAs. As a result, high-stakes summative GBAs will likely be more focused on relatively well understood, often simpler, mechanics, while formative and learning applications can be in a far more developmental mode.

3.8 Summary

Trailing game-based learning, game-based assessment (GBA) has quickly grown from a grant-supported research niche for formative assessment to a venture-backed start-up industry, initially taking hold in the personnel selection markets. While there are still many challenges to overcome in terms of generalizability, fairness, and scalability, for those use cases where an outsize interest is afforded to engagement (e.g., low-stakes drop-from-the-sky assessments) and context (e.g., individual licensing or certification exams), GBAs provide a compelling paradigm for fulfilling those interests. In addition, for use cases where the construct of interest is inherently complex and situated, digital environments with higher degrees of interactivity may prove to be the most fair, valid, and scalable way to assessment. This includes constructs often associated with a new economy, such as collaborative problem-solving, cross-cultural competence, and global citizenship. In contrast, the case for GBAs in an individual achievement assessment (e.g., admissions) is rather poor due to the critical levels of security, generalizability based on an abstract level of inference, and repeatability placed upon those assessments. Finding a balance for when GBAs make sense is challenging as society's desire for different types of information about human knowledge, skills, and abilities shifts constantly.

In this chapter, we have discussed a number of perspectives, motivations, tensions, and design considerations for summative GBAs and attempted to answer questions related to what the core mechanics of games are and how they could relate to summative assessments. We discussed how games are emergent systems that require meaningful interaction and are governed by rules. We showed how games are built for interaction and engagement and provide complex, contextualized environments. Therefore, games may be particularly compatible with assessment when motivation is at stake or more complex skills and cognitive processes are of interest. Subsequently, we discussed how design trade-offs are at play when developing GBAs. Foremost, competing goals exist in terms of the need for standardization, generalizability, and fairness in assessment versus the need for context, compelling narratives, and freedom to explore in games. In addition, assessments typically serve broad audiences, whereas games focus on specific market segments, and lastly, the development speeds in both fields are typically very different. There are, however, many commonalities and respective strengths across the assessment and game fields that are important to draw from when designing GBAs. Foremost, it is important to use a principled design framework to appropriately develop a GBA based on strong connections between game mechanics and assessment goals. Principles of good analysis consistent with the design, the reuse of system components, and strong reliance on frequent, iterative empiricism are also important to build robust GBAs.

Whether the case for summative GBAs is compelling enough is in large part a question of whether the required investment, which tends to be sizable, is worth it. Are the types of claims that can be made based on GBAs but not in other ways valuable enough (i.e., in terms of warrants or certifications that institutions or employers are willing to associate with this) that test takers are willing or compelled to pay for it? For those situations in which GBAs do not necessarily provide unique measurement claims, could they still provide acceptable and fair alternative assessment modes that institutions and employers are willing to put warrants against and individual test takers are willing to pay a premium for because they provide a better or fairer opportunity to show proficiency for that test taker? It appears that there is already some empirical evidence that employers see GBA as a worthwhile tool to aid in their recruitment efforts (e.g., Tsang, 2017). At the same time, efficacy research and evidence for GBAs is still exceedingly sparse. It appears that, first, psychometric and cognitive research is needed to better define and validate the mechanics (i.e., design), evidence, and claims that are intended to be brought together. Second, this appears to be a particularly appropriate area for partnerships as well-developed traditionally disjointed fields are brought together, and therefore, developing and cultivating those partnerships is equally important.

Despite attempts to provide a comprehensive perspective on summative GBAs, there are several topics that we have not attended to satisfactory and that deserve more detailed consideration. First, the topic of engagement and motivation in GBAs is unchartered and untamed territory. We know that there are mechanics that increase engagement and motivation for specific learners, and we usually strive to improve intrinsic motivation for learning. However, relatively little is known about the transfer between extrinsic and intrinsic motivators, how engagement mechanics in typical commercial games transfer to learning or assessment environments, and how different levels of engagement relate to learning and assessment outcomes. Second, studying and showing validity and efficacy evidence remains a relatively weak area in GBA, which is in part due to the level of maturity of the applications and companies involved (Carolan & Zielezinski, 2019). Third, following the notion that assessment and game mechanics need to be aligned, there is a lot to be explored about different game types and genres and how they relate to different assessment goals. Fourth, we have been relatively mute about adaptive and personalized assessment (and learning). In many ways, games are the ultimate personalized environment at scale, using levels, power-ups, and unlocking of features and areas to confront players with increasingly complex and challenging tasks while also ensuring success for beginners. The goals for summative adaptive assessment are quite different and mostly related to efficiency. However, it would be important to better understand where the two are compatible. Finally, we touched on accessibility briefly, doing this critical topic grave injustice. The rich, visually stunning, and highly interactive and responsive environments that are typically associated with games are difficult to reconcile with the goal of equitable assessments, including for people who are blind or have low vision. That is not to say that GBAs should be nonvisual only, but we should strive to develop alternative modes and assistive technologies that can meet similar goals of engagement and measurement.

Acknowledgements On last count, no fewer than seven anonymous reviewers at ETS and abroad have provided countless comments and edits on previous versions of this work. We are deeply indebted to their valuable and perspective-altering contributions without which this work would never have seen the light of day.

References

Almond, R., Steinberg, L., & Mislevy, R. S. (2002). Enhancing the design and delivery of assessment systems: A four process architecture. *The Journal of Technology and Assessment, 5*(1), 1–63.

Attali, Y., & Powers, D. (2008). *Effect of immediate feedback and revision on psychometric properties of open-ended GRE subject test items (Research Report 08-21)*. Princeton, NJ: Educational Testing Service.

Baker, R. S. J. (2010). Data mining for education. In B. McGaw, P. Peterson, & E. Baker (Eds.), *International encyclopedia of education* (3rd ed.). Oxford, UK: Elsevier.

Barab, S. A., Gresalfi, M. S., & Ingram-Goble, A. (2010). Transformational play: Using games to position person, content, and context. *Educational Researcher, 39*(7), 525–536.

Bauer, M., Wylie, C., Jackson, T., Mislevy, R., Hoffman-John, E., & John, M. (2017). Why video games can be a good fit for formative assessment. *Journal of Applied Testing Technology, 18*(S1), 19–31.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement, 8*(2–3), 70–91.

Carolan, J., & Zielezinski, M. B. (2019). *Debunking the 'gold standard' myths in edtech efficacy.* Retrieved from: https://www.edsurge.com/news/2019-05-21-debunking-the-gold-standard-myths-in-edtech-efficacy

Csíkszentmihályi, M. (1975). *Beyond boredom and anxiety*. San Francisco, CA: Jossey-Bass.

Csíkszentmihályi, M. (1990). *Flow: The psychology of optimal experience*. New York, NY: Harper Collins.

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*, 227–268.

DiCerbo, K. E., & Behrens, J. T. (2012). Implications of the Digital Ocean on current and future assessment. In R. Lissitz (Ed.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 273–306). Charlotte, NC: Information Age Publishing.

Dorans, N. J. (2012). The contestant perspective on taking tests: Emanations from the statue within. *Educational Measurement: Issues and Practice, 31*(4), 20–37.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*, 289–303.

Electronic Arts. (2018). *The Sims 4 [computer software]*. Redwood City, CA: Electronic Arts.

Ercikan, K., & Pellegrino, J. W. (2017). *Validation of score meaning for the next generation of assessments: The use of response processes* (The NCME applications of educational assessment and measurement book series). Washington, DC: National Council on Measurement in Education.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27–32.

Gee, J. P. (2007). *Good video games + good learning*. New York, NY: Peter Lang.

Greeno, J. G. (1998). The situativity of knowing, learning, and research. *American Psychologist, 53*, 5–26.

Holland, P. W. (1994). Measurements or contests? Comments on Zwick, bond and Allen/Donoghue. In *Proceedings of the social statistics section of the American Statistical Association* (pp. 27–29). Alexandria, VA: American Statistical Association.

John, M. (2014). *Gamification is dead, long live games for learning.* Retrieved from http://techcrunch.com/2014/10/05/gamification-is-dead-long-live-games-for-learning/

Kerr, D., & Chung, G. K. W. K. (2012). Identifying key features of student performance in educational video games and simulations through cluster analysis. *Journal of Educational Data Mining, 4*(1), 144–182.

Klopfer, E., Osterweil, S., & Salen, K. (2009). *Moving learning games forward: Obstacles, opportunities & openness.* Retrieved from http://education.mit.edu/papers/MovingLearningGamesForward_EdArcade.pdf

Koster, R. (2014). *A theory of fun for game design*. Sebastopol, CA: O'Reilly Media.

Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment for education. Theory and applications*. Cambridge, MA: Cambridge University Press.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Research, 23*(2), 13–23.

Metcalf, S. J., Kamarainen, A., Tutwiler, M. S., Grotzer, T. A., & Dede, C. J. (2011). Ecosystem science learning via multi-user virtual environments. *International Journal of Gaming and Computer-Mediated Simulations., 3*(1), 86–90.

Mislevy, R. J., Behrens, J. T., DiCerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence centered design, psychometrics, and data mining. *Journal of Educational Data Mining, 4*, 11–48. Retrieved from http://www.educationaldatamining.org/JEDM/images/articles/vol4/issue1/MislevyEtAlVol4Issue1P11_48.pdf

Mislevy, R. J., Corrigan, S., Oranje, A., DiCerbo, K. E., Bauer, M. I., von Davier, A., & John, M. (2016). Psychometrics and game-based assessment. In F. Drasgow (Ed.), *Technology and testing* (pp. 23–48). New York, NY: Routledge.

Mislevy, R. J., Haertel, G., Cheng, B. H., Ructtinger, L., DeBarger, A., Murray, E., … Vendlinski, T. (2013). A "conditional" sense of fairness in assessment. *Educational Research and Evaluation, 19*, 121–140.

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A., Hao, J., Corrigan, S., … John, M. (2014). *Psychometric considerations in game-based assessment*. Redwood City, CA: GlassLab.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–67.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Johnson, L., & Almond, R. A. (2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15*, 363–378.

Moustaki, I., & Knott, M. (2000). Generalised latent trait models. *Psychometrika, 65*, 391–411.

National Assessment Governing Board. (2014). *Science framework for the 2015 National Assessment of educational Progress*. Washington, DC: National Assessment Governing Board.

Partnership for 21st Century Skills. (2011). *Framework for 21st century learning*. Retrieved from P21.org

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Ramanarayanan, V., Evanini, K., & Tsuprun, E. (2019). Beyond monologues: Automated processing of conversational speech. In K. Zechner & K. Evanini (Eds.), *Handbook of automated speech scoring*. London, UK: Routledge.

Rideout, V. (2014). *Learning at home: Families' educational media use in America*. New York, NY: The Joan Ganz Cooney Center.

Salen, K., & Zimmerman, E. (2003). *Rules of play: Game design fundamentals*. Cambridge, MA: MIT Press.

Schell, J. (2008). *The art of game design*. Boca Raton, FL: CRC Francis & Taylor.

Schmit, M. J., & Ryan, A. (1992). Test-taking dispositions: A missing link? *Journal of Applied Psychology, 77*, 629–637.

Scott, L. A. (2017). 21st century skills early learning. *Partnership for 21st Century Learning.* Retrieved from http://www.p21.org/storage/documents/EarlyLearning_Framework/P21_ELF_Framework_Final.pdf

Shaffer, D. W. (2006). *How computer games help children learn*. New York, NY: Palgrave Macmillan.

Shute, V. J., Ventura, M., & Kim, Y.-J. (2013). Assessment and learning of qualitative physics in Newton's playground. *The Journal of Educational Research, 106*, 423–430.

Stecher, B. M., & Hamilton, L. S. (2014). *Measuring hard-to-measure student competencies: A research and development plan*. Santa Monica, CA: RAND Corporation.

The Entertainment Software Association. (2013). *Essential facts about the computer and video game industry*. Washington, DC: The Entertainment Software Association.

Tsang, A. (2017). Want to work for Jaguar Land Rover? Start playing phone games. *The New York Times.* Retrieved from https://www.nytimes.com/2017/06/19/business/jaguar-land-rover-app-puzzles.html

U.S. Army (2012). http://www.military.com/Entertainment/Games/

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361–375.

Chapter 4
Stealth Assessment Embedded in Game-Based Learning to Measure Soft Skills: A Critical Review

Xinyue Ren

4.1 Introduction

With the development of technologies and popularization of digital devices, learning styles and delivery models have changed accordingly. The ideas of multimedia learning, computer-assisted learning, and technology-based learning have been increasingly adopted by instructors to enhance their students' learning attitudes, experiences, and outcomes (Hwa, 2018; Lock, Kim, Koh, & Wilcox, 2018). Among a variety of e-learning products, digital/video games are regarded as one effective digital tool to enhance active and engaging learning experience.

The idea of edutainment also incorporates the dual aspects of video games in performing both educational and entertainment purposes, such as serious games (All, Patricia Nunez Castellar, & Van Looy, 2016; Caballero-Hernandez, Palomo-Duarte, & Dodero, 2017; Hwa, 2018). Under appropriate instructional design, these games can be viewed as an assistive learning tool to complement the weaknesses of the formal schooling system (Borji & Khaldi, 2014). According to previous research, digital games have been widely applied in various teaching contexts to successfully increase students' learning motivations and promote their academic achievements (All, Patricia Nunez Castellar, & Van Looy, 2015; Hwa, 2018).

Accordingly, game-based assessments (GBA) can be embedded in digital games to evaluate learners' learning outcomes, such as soft skills. However, limited research has focused on how to effectively design and implement game-based assessments in virtual learning environments, such as stealth assessment, and few studies have addressed how to interpret the data collected from stealth assessment to support student learning (Ke & Shute, 2015). Therefore, the chapter aims to contribute to the knowledge in the field of GBA and lay a foundation for the future research. The author will first review how stealth assessment has been used to

X. Ren (✉)
Ohio University, Athens, OH, USA
e-mail: xr868414@ohio.edu

D. Ifenthaler, Y. J. Kim (eds.), *Game-Based Assessment Revisited*, Advances in Game-Based Learning, https://doi.org/10.1007/978-3-030-15569-8_4

measure soft skills in previous studies and then discuss and analyze effective practices or procedures of stealth assessment embedment for future reference. The chapter will be divided into five parts: (1) the affordances of games on learning experiences and outcomes, (2) what is learning assessment and what is GBA, (3) how stealth assessment was designed and implemented to measure soft skills, (4) critical analysis of effective strategies and procedures, and (5) recommendations for the future GBA practices.

4.2 The Affordances of GBL

Various features of digital games allow learners to develop active and engaging learning experiences. First, GBL can potentially promote active participation among learners (Hwa, 2018). Many studies showed that the degree of control in learning may influence students' perceptions, motivations, and learning outcomes (Snow, Allen, Jacovina, & McNamara, 2014). In GBL environments, learners can enjoy more autonomy and freedom than in a conventional classroom. This means that games become an agency in which students can self-regulate their learning plans, make learning decisions, and actively engage in various learning activities (All et al., 2015; Snow et al., 2014). As a result, when learners can enjoy more autonomy to control their learning process, they are more likely to feel motivated and produce positive attitudes.

Second, a game-based environment can produce an engaging learning experience for learners (All et al., 2015; Petrovica & Anohina-Naumeca, 2017; Shute & Ventura, 2013). Emotions and attitudes often influence how learners perceive and behave (Zull, 2002). Positive attitudes can potentially increase students' motivation to learn new information. While playing video games, learners are more likely to immerse themselves in gaming environments to produce a flow state (Kiili, 2005). The flow state refers to a period of time in which learners completely engage in the GBL activities (Kiili, 2005). Other game features, such as simulation and role-playing, in digital games can also promote engaging and authentic learning experiences.

GBL can also produce customized learning experiences for learners. Many features and elements in digital games include characters/avatars, narratives, languages, images, and sounds. Usually, users are allowed to make their own choices in playing the games. In GBL environments, learners can make decisions about whether to engage in a learning activity or replay a game (Snow et al., 2014). However, many studies also showed the limitations of user control in producing positive learning outcomes. For instance, because of different capabilities and preferences among users, many learners may not be comfortable with making their own learning decisions while learning in GBL environments (Snow et al., 2014). Learners may lose their motivation or lack appropriate guidance in personalized learning experiences (Tsai, Tsai, & Lin, 2015).

Well-designed digital games can provide students opportunities to learn by doing (Hwa, 2018; Lock et al., 2018). Zull (2002) believed that the learning cycle includes abstract hypothesis, active testing, concrete experience, and reflective observation. Similarly, Dewey (1910) also proposed five steps to complete a thinking process: containing problem discovering, observing the problem, coming up with solutions, identifying the reasoning behind the solutions, and testing and verifying the solutions. For instance, in order to develop critical thinking skills, experiments are needed to provide learners opportunities to become knowledge producers. In terms of the safe environment provided in digital games, students can freely conduct experiments and test their ideas (Perini, Luglietti, Margoudi, Oliveira, & Taisch, 2018). Dewey stated that experiencing is prior to knowing (Hickman, 1990). Therefore, various hands-on activities can produce experiential learning experiences, which can effectively support students' learning outcomes (Brown & Green, 2016; Lock et al., 2018).

Well-designed digital games may also promote the constructivist learning experience (Flynn, Bacon, & Dastbaz, 2010). Gaming environments are often viewed as comfortable and welcome places in which learners can interact with others. They are welcomed to express themselves and contribute their experiences to the gaming community. For example, learners may produce knowledge according to their individual cultural backgrounds and life experiences. Learners are able to control and create their own stories in the virtual environment. Therefore, in digital games, especially multiplayer games, rather than emphasizing producing the standardized learning outcomes, learners are able to create diverse learning environments and experience social and peer learning (Snead, 2014).

In conclusion, various benefits of digital games give learners opportunities to experience active and engaging learning experiences. However, it is important to realize the limitations of digital games in enhancing students' learning experiences. Many elements may have an impact on students' GBL experience, such as their gameplaying experience, perceptions, learning styles, motivations, and quality of digital games. Hence, learning experiences may vary according to learners' characteristics and experiences (Mayer, Warmelink, & Bekebrede, 2013).

4.3 What Is GBA?

In formal educational institutions, evaluation methods, such as standardized or high-stake tests, are often used to measure students' learning progress and outcomes. However, these measurements are problematic in their indication of learners' authentic academic achievements. They are usually used to provide final judgments, rather than to support students during their learning (Shute & Ventura, 2013). These evaluation approaches are limited in assessing a higher level of thinking skills and complex competencies. As a result, these traditional evaluations often fail to provide effective feedback during the learning and producing of valid and reliable findings in order to inform instructors of how to enhance students' learning experiences and outcomes (De Klerk & Kato, 2017). Therefore, in order to engage and empower learners, there is a need to consider designing and developing other alternative assessments.

4.3.1 Assessment

Assessment plays an important role in supporting student learning in formal education (Lock et al., 2018). Assessment often involves data collection, analysis, and interpretation (Shute & Ventura, 2013). Usually, for pedagogical purposes, assessments include formative, summative, and confirmative evaluations, and each is used at different stages for various purposes (Caballero-Hernandez et al., 2017; Morrison, Ross, Kemp, & Kalman, 2010). For instance, formative assessment refers to “the quality control of the development process” (Morrison et al., 2010, p. 254). This type of evaluation is often used during the curriculum development, with aims of identifying the effectiveness of learning materials and learners’ capabilities and providing timely feedback to support student learning. Feedback from formative evaluation can be used as evidence to revise the learning materials or objectives (Faizan, Löffler, Heininger, Utesch, & Krcmar, 2019). Formative assessment also indicates the benefits to better accommodate the needs of learners to enhance their learning outcomes (Morrison et al., 2010; Shute & Ventura, 2013).

Summative assessment refers to the method used to measure the final learning outcomes of a module or a course. It aims to measure several aspects, including the efficiency of learning materials, the cost of course development, and the learners’ perceptions (Morrison et al., 2010). Results from summative assessment can indicate whether the expected learning objectives are achieved.

Confirmative assessment is also known as a follow-up or future-oriented evaluation with an aim to measure the effectiveness of a course over time or the continuity of learners’ performance (Dessinger & Moseley, 2003; Morrison et al., 2010). It can be divided into two types: learner and context oriented. Because of the changes in learners’ characteristics and learning contexts over time, the success of courses may not remain the same in a long run. In this case, a confirmative evaluation is necessary to obtain follow-up data to help instructors further understand the quality of their curriculum design. They may perform continuous revisions accordingly to sustain the effectiveness of their courses.

In conclusion, different types of assessment methods address various features and purposes. Instructors may choose the assessment methods according to their specific contexts and objectives. Most importantly, when designing and developing assessments, instructors should ensure the alignment between predefined learning objectives and tasks (De Klerk & Kato, 2017; Lock et al., 2018; Morrison et al., 2010).

4.3.2 Various Categories of GBA

When appropriately integrated and guided, well-designed games may serve as a pedagogical platform to motivate learners in various learning settings. Evaluation approaches, mainly external evaluation instruments, can be applied to assess the learned knowledge before, during, or after playing the games (Caballero-Hernandez et al., 2017; Petrovica & Anohina-Naumeca, 2017; Shute & Ventura, 2013). However, many researchers realized the limitations of relying on these approaches to produce authentic knowledge assessment results. They believed that the potentials of games make them become suitable tools to perform educational and assessment purposes simultaneously (Kim & Shute, 2015; Petrovica & Anohina-Naumeca, 2017).

According to various evaluation purposes, assessments in gameplaying environments can be categorized into several types (Caballero-Hernandez et al., 2017). For example, some assessment methods can be used to identify learners' prior knowledge, such as diagnostic tasks, or to provide feedback on learners' self-paced learning skills, such as integrative tasks. According to different implementation strategies, assessments can rely on game scoring and in-game or out-game instruments. For assessment integration, methods may include monitoring accomplished levels, adding missions and quests, and comparing learners' game performance with assessment models, quizzes, and peer evaluation. Other types of assessments are developed based on different stages. For instance, assessments can be used to analyze learners' performance in process or when they reach the game goals. Teachers may observe learners in gameplaying conditions to produce evaluation results. Therefore, instructors may consider their specific teaching contexts and learners' experiences while making decisions on assessment methods and integration strategies in a GBL environment.

4.3.3 Stealth Assessment

In order to validly measure soft skills, it is important to rely on performance-based data (Shute & Ventura, 2013). Authentic performance can indicate the transferability of these competencies in a real-world situation, which can produce reliable feedback and findings (All et al., 2015). However, it is often challenging to measure competencies, especially soft skills, in a realistic situation, or accurately interpret the data. In formal learning settings, instructors may face difficulties creating situated scenarios to measure learners' performances. Owing to the elements of digital games, such as simulation and role-playing, instructors may rely on GBA to better evaluate and infer learners' performances in virtual environments (Faizan et al., 2019).

Generally speaking, GBA includes internal and external assessments (Caballero-Hernandez et al., 2017; Petrovica & Anohina-Naumeca, 2017). Internal assessment, also known as stealth assessment, refers to an embedded assessment method to measure students' performances in gameplaying environments. Stealth assessment is an invisible and ubiquitous evaluation method, aiming to support student learning. Without disruption, learners can maintain their flow state (De Klerk & Kato, 2017; Petrovica & Anohina-Naumeca, 2017; Shute & Ventura, 2013).

Gameplaying can potentially indicate users' actual behaviors, skills, and competencies when facing problems (Shute, Wang, Greiff, Zhao, & Moore, 2016). While being evaluated by stealth assessment, learners can reduce their test anxiety and perform in an authentic manner (de-Juan-Ripoll et al., 2018; DeRosier & Thomas, 2018). Their performances, interactions, and solutions will become data and evidence to indicate their authentic competencies when completing various tasks to reach the final goals in games (De Klerk & Kato, 2017; Shute et al., 2016; Shute & Ventura, 2013; Snow et al., 2014).

Hence, this kind of assessment can be embedded in GBL environments to validly and reliably monitor learners' behaviors, capabilities, and outcomes (Shute et al., 2016; Shute & Ventura, 2013). The assessment findings can potentially inform instructors to make decisions concerning the improvement of learning materials or the revision of learning objectives to enhance student learning.

4.3.4 Benefits of GBA

In terms of the benefits of GBL in promoting active and engaging learning experiences, it is reasonable to discuss the use of GBL environments to complement the restrictions of traditional formal learning settings. Similarly, GBA demonstrates the possibility of evaluating skills in a realistic situation to produce valid and reliable results, which is often impractical through conventional assessment methods (De Klerk & Kato, 2017). Therefore, GBA presents the possibility of supplementing the weaknesses of standardized and high-stake tests.

GBA is increasingly viewed as a beneficial tool, which can bring many advantages. First, GBA can be used to evaluate many competencies which cannot be appropriately tested in paper-based assessment methods (DiCerbo, 2017). In digital games, learners are required to complete a series of tasks to achieve the goals. With the development of technologies, well-designed digital games can provide learners with interactive and simulation-based learning experiences (Perini et al., 2018). In this way, learners are able to react, make decisions, or take actions according to the realistic settings they are facing (De Klerk & Kato, 2017; DiCerbo, 2017). Their reactions and actions will be recorded and assessed to indicate their capabilities and valued competencies.

Second, GBA can provide positive testing experiences for learners (DiCerbo, 2017). Many conventional assessment methods negatively influence students' learning experiences, such as test anxiety (De Klerk & Kato, 2017; Petrovica & Anohina-Naumeca, 2017). The unpleased experiences may have a negative impact on their test performance; as a result, these traditional tests cannot be used to indicate reliable and valid outcomes. Alternatively, while playing games, GBA can potentially maintain learners' flow state to reduce their test anxiety. Moreover, without external interruption, learners are able to perform in an authentic manner, and assessments may produce accurate and valid results (Kim, Almond, & Shute, 2016).

Moreover, GBA embedded in the game-based environment may record and collect small and big data to increase the validity of measurement results. GBL environment can be used to track and capture learners' actions and decisions (Alcañiz, Parra, & Giglioli, 2018; DeRosier & Thomas, 2018). In GBL environments, learners are able to perform the larger number of tasks than that in traditional learning settings (De Klerk & Kato, 2017). Various performances can produce large numbers of data to increase the trustworthiness of assessments. For instance, while playing games, learners can produce large quantities of process data, such as their log data, clicks, moves, and other behaviors that are unobservable through conventional methods (DeRosier & Thomas, 2018; Faizan et al., 2019; Ke & Shute, 2015; Snow, Likens, Allen, & McNamara, 2016). This information can be recorded stealthily and constantly to provide small or big data, and they can be further analyzed to produce numerous interesting outcomes. Meanwhile, the results may help instructors better understand how students perceive the information, how they behave to reach the learning objectives, and their personal improvements.

In conclusion, in GBL environments, GBA indicates many benefits. For instance, GBA can be embedded to evaluate many competencies which cannot be appropriately tested by conventional assessment methods. GBA can potentially increase the validity of results through analyzing authentic performances and data. However, people often face the challenges of accurately interpreting the data collected from implementing GBA and building relationships among the data. For example, the researchers need to first identify which variables are useful to indicate different capabilities (De Klerk & Kato, 2017; DiCerbo, 2017). However, multiple variables may also complicate the data analysis and interpretation. Therefore, De Klerk and Kato (2017) believe that it is often challenging for instructors and researchers to conclude the results by relying on GBA.

4.4 Stealth Assessment and Soft Skills

Many researchers believe the necessity of developing twenty-first-century competencies (Shute & Ventura, 2013). However, soft skills are often ignored in curriculum development and assessment in formal educational institutions (Lock et al., 2018). Soft skills are competencies that are valued to be successful in the twenty-first century, including critical thinking, creativity, problem-solving, time management, information literacy, team work, and others (De Klerk & Kato, 2017; Faizan et al., 2019; Lock et al., 2018; Shute et al., 2016; Shute & Ventura, 2013). Because of the limitations of curriculum and assessment in formal education, many competencies, such as problem-solving skills and creativity, cannot be appropriately and accurately measured by traditional evaluation approaches (Kim et al., 2016).

For instance, problem-solving skills cannot be effectively assessed to reflect on a real-world situation through predefined conventional assessment methods (Shute & Ventura, 2013). Creativity cannot be directly tested in paper-based assessments. Instead, these competencies should be divided into small skillsets or psychometric components. Researchers may analyze these variables to produce valid inferences about students' knowledge and competencies. Under various circumstances, digital games can be viewed as an alternative platform to provide simulated environments to develop and evaluate these soft skills in real time (Faizan et al., 2019).

In order to tell whether learners are able to successfully gain these competencies through various learning activities online, there is a need to develop appropriate assessment methods to support their learning. In terms of the benefits of formative assessment in providing feedback during the learning process to enhance learning outcomes, the author will mainly focus on stealth assessment, one type of formative assessment in GBL environments, in the following section. Through investigating previous studies, the author will further discuss GBA design and implementation strategies, the advantages of stealth assessment in measuring soft skills, stealth assessment implementation, and analysis.

4.4.1 GBA Design Strategies

The evidence-centered design (ECD) is a framework that can be applied to produce valid and evidentiary arguments in GBA (De Klerk & Kato, 2017; DiCerbo, 2017; Kim et al., 2016; Shute & Ventura, 2013). The framework contains several models, including student/competency model, task model, evidence model (demonstrated by scoring model and measurement model), and assembly model (De Klerk & Kato, 2017; DiCerbo, 2017; Kim et al., 2016; Shute & Ventura, 2013). Each model deals with a specific purpose. For instance, the competency model focuses on the variables that the researchers aim to measure, and the task model may be used to design in-game tasks to measure the defined variables (DeRosier & Thomas, 2018; Kim et al., 2016). Usually, the design process involves several iterative cycles, which include hypothesis, prototype, test, and revise (DiCerbo, 2017). Different models interrelate with each other to build solid assessment arguments.

For example, DiCerbo (2017) introduced a GBA approach to enhance third grade students' understanding of geometric measurement. First, the researchers investigated relative research and standards to define what knowledge should be contained to meet the reasonable learning objectives. Next, they understood learners' prior knowledge in the field of geometric measurement. According to the learning objectives and learners' knowledge, they designed four aligned tasks at different levels and asked the participants to test these tasks. After each iteration, the evidence model was used to indicate participants' performances and the effectiveness of prototypes. As a result, assessment developers can redesign the tasks after several iterative tests and analyses, and the redesigned tasks may better align assessment goals (focal competencies) and learners' performances (Shute & Ventura, 2013).

Kim et al. (2016) further described the procedures they used to develop GBA based on Bayes nets. The procedures can be meaningful for future reference, which include (1) identifying variables and their relationships to align with the assessment purposes, (2) building the structure of network containing variables and observables, (3) distinguishing the level of each indicator from others, (4) deciding appropriate values or levels for each parameter, (5) taking the difficulty levels of games into consideration when assigning values to variables, (6) calculating and scoring variables, (7) calibrating the conditional probability tables (CPTs) and improving the Bayes network based on the pilot study, and (8) reviewing and solving unexpected problems based on collected evidence to increase the validity of assessment.

Moreover, DiCerbo (2017) states that choice is an important component in digital GBA, which can be used to distinguish games from other multimedia learning resources, such as tutorials. On the one hand, assessment developers should assume that learners may fail when performing GBA. On the other hand, they need to integrate easier options to maintain their engaging experiences. Therefore, designers should develop and balance different levels of choice to accommodate various capabilities of learners.

All in all, the design of GBA is a complicated and time-consuming process. Instructors should know about how challenging it is to convert games into GBA or to embed assessments to existing games before taking GBA into consideration in their courses. Although ECD and Bayes network are not the only design strategies, they may be helpful to promote the GBA design process.

4.4.2 GBA Practices

After understanding the benefits of stealth assessment, the following section will discuss effective practices of implementing and interpreting stealth assessment. After researching more than 400 papers, Caballero-Hernandez et al. (2017) determined the most popular methods among a variety of assessment types in GBL environments. For example, according to various aims of the assessment, a majority of researchers integrated formative assessments to produce feedback to support the learning process. In terms of the assessment implementation, Caballero-Hernandez et al. (2017) found three frequently used methods, including game scoring, in-game assessment, and game scoring with external assessment. For assessment integration, monitoring of accomplished states and levels was frequently used to identify players' performances and skills (Ke & Shute, 2015). Moreover, in-process assessment was used as the main method with which to indicate players' performance during the gameplay.

Therefore, the effective practices of stealth assessment often rely on the combination of various types of assessment instruments, including formative, embedded, and in-process assessments, as well as assessment methods, including collecting data from game scoring and monitoring. Their combination may increase the validity of GBA to better measure players' competencies and identify learners' performance during the gameplay. Moreover, when implementing GBA instruments, instructors should consider students as an important part in GBL environments (Lock et al., 2018). In order to increase the effectiveness of GBA, instructors need to help students understand the affordances of digital games and the benefits of GBA. Instructors may also regularly check to ensure the alignment between assessments and learning objectives (Lock et al., 2018).

4.4.3 Stealth Assessment on Soft Skills

As mentioned before, GBA is increasingly popular in the field of education. Stealth assessment indicates its advantages in reducing the interruption of test anxiety to maintain learners' flow state. In this way, when facing problems in GBL environments, learners are more likely to conduct their performances and make their decisions in authentic manners. According to previous studies, stealth assessment was commonly applied in different subject areas to produce more reliable and valid results compared to other traditional assessment methods. In terms of the importance of soft skills in the twenty-first century, there is a need to discuss how to effectively integrate stealth assessment to measure these valued competencies. Thus, in the following section, three studies will be analyzed to address the integration of stealth assessment to measure soft skills.

First, Mayer (2018) introduced the use of a serious game to train and assess teamwork. Team performance and quality are often too complicated to be measured in a real-world scenario. In order to rely on serious games to increase the efficacy of training and assessment, Mayer (2018) believed that determining learning objectives and the validity of the serious game were necessary. Traditionally, researchers had to collect data through observation and analyze the data thorough structured psychometric approaches. Due to the advantages of digital games, stealth assessment can be used to collect and analyze data invisibly while playing the game. As a result, the data can be valid enough to serve as an indication of team performance and quality.

A multiplayer digital game, *TEAMUP*, was used for teamwork training and assessment purposes. Participants (n = 424) were asked to complete pre- and post-game surveys to establish the relationship between in-game assessments and teamwork measurements. After completing the game, game scores and in-game performances were analyzed to indicate their teamwork. Mayer (2018) conducted a regression analysis to conclude that in-game performance indicators, such as error, were significantly correlated to team performance. However, other performances, such as time and distance, were influenced by participants' age and gaming experience.

Shute et al. (2016) introduced the use of stealth assessment to measure problem-solving skills in a GBL environment. Problem-solving skills play a functional role in people's life, work, and study. However, because of the limitations in school contexts, structured curriculum and evaluation methods cannot be flexibly converted to develop and assess problem-solving skills. Therefore, researchers investigated the embedment of a digital game, *Use Your Brainz*, to measure middle school students' problem-solving skills. At first, all participants ($n = 47$) were required to play the game 3 days. After completing the game, the participants were asked to complete two problem-solving skill tests, Raven's Progressive Matrices and MicroDYN, and a demographic survey. The researchers further developed a competency model and determined in-game performance indicators and model variables to measure problem-solving skills by investigating previous literatures.

The results of a multiple regression showed that in-game performance indicators were significantly correlated with two external tests, which indicates the validity of the assessment of problem-solving skills. As a result, the researchers believed that stealth assessment is useful to indicate not only students' problem-solving skills but also different levels of each indicator. Accordingly, instructors may provide suitable remediations to support and improve their students' problem-solving skills. The researchers further mentioned that it is crucial to select appropriate external measures to align with problem-solving skills and stealth assessment.

Shute and Ventura (2013) developed a simulation-based game, called *Newton's Playground*, and embedded stealth assessment to evaluate three competencies, including creativity, conscientiousness, and physics understanding. Creativity refers to the ability to create novel, quality, and appropriate approaches to solve a task. At first, the researchers created different levels of the problems. In order to increase the validity of creativity measure, the researchers developed a competency model of creativity based on reliable creativity tests. Indicators include cognitive skills, such as fluency, flexibility, and originality, and dispositions, such as openness and risk taking. The participants ($n = 150$) were required to complete a pretest and posttest. Tutorials were provided to help students be familiar with the game. As a result, the competency estimates were correlated with the results of external measures.

Thus, stealth assessment can be effectively used to measure soft skills, such as teamwork, problem-solving skills, and creativity. However, GBA indicates its limitations in accurately measuring individual competencies. Learners' backgrounds, such as their age and gaming experiences, may influence their in-game performances, and random errors may affect the accuracy of their performance data and outcomes in gaming environments (DeRosier & Thomas, 2018). For example, experienced gameplayers may perform tasks better than those without much gaming experience. In order to produce reliable results, researchers may consider controlling these confounding variables when integrating GBA to analyze players' skills and competencies.

4.5 Critical Analysis

GBL is beneficial in providing learners with active and engaging learning experiences. The idea of assessment embedment in GBL environments is increasingly developing in the field of education. Many soft skills are highly valued in the twenty-first century; however, limited instruments can be used to measure learners' performances. Among various types of GBA, stealth assessment is viewed as an effective method to unobtrusively measure learners' behaviors, with an aim of supporting student learning. Thus, stealth assessment is suitable to measure some soft skills which are often too complicated to be accurately measured by conventional evaluation methods. Due to limited research focusing on the use of stealth assessment as well as the design and implementation strategies to measure soft skills, there is a need to review and analyze previous studies for the future reference.

Many competencies and soft skills are multidimensional concepts. It is often difficult to measure them in a traditional single-dimensional manner (Kim et al., 2016). According to previous studies, ECD and Bayes network are two important frameworks to produce valid assessment instruments in complex learning settings, such as GBL environments. Many researchers highlighted the necessity of iterative designs while developing GBA. An initial product may not be the perfect one. Continuous improvement and adjustment may better accommodate various learning types and contexts, which can be used to increase the validity of GBA.

Previous studies also indicated some strategies when developing GBA. For example, researchers may develop reliable competency models to indicate various in-game performances. Indicators and variables of the competency model can be effectively used to analyze the multifaceted construct of each competency. While embedding in-game assessment, external measures, such as surveys or tests, are needed to validate stealth assessment. Moreover, it is necessary to align reliable standards, assessment, and measurable learning objectives when developing GBA. However, some external variables may influence learners' gameplaying experiences and performances, such as gaming experiences or performance errors. Therefore, it is important for researchers to carefully control the influence of these variables and reduce their impact on the accuracy of the results.

All in all, not every game is suitable to serve as a pedagogical tool, and not every game can be converted into a GBA instrument. It is often time-consuming to develop valid in-game assessment methods. Hence, the practitioners should take some factors into consideration before implementing strategies to embed assessments in GBL environments. For example, they may consider the purpose of assessment and identify the competencies which need to be measured before deciding whether it is reasonable to embed GBA as an instrument to measure students' in-game performances or learning outcomes. Although ECD and Bayes network can be helpful to develop in-game assessment tools, they are not the only options. Instructors may design GBA based on their specific situations, such as learners' experiences and learning contexts.

4.6 Recommendations for Future GBA Practices

After understanding how to design and implement GBA, several recommendations for the future practices will be discussed. First, it is important to notice that the development of GBA is grueling and time-consuming. In order to produce valid and reliable assessment prototypes, researchers often need to continually design, test, and revise, indicating an iterative cycle of designing and developing GBA. Sometimes, players may face the problems which the designers did not expect. Under this circumstance, teamwork, collaborations among professionals in different areas, is necessary. The design team may include subject matter experts (SMEs) or content experts, game developers, technologists, and assessment designers (DiCerbo, 2017).

In addition, assessment developers may take learners' differences into consideration when developing GBA (All et al., 2016). Learners are different from each other and from various backgrounds, which may lead to different levels of skills and prior knowledge. These differences will affect their emotions, perceptions, and learning experiences in game-based environments. Therefore, in order to ensure the accuracy of in-game performance data, team members may consider how to maintain fairness among learners from diverse backgrounds and reduce the impact of these differences on their outcomes (Kim & Shute, 2015).

Instructors may also consider adopting additional components to enhance GBL outcomes and assessment. Gameplaying can motivate students to produce active and engaging learning experiences. However, gameplaying alone is not enough to effectively let learners obtain knowledge and to assess their capabilities. Therefore, instructors should rethink about the role of digital games and in-game assessment in their specific contexts. For example, digital games may play a complementary role to traditional teaching methods. Instructors may consider integrating other strategies to enhance students' learning outcomes and the effectiveness of measurements, such as group discussions or debriefing sessions.

Institutions should provide support services to promote the development of GBA. For example, technical support should be available to promote learners' GBL experiences. Institutions may also provide benign environments where instructors can take risks to develop innovative assessment strategies (Lock et al., 2018). Moreover, administrators may adjust institutional policies to encourage collaborations among various parties to promote curriculum development and GBA design.

4.7 Limitations and Recommendations

The chapter aims to lay the foundations for the future practices and research and mainly discusses the use of stealth assessment from a theoretical angle. Researchers may consider building on the theoretical base to conduct studies to investigate the effectiveness of stealth assessment to measure soft skills in various contexts. The author only discusses how to embed stealth assessment as a formative assessment method to provide learners with feedback during their learning process. Future research may investigate how to embed summative or confirmative assessments in GBL environments. Other researchers may discuss how to balance game design and the effectiveness of GBA, such as how to maintain game design considerations while improving the validity and reliability of assessments.

The chapter exclusively focuses on how stealth assessment was used to measure soft skills, including teamwork, problem-solving, and creativity. Future research may investigate the integration of stealth assessment to measure other types of soft skills, such as leadership, time management, and critical thinking. Other researchers may also focus on the effectiveness of other types of GBA, such as external assessments and game scoring, to measure soft skills or the limitations of GBA on soft skills measurement. Moreover, with the development of technologies, researchers may study the effectiveness of stealth assessment in virtually situated environments, such as virtual reality (VR). Due to different genres of digital games, future research may compare and contrast design and implementation strategies between serious games and commercial games. Learners' characteristics may influence the effectiveness of GBL and GBA; further research may investigate the correlations between in-game performances and different variables, such as learners' readiness, comfort level, and motivation.

4.8 Conclusion

With the development of technology, digital games have been widely applied in learning settings to enhance students' learning outcomes. However, GBA is still underdeveloped in the field of education. In many formal educational institutions, instructors often rely on traditional summative assessment methods to evaluate students' academic performance and the effectiveness of learning materials. In GBL environments, many instructors may choose external assessment tools to measure students' learning outcomes. However, these types of assessment indicate some limitations in producing accurate and reliable results, especially when real-time performance-based data is necessary to interpret students' competencies (Alcañiz et al., 2018). Therefore, to address the weaknesses, there is a need to develop alternative assessment tools to increase the validity of measurement. One example of GBA is called stealth assessment, which is beneficial in collecting students' performance data invisibly and providing reliable feedback to enhance their learning outcomes.

In the twenty-first century, some competencies, such as problem-solving, critical thinking, and teamwork, are highly valued. Stealth assessment is useful to evaluate learners' in-game performances, but limited research has focused on the integration of stealth assessment to measure soft skills. Therefore, in the chapter, the author reviewed previous studies and further discussed effective practices and procedures of implementing stealth assessment to measure soft skills. Eventually, the author concludes several strategies to design and implement GBA in the future practices. For instance, instructors should know about the purpose of assessment or what is the competency to be measured. ECD and Bayes network may be appropriately used to develop GBA, but instructors should take their specific contexts into consideration. Assessment needs to align with measurable variables and learning objectives, so it is important to rely on external measures to increase the validity of stealth assessment. In this case, the indicators and variables of the reliable competency models can be used to effectively measure learners' in-game performances. Moreover, instructors should understand how challenging and time-consuming the process is. ECD and Bayes network include testing and revision components; thus, it is necessary for instructors to work with other experts, such as SMEs in assessment and game developers, to constantly make reasonable adjustments to ensure the reliability of GBA instruments.

Through invisibly collecting students' authentic performance-based data, stealth assessment indicates its benefits to reduce their test anxiety and increase the accuracy of evaluation results. In this way, the results may inform instructors of providing reasonable remediations to support student leaning. However, there is no perfect model of embedding stealth assessment to measure soft skills in GBL environments. Instructors may develop or refer to appropriate GBA development frameworks according to their students' needs and learning contexts. With the development of technologies, high-quality digital games may support interactive and immersive learning environments in which instructors can better conduct training and assessment. Virtual reality and simulation technologies can also be applied to train and measure complex competencies (Alcañiz et al., 2018; de-Juan-Ripoll et al., 2018). Moreover, due to the implicit nature of stealth assessment, instructors should realize the ethical issues involved in the measurement and handle personal data in a responsible way.

In conclusion, the chapter aims to review previous GBA design practices and analyze strategies and procedures of embedding stealth assessment to measure soft skills. The findings from this chapter will contribute to the body of knowledge in the field of GBA and lay a solid foundation for the future research and design practices. Due to the complicated process of designing GBA, it remains underdeveloped in the field of education and educational evaluation. However, because of the promising future of GBA to engage and empower students, there is a need to encourage the adoption of GBA to classrooms. The chapter intends to provide meaningful information to advocate GBA design and implementation in the future teaching and assessment activities. However, instructors should always keep in mind that digital games or in-game assessment are one example of many tools to facilitate students to achieve learning objectives. They should select the technology for the sake of the learning benefits rather than technology itself.

References

Alcañiz, M., Parra, E., & Giglioli, I. A. C. (2018). Virtual reality as an emerging methodology for leadership assessment and training. *Frontiers in Psychology, 9*, 1658. Retrieved from https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01658/full

All, A., Patricia Nunez Castellar, E., & Van Looy, J. (2015). Towards a conceptual framework for assessing the effectiveness of digital game-based learning. *Computers & Education, 88*, 29–37.

All, A., Patricia Nunez Castellar, E., & Van Looy, J. (2016). Assessing the effectiveness of digital game-based learning: Best practices. *Computers & Education, 92-93*, 90–103.

Borji, Y. E., & Khaldi, M. (2014). Comparative study to develop a tool for the quality assessment of serious games intended to be used in education. *International Journal of Emerging Technologies in Learning, 10*, 50–55.

Brown, A., & Green, T. (2016). Virtual reality: Low-cost tools and resources for the classroom. *TechTrends, 60*(5), 517–519.

Caballero-Hernandez, J. A., Palomo-Duarte, M., & Dodero, J. M. (2017). Skill assessment in learning experiences based on serious games: A systematic mapping study. *Computers & Education, 113*, 42–60.

De Klerk, S., & Kato, P. M. (2017). The future value of serious games for assessment: Where do we go now. *Journal of Applied Testing Technology, 18*(S1), 32–37.

de-Juan-Ripoll, C., Soler-Domínguez, J. L., Guixeres, J., Contero, M., Álvarez Gutiérrez, N., & Alcañiz, M. (2018). Virtual reality as a new approach for risk taking assessment. *Frontiers in Psychology, 9*, 1–8.

DeRosier, M. E., & Thomas, J. M. (2018). Establishing the criterion validity of Zoo U's game-based social emotional skills assessment for school-based outcomes. *Journal of Applied Developmental Psychology, 55*, 52–61.

Dessinger, J. C., & Moseley, J. L. (2003). *Confirmative evaluation: Practical strategies for valuing continuous improvement*. Hoboken, NJ: Wiley.

Dewey, J. (1910). *How we think*. New York, NY: D. C. Heath and Company.

DiCerbo, K. E. (2017). Building the evidentiary argument in game-based assessment. *Journal of Applied Testing Technology, 18*(S1), 7–18.

Faizan, N., Löffler, A., Heininger, R., Utesch, M., & Krcmar, H. (2019). Classification of evaluation methods for the effective assessment of simulation games: Results from a literature review. *International Journal of Engineering Pedagogy, 9*(1), 19–33.

Flynn, R., Bacon, E., & Dastbaz, M. (2010). Impact of learning paradigms in games design: How the theory of learning has influenced the design of video games. *Global Learn Asia Pacific, 2010*(1), 1550–1556.

Hickman, L. (1990). *John Dewey's pragmatic technology*. Bloomington, IN: Indiana University Press.

Hwa, S. P. (2018). Pedagogical change in mathematics learning: Harnessing the power of digital game-based learning. *Educational Technology & Society, 21*(4), 259–276.

Ke, F., & Shute, V. (2015). Design of game-based stealth assessment and learning support. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics* (pp. 301–318). Cham, Switzerland: Springer.

Kiili, K. (2005). Digital game-based learning: Towards an experiential gaming model. *Internet and Higher Education, 8*(1), 13–24.

Kim, Y. J., Almond, R. G., & Shute, V. J. (2016). Applying evidence-centered design for the development of game-based assessments in physics playground. *International Journal of Testing, 16*(2), 142–163.

Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Computers & Education, 87*, 340–356.

Lock, J., Kim, B., Koh, K., & Wilcox, G. (2018). Navigating the tensions of innovative assessment and pedagogy in higher education. *Canadian Journal for the Scholarship of Teaching and Learning, 9*(1), 8–18.

Mayer, I. (2018). Assessment of teams in a digital game environment. *Simulation & Gaming, 49*(6), 602–619.

Mayer, I., Warmelink, H., & Bekebrede, G. (2013). Learning in a game-based virtual environment: A comparative evaluation in higher education. *European Journal of Engineering Education, 38*(1), 85–106.

Morrison, G. R., Ross, S. M., Kemp, J. E., & Kalman, H. (2010). *Designing effective instruction*. Hoboken, NJ: Wiley.

Perini, S., Luglietti, R., Margoudi, M., Oliveira, M., & Taisch, M. (2018). Learning and motivational effects of digital game-based learning (DGBL) for manufacturing education –the life cycle assessment (LCA) game. *Computers in Industry, 102*, 40–49.

Petrovica, S., & Anohina-Naumeca, A. (2017). The adaptation approach for affective game-based assessment. *Applied Computer Systems, 22*, 13–20.

Shute, V. J., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. Cambridge, MA: MIT.

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior, 63*, 106–117.

Snead, J. (Director). (2014). *Video games: The movie* [Motion picture]. Dallas, TX: Media Juice Studios Ltd.

Snow, E. L., Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2014). Does agency matter?: Exploring the impact of controlled behaviors within a game-based environment. *Computers & Education, 82*, 378–392.

Snow, E. L., Likens, A. D., Allen, L. K., & McNamara, D. S. (2016). Taking control: Stealth assessment of deterministic behaviors within a game-based system. *International Journal of Artificial Intelligence in Education, 26*(4), 1011–1032.

Tsai, F., Tsai, C., & Lin, K. (2015). The evaluation of different gaming modes and feedback types on game-based formative assessment in an online learning environment. *Computers & Education, 81*, 259–269.

Zull, J. E. (2002). *The art of changing the brain: Enriching teaching by exploring the biology of learning*. Sterling, VA: Stylus Publishing.

Chapter 5
Intrinsic Motivation in Game-Based Learning Environments

T. Fulya Eyupoglu and John L. Nietfeld

5.1 Introduction

With the increased use of game-based learning environments (GBLEs) in educational settings, more attention has been directed toward understanding the motivational benefits and challenges presented in those environments. In educational research, the study of motivation has a rich literature in traditional learning environments and a growing presence in digital learning environments. Motivation has been defined as physiological processes involved in the direction, vigor, and persistence of behavior (Moos & Marroquin, 2010). The end state that educators desire to achieve is a motivated learner who is self-determined and driven by his or her own desire (Garris, Ahlers, & Driskell, 2002).

GBLEs are appealing because of their potential as motivational learning tools (Ke, 2009), and to date, many different motivational constructs (i.e., goal orientations, self-efficacy, interest) have been investigated. Yet, findings related to the impact of motivation on learning in GBLEs have been mixed (Clark, Tanner-Smith, & Killingsworth, 2016; Wouters, van Nimwegen, Oostendorp, & van der Spek, 2013). Within this literature, there has been particular emphasis paid to the construct of intrinsic motivation (Garris et al., 2002). Intrinsic motivation has been characterized as performing an activity for the inherent satisfaction of the activity itself (Ryan & Deci, 2000), as opposed to extrinsic motivation that represents a desire to engage in behavior due to external incentives, such as money, grades, and praise (Moos & Marroquin, 2010).

The work of Malone and Lepper (1987), which proposed a link between motivation and intrinsic learning, has been adopted by many studies of GBLEs primarily focusing on four factors: challenge, curiosity, control, and fantasy (Huang, Huang, & Tschopp, 2010). However, studies also focus on other intrinsic motivation factors

T. Fulya Eyupoglu (✉) · J. L. Nietfeld
North Carolina State University, Raleigh, NC, USA
e-mail: fulya_eyupoglu@ncsu.edu

D. Ifenthaler, Y. J. Kim (eds.), *Game-Based Assessment Revisited*, Advances in Game-Based Learning, https://doi.org/10.1007/978-3-030-15569-8_5

such as interest (Admiraal, Huizenga, Akkerman, & Ten Dam, 2011; Hamari et al., 2016) and concentration (Ronimus, Kujala, Tolvanen, & Lyytinen, 2014). Moreover, studies in GBLEs tend to use engagement and intrinsic motivation as interchangeable constructs. For example, Ronimus et al. (2014) used the rating of enjoyment as a predictor of student's engagement as well as a measure of intrinsic motivation in the game *GraphoGame*.

This wide variation in the operationalization of intrinsic motivation as well as the employment of numerous brief and unstandardized questionnaires in the assessment of intrinsic motivation has led to a lack of clarity across studies (Brockmyer et al., 2009). Understanding the composition of intrinsically motivated user experiences and how to evaluate these experiences is necessary in the design of interactive systems (O'Brien & Toms, 2010). Thus, agreeing upon definitional and measurement approaches for intrinsic motivation in GBLEs may be a crucial step in understanding the attraction of games and the effectiveness of instruction (Garris et al., 2002).

The purpose of this chapter is to examine how researchers have implemented and assessed intrinsic motivation in GBLEs and to draw some conclusions and recommendations for future research. The following sections will provide a review and analysis of the following issues related to intrinsic motivation in GBLEs: (1) theoretical frameworks and definitions used to examine intrinsic motivation in GBLEs, (2) measurement issues in assessing intrinsic motivation, (3) the relationship between intrinsic motivation and learning and performance outcomes in GBLEs, and (4) the impact of GBLE components on intrinsic motivation. The review will close with conclusions and future directions for the field.

5.2 Theoretical Frameworks and Definitions for Intrinsic Motivation in GBLEs

Malone's (1981) work attempting to determine the elements of intrinsic motivation relevant to games set the stage for the study of intrinsic motivation in digital gaming environments. From a theoretical perspective, this work contributed to prevalent motivational theories that now inform serious-game development. Current studies that measure some form of intrinsic motivation as a primary dependent variable predominately situate themselves within either Self-Determination Theory (SDT) (Ryan & Deci, 2000) or the theory of flow (Csikszentmihalyi, 1975), a related construct that contains a significant degree of overlap with the elements of intrinsic motivation described above. Below is a more detailed description of these theories.

Malone (1981) proposed a theory of intrinsic motivation suggesting that games are rewarding due to a combination of challenge, fantasy, and curiosity, in which challenge is dependent on the degree of difficulty and level of uncertainty to drive players (Iacovides, Aczel, Scanlon, Taylor, & Woods, 2011). According to Malone and Lepper (1987), there is an optimal level of challenge desired by players that is neither too difficult nor too easy (Garris et al., 2002). Fantasy is defined as the way players imagine themselves in the game by using intense, realistic images (Iacovides et al., 2011) that allow users to experience real-world processes from different perspectives (Garris et al., 2002). The third aspect, curiosity, stimulates players to learn more through their senses (e.g., using light or sound) or cognition and involves mystery or puzzlement (Whitton, 2011). Curiosity increases players' desire to keep playing the game and to learn more about upcoming actions. As there are social factors impacting motivation, Malone and Lepper (1987) later added the elements of control, recognition, competition, and cooperation (Iacovides et al., 2011). Thus, researchers have theorized seven individual elements: challenge, fantasy, curiosity, control, competition, cooperation, and recognition that promote intrinsic motivation (Admiraal et al., 2011).

5.2.1 Studies Based on Self-Determination Theory

SDT is a comprehensive theoretical framework that has the potential for clarifying definitional ambiguities and related measurement issues within studies of intrinsic motivation in GBLEs (Ryan & Deci, 2000). SDT addresses factors that either promote or threaten intrinsic motivation (Ryan, Rigby, & Przybylski, 2006). The theory is focused on the satisfaction of three innate psychological needs: autonomy, competence, and relatedness, that lead to enhanced self-motivation accompanied by commitment, effort, and high-quality performance (Ryan & Deci, 2000). A subtheory of SDT, cognitive evaluation theory (CET), suggests that events and conditions that enhance a person's sense of autonomy and competence support intrinsic motivation (Ryan et al., 2006). While autonomy provides players the freedom to make in-game choices, competence is the ability to effectively perform the behavior (Peng, Lin, Pfeiffer, & Winn, 2012). The third psychological need within SDT is relatedness, which refers to a player's feeling of belonging in the learning environment (Eseryel, Law, Ifenthaler, Ge, & Miller, 2014).

A significant amount of work has examined intrinsic motivation using SDT as a theoretical framework with commercial games as opposed to GBLEs (Przybylski, Rigby, & Ryan, 2010; Sheldon & Filak, 2008; Tamborini, Bowman, Eden, Grizzard, & Organ, 2010). A full review of motivation in commercial games is beyond the focus of this chapter; however, we highlight the Ryan et al. (2006) study that used SDT to examine motivation across different gaming contexts for illustrative purposes. In that study, intrinsic motivation was described as the core element of motivation relevant to computer games and defined as "inherent satisfactions derived from action" (Ryan & Deci, 2000). Intrinsic motivation was operationalized as autonomy, competence, presence, and relatedness in the game environment. Results indicated that perceived in-game autonomy and competence were associated with game enjoyment, preferences, and changes in well-being before and after gameplay. Competence and autonomy perceptions were also related to the intuitive nature of game controls, and the sense of presence or immersion in participants' gameplay experiences. Moreover, autonomy, competence, and relatedness independently predicted enjoyment and future gameplay.

Eseryel et al. (2014) and Chen and Law (2016) employed SDT within studies of GBLEs; however, both utilized interest as a reflection of intrinsic motivation and suggested that competence and autonomy are needed to maintain intrinsic motivation. In addition, Eseryel et al. (2014) discussed the importance of relatedness to foster intrinsic motivation. However, neither study provided an explicit definition of intrinsic motivation. Moreover, Hawlitschek and Joeckel (2017) utilized SDT to discuss the effects of extrinsic incentives on intrinsic motivation but included a definition of intrinsic motivation referring to flow experience. The authors discussed self-efficacy, autonomy, and social environment as important factors for intrinsic motivation. Finally, Burgers, Eden, van Engelenburg, and Buningh (2015) defined intrinsic motivation as the motivation to pursue an activity for its own sake and framed the study within the CET framework.

5.2.2 Studies Based on Flow Theory

The Csikszentmihalyi (1975) flow theory is a prevalent theory of motivation that has been integrated within GBLEs. Flow theory is dependent on the balanced relationship of skills of the user and the challenges in an activity, where one's skills are neither overmatched nor underutilized to meet a given challenge (Shernoff, Csikszentmihalyi, Shneider, & Shernoff, 2003). The achievement of balance leads to feelings of pleasure and time distortion occurring simultaneously (Bouvier, Lavou, & Sehaba, 2014). The flow state of users is distinguished by a high level of enjoyment and fulfillment (Bellotti, Kapralos, Lee, Moreno-Ger, & Berta, 2013). In addition, the person in the state of flow has an intrinsically rewarding experience with clear goals and is involved in a high degree of concentration and a sense of personal control (Boyle, Connolly, Hainey, & Boyle, 2012). Csikszentmihalyi and Csikszentmihalyi (1992) described flow as a technical term in the field of intrinsic motivation; thus, numerous studies have adopted the term flow as a synonym with intrinsic motivation.

The bulk of the studies examining intrinsic motivation in GBLEs have used the theory of flow as a theoretical framework. However, even within these studies the definitions and individual factors used to investigate intrinsic motivation have differed widely. For instance, Admiraal et al. (2011) and Huizenga, Admiraal, Akkerman, and Dam (2009) discussed the work of Malone and Lepper (1987) and the factors contributing to intrinsic motivation as challenge, curiosity, control, fantasy, competition, cooperation, and recognition in the game *Frequency 1550*. However, while the study of Admiraal et al. (2011) investigated only the cooperation and competition factors of intrinsic motivation, Huizenga et al. (2009) did not refer to any individual elements of intrinsic motivation. Hamari et al. (2016) also investigated the impact of flow, operationalized as heightened challenge and skills, on engagement in the physics-based games, *Quantum Spectre* and *Spumone*. The authors described intrinsic motivation as a reflection of interest but did not discuss any other indicators of intrinsic motivation in their study. Although all three studies related flow theory to intrinsic motivation, no clear definition of intrinsic motivation was provided. Similarly, Sung, Hwang, and Yen (2015) discussed the theory of flow and measured intrinsic motivation, yet no definition of intrinsic motivation was mentioned in the study.

Erhel and Jamet (2013) investigated the impact of instructions and feedback on multiple components of motivation (goal orientations and intrinsic motivation) using flow theory in a simulation game, *ASTRA*. The authors adopted Deci and Ryan's (2000) definition of intrinsic motivation as "inner desire to engage in a task out of interest or amusement or because of the challenge it poses" (p. 157). However, no specific intrinsic motivation construct related to flow theory was investigated in the study. Likewise, Ronimus et al. (2014) adopted Deci and Ryan's (2000) definition of intrinsic motivation as "a situation where actions are performed in the absence of any apparent external contingency, that is, an intrinsically motivated person finds the activity rewarding in itself and does not expect to gain anything particular, such as extrinsic rewards, from it" using the web-based game, *GraphoGame* (p. 238). While the authors discussed intrinsic motivation factors by Malone (1980) and Sweetser and Wyeth (2005) that included concentration, challenge, control, clear goals, feedback, immersion, and social interaction, their measures were limited to the impact of the level of challenge and reward upon on students' motivation. Moreover, the terms engagement and motivation were used interchangeably throughout the study to describe links between flow theory and intrinsic motivation.

More recently, Sweetser and Wyeth (2005) have developed a model, GameFlow, which consists of eight core elements overlapped with the elements of flow: concentration, challenge, skills, control, clear goals, feedback, immersion, and social interaction. Several researchers (e.g., Chen, 2007; Fu, Su, & Yu, 2009) have utilized the model to explain how to facilitate flow experiences in computer games (Fang, Zhang, & Chan, 2013). However, despite being used to evaluate a variety of games and applications, including educational games, the model has not been validated and, no operationalization is proposed for converting the model into a measure (Sweetser et al., 2017).

It should also be noted that, while numerous studies emphasizing intrinsic motivation have been contextualized in the predominant theories described above, a significant number have lacked theoretical frameworks (e.g., Vos, van der Meijden, & Denessen, 2011; Chen, Wang, & Lin, 2015; Sung, Hwang, Lin, & Hong, 2017; Liao, Chen, and Shih (2019). For example, Tuzun, Yılmaz-Soylu, Karakuş, Inal, and Kızılkaya (2009) examined the effects of games on primary school students' achievement and intrinsic motivation in geography learning with the game *Quest Atlantis*. Chen (2018) explored how different contexts of gameplay (individual, collaboration, and competition) affected learning outcomes and intrinsic motivation of middle-school students in a science game, *SumMagic*. Although intrinsic motivation was specifically investigated in both studies, no theoretical background was presented. Also, the authors did not conceptualize intrinsic motivation.

5.3 The Measurement of Intrinsic Motivation in Game-Based Learning

Given the interest in measuring intrinsic motivation in GBLEs and the accompanying assumption that intrinsic motivation will be heightened in such environments, the lack of attention dedicated to measuring intrinsic motivation in the literature is surprising. The field is currently saturated with self-report measures, mostly functioning as the sole measure of intrinsic motivation, with scant evidence for validity.

Deci and Ryan's (2000) Intrinsic Motivation Inventory (IMI) is one of the few instruments that have emerged from the hodgepodge of self-report measures employed by the field. Although a number of researchers have aimed for consistency across studies by using the IMI, there has been considerable variance in its application. For example, out of 45 items, Eseryel et al. (2014) administered 25 adapted items, across four subscales (interest, perceived competence, perceived autonomy, and perceived relatedness). The reliability for all subscales according to Cronbach's alpha was 0.92, 0.77, 0.72, and 0.77, respectively. On the other hand, Chen (2018) and Vos et al. (2011) administered 14 adapted items across four subscales (interest, perceived competence, tension, and perceived value) and three subscales (interest, perceived competence, and effort), respectively. Chen (2018) reported the Cronbach's alpha value of the subscales as 0.86, 0.81, 0.77, 0.83, respectively. Furthermore, Liao et al. (2019) reported using 14 items from the IMI; however, the specific subscales were not identified. Reliability of the intrinsic motivation scale was indicated by its Cronbach's alpha of 0.82.

The Motivated Strategies for Learning Questionnaire (MSLQ) has also been used to measure intrinsic motivation in GBLE studies. As with the IMI, studies employing the MSLQ have varied significantly in the administration of the inventory. For instance, Chen et al. (2015) included four items from the MSLQ by Pintrich, Smith, Garcia, and McKeachie. The reliability for intrinsic motivation according to Cronbach's α was 0.86. Despite indicating that the survey had sufficient validity, the authors have not described how they ensured the validity of the survey. Sung et al. (2015) and Sung et al. (2017) also included seven items from the MSLQ by Pintrich and DeGroot. The authors did not discuss how and why they adapted those items from the MSLQ. Sung et al. (2015) reported the Cronbach's alpha value of the questionnaire as 0.88. The authors had each item reviewed by a researcher with more than 10 years' experience of studying learning and motivation as a check for validity.

The reliability of the questionnaires was provided in the above-mentioned studies, yet other studies (e.g., Chen & Law, 2016; Hamari et al., 2016; Sung et al., 2017; Vos et al., 2011) did not report reliability. Moreover, evidence for validity has rarely been reported. Despite adapting the same inventory, the researchers also used different scales in their surveys, a 5-point Likert versus a 7-point Likert scale.

Even greater variation in measurement approaches is evident in other studies of intrinsic motivation. Admiraal et al. (2011) and Huizenga et al. (2009) measured intrinsic motivation with a 6-item, 5-point Likert scale questionnaire adopted by

Cito for use with the *Frequency 1550* game. The individual intrinsic motivation elements investigated were challenge, curiosity, control, fantasy, competition, cooperation, and recognition. Chen and Law (2016) included an intrinsic motivation survey developed by Lin, Atkinson, Christopherson, Joseph, and Harrison in the science game *Carrot Land.* The survey was composed of seven questions with three subscales—interest, competence, and autonomy. Moreover, studies failed to measure variables of interest despite emphasizing the importance of those variables for intrinsic motivation. Hawlitschek and Joeckel (2017) used a modified questionnaire, adapted from Isen and Reeve (2005), to measure intrinsic motivation in the game, *1961*. Six items measured situational interest, curiosity, and fun as individual factors of intrinsic motivation. Although the authors discussed the importance of self-efficacy and autonomy for the intrinsic motivation, they did not measure these variables in the study.

Significant variation in the administration of surveys is also common across studies. Some studies have administered intrinsic motivation surveys before and after gameplay, while some have administered the surveys only after the gameplay. For instance, Ronimus et al. (2014) used self-report surveys to measure ratings of enjoyment and interest at the end of each gameplay session in the *GraphoGame*. A single interest question asked students how much the player liked school tasks that involved reading at the beginning and end of the gameplay. In addition, a single-item enjoyment question asked the students to rate how they had enjoyed playing the game, only at the end of each play session. The students answered these two questions by clicking one of the five faces on the screen, ranging from having a big smile to a big frown. Neither reliability nor validity was reported. Two items were also sent to the parents at the end of the study to evaluate children's motivation and concentration that asked, "How eagerly did the child play *GraphoGame* during the study?" and "How well did the child concentrate while playing *GraphoGame*?"

The study by Tuzun et al. (2009) was unique in that they employed both quantitative and qualitative measures of intrinsic motivation in their study. The authors developed a motivation scale based on the work of Lepper, Corpus, and Iyengar (2005) to study the effects of games on primary school students' intrinsic motivation in the game *Quest Atlantis*. Nine items were included in a 5-point Likert scale type questionnaire. The subscales of the questionnaire included: the desire for challenging tasks, doing tasks for personal curiosity, and desire for independent mastery. In addition, the authors also asked students four open-ended questions at the end of the gameplay to evaluate learning and motivation. Sample questions were: "How and where may you use the information you obtained?" "How did you feel while collecting information in the Global Village; was it fun or boring?" Unfortunately, triangulation of the data from the open-ended items and the questionnaires was not provided.

Hou (2015) and Tsai, Huang, Hou, Hsu, and Chiou (2016) adapted a questionnaire by Kiili (2006) to measure flow state that consisted of the sub-dimensions challenge–skill balance, clear goals, unambiguous feedback, control, action-awareness merging, concentration on the event at hand, transformation of time, autotelic experience, and loss of self-consciousness. Hou (2015) and Tsai et al.

(2016) reported the Cronbach's alpha value of the questionnaire as 0.94 and 0.84, respectively. In addition to self-report measures, Hou (2015) used multi-approach analysis, by integrating cluster analysis with sequential behavioral pattern analysis, to explore college students' flow experiences and learning behaviors in the simulation game with situated-learning context *Perfect PAPA II*. After students' gaming process were all recorded, their learning behaviors were analyzed with cluster and sequential analysis. This approach is unique in providing in-depth analysis of learners' behavioral patterns during gameplay. Correspondingly, Tsai et al. (2016) used eye-tracking technology to measure the participants' flow experience in *Escape the Lab*. University students' visual behaviors and computer screens were tracked and recorded by the eye-tracking system and the patterns of the visual attention distributions illustrated by heat maps analyses during gameplay.

In summary, it is evident from the review above that self-reports dominate measurement approaches for intrinsic motivation in GBLEs. Other emergent potential sources of measurement include trace data and psychophysiological measures. However, Kivikangas et al. (2011) reviewed psychophysiological measures (e.g., heart rate measured with electrocardiograph, ECG) in digital game research and concluded that these measures do not yet form a common collective field due to significant variation in scientific backgrounds and research purposes. The authors suggested using psychophysiological methods combined with other methods (e.g., self-report and observational data) to add significant precision to studying the gaming experience. Thus, the conformity of the current literature in emphasizing self-reports is one measurement issue for the field to focus on while another is the lack of a corpus of validity evidence for such measures.

5.4 Intrinsic Motivation and Learning and Performance

The majority of previous research on intrinsic motivation in GBLEs has revealed positive effects between measures of intrinsic motivation and learning and performance, with a few findings to the contrary. For example, Vos et al. (2011) found advantages for intrinsic motivation as measured by perceived competence, effort, and interest items. They had fifth and sixth graders either play a simple drag-and-drop game related to understanding Dutch proverbs or create and play their own version of the game. Results showed advantages for the construction group for all three facets of intrinsic motivation and also for self-report deep learning strategies. However, the game environment was limited in its complexity and re-playability thus limiting generalizations. Papastergiou (2009) found that students (16–17 years old) who played a GBLE to teach basic computer memory concepts performed better on a knowledge posttest but also showed greater motivation as measured single-item measures of interest, enjoyment, and engagement summed to represent "overall appeal."

Similarly, Eseryel et al. (2014) demonstrated that learners' motivation determined ninth-grade students' development of complex problem-solving competencies via their engagement during the game *McLarin's Adventure*. The authors found that the challenges associated with complex problem-solving led to an increase in student motivation and engagement. Sung et al. (2015) evaluated the effects of a contextual health educational digital game on fourth-grade students' intrinsic motivation, flow experience, and learning achievement. The authors compared the effects of a contextual digital game versus conventional technology-enhanced learning approach, that is, e-books. Results showed that the contextual game-based learning approach improved the students' intrinsic motivation, achievement, and problem-solving competencies more than those who learned with the conventional e-learning approach.

Likewise, Hamari et al. (2016) showed the impact of flow (operationalized as heightened challenge and skill) on learning in the physics-based games, *Quantum Spectre* and *Spumone* for high school students. The authors used a psychometric survey to measure the participants' concentration, enjoyment, interest, challenge, skills, and immersion. The results indicated that challenge had a positive effect on learning both directly and via increased engagement. However, being skilled in the game did not affect learning directly but rather as a mediation effect through engagement. Tuzun et al. (2009) showed that students (7–14 years old) made significant learning gains by participating in the *Quest Atlantis* and showed statistically significant higher intrinsic motivation (desire for challenging tasks, doing tasks for personal curiosity, and desire for independent mastery) as compared to students in a traditional school environment.

Contrary to the above findings, Admiraal et al. (2011) found that intrinsic motivation, described as flow, was not related to students' (12–16 years old) learning outcomes after taking educational level into consideration in the game *Frequency 1550*. The higher the educational level of students, the more flow their teams showed, and the more students learned about history content in the game. Team flow was related to group performance in the game, but not related to student learning outcomes. Correspondingly, Huizenga et al. (2009) found no significant differences between students (12–16 years old) playing the game *Frequency 1550* versus attending regular lessons with respect to motivation. The authors attributed these findings to a lack of motivation with one-day gameplay and technical problems during the gameplay. However, the results showed students who played the game gained significantly more knowledge than those students who received regular project-based instruction. Sung et al. (2017) examined the effectiveness of an experiential game-based learning approach to promote fifth-grade students' learning outcomes and motivation. They found that the students who learned with the experiential gaming mode showed higher intrinsic motivation, better conceptions of deep learning strategies, and higher acceptance of the learning technology than those learning with the conventional technology-enhanced learning approach. However, no significant difference was found between the learning achievements of the two groups.

Finally, Barzilai and Blau (2014) analyzed data of 6–14 year olds from a larger data collection to examine the impact of pre or post scaffolds using a business simulation game called *Shakshouka Restaurant*. Flow and perceived enjoyment were largely unrelated to achievement in the study, with only flow showing a direct relationship with problem-solving in the "play and study" condition where the study scaffold followed gameplay. Interestingly, the study and play condition, where the scaffold was presented before gameplay, did not negatively impact flow or enjoyment. The authors attributed this finding to the fact that the scaffold was directly connected to the game narrative.

5.5 Intrinsic Motivation and GBLE Components

GBLEs provide a myriad of components ranging from instructions, scaffolds, features that impose cognitive load, game narratives, academic content to game mechanics, and variations in the user interface that have the potential to impact intrinsic motivation. Yet, there is a paucity of studies making direct comparisons between such components and intrinsic motivation, particularly within experimental designs allowing for causal conclusions. However, some early insights have been gained and it is expected that momentum will rapidly increase understanding in this area. Chen and Law (2016) found a negative relationship between hard scaffolds (writing prompts) and intrinsic motivation as measured by competence, autonomy, and interest items. The study examined seventh graders playing a game called *Carrot Land* that focused on force and motion curriculum. Soft scaffolds (collaboration) had a negative effect for competence but no relationship with autonomy or interest. However, when both types of scaffolds were included together rather than separately there was a positive effect over the two isolated scaffold conditions. Chen (2018) also examined whether the presence or absence of collaboration and competition affects learning outcomes and intrinsic motivation of middle school students in a science game, *SumMagic.* The results revealed that students who played the game collaboratively with intergroup competition showed a significantly higher interest in learning, a higher value of the learning, and lower tension during the learning process than those who played individually. Similarly, Liao et al. (2019) explored the effects of using an instructional video and collaboration on achievement, intrinsic motivation, cognitive load, and science learning behaviors of seventh-grade students within the game *SumMagic*. They also found that collaborative learning increased students' intrinsic motivation. Also, integrating instructional video use had an increase on students' achievement in gameplay. Although integrating instructional video use had no effects on intrinsic motivation, integrating instructional video in combination with collaborative gameplay improved intrinsic motivation and achievement.

Erhel and Jamet (2013) explored the effects of learning instruction versus entertainment instruction on the intrinsic motivation of college students (18–26 years old) in a simulation game, *ASTRA*. Results indicated that the learning instruction condition (instructions stressed ASTRA's educational dimension, presenting it as a learning module) elicited deeper learning than the entertainment condition (instructions stressed ASTRA's playful dimension, presenting it as a game). Choice of an entertainment instruction appeared to hinder learning; however, this effect was nullified by the addition of feedback that supplied the correct answers. On the other hand, no effect of instruction on the other learning goals or intrinsic motivation was shown.

Hawlitschek and Joeckel (2017) examined the effects of integrating instructional support on students' (13–17 years old) intrinsic motivation, cognitive load, and learning with a digital educational game, *1961*. The authors applied an explicit instruction learning prompt before gameplay: "Your task is to play an educational game. Afterwards, you will be asked some questions about the learning content! Try to learn as much as possible!" Students in the group with learning instruction achieved lower scores in the transfer knowledge test with the results attributed to an increase in extraneous cognitive load. However, the authors did not find an effect of the learning instruction on intrinsic motivation. Burgers et al. (2015) explored the role of feedback on intrinsic motivation and future play in an educational brain-training game, *Concentration*. Results demonstrated that evaluative feedback increased, but comparative feedback decreased future gameplay. Moreover, positive feedback increased intrinsic motivation by satisfying competence and autonomy needs leading to long-term motivation and play whereas negative feedback only motivated players to repair poor short-term performances.

5.6 Discussion and Conclusion

This chapter has reviewed studies that have implemented and assessed intrinsic motivation in GBLEs. Given the lack of other existing reviews related to this topic in the literature, we provide some overall conclusions synthesizing our findings followed by future research recommendations (see below) to guide studies that explore intrinsic motivation in GBLEs. The following conclusions relate to issues of definitional clarity, approaches to measurement, and directionality of causal relations.

The studies reviewed above have revealed little consensus in the conception of intrinsic motivation in GBLEs even among investigations using the same theoretical framework. In order to create more effective GBLEs that increase or facilitate intrinsic motivation, researchers would likely benefit from developing a consensus on the definition of intrinsic motivation and to then subsequently consider more sophisticated measurement approaches. Reviews in this chapter revealed common flaws seen in studies of GBLEs including the insufficient conceptualization of intrinsic motivation and the lack of comparability across studies due to inconsistent use of measures. Currently, definitions, when provided, are partially overlapping but with enough variation to lead to important differences in findings across studies. Many studies noted the individual elements (i.e., challenge, curiosity, control, fantasy, competition, cooperation, and recognition) contributing to intrinsic motivation in GBLEs but neglected to investigate these factors. The lack of consistency in defining intrinsic motivation, even when referencing other publications in terms of how intrinsic motivation is contributing to learning in GBLEs, made it unclear as to whether authors were referring to the same construct across studies. Moreover, terminology related to engagement, enjoyment, immersion, presence, and motivation have been associated with the flow and frequently used interchangeably, creating more confusion in the literature (Hamari et al., 2016).

Self-report questionnaires are the dominant indicators of intrinsic motivation and the variation across studies in questionnaires and items is great. This may be due to the fact that it is not easy to integrate standard questionnaires across very different GBLEs (Hsieh, Lin, & Hou, 2015). Variation in questionnaires notwithstanding, there was also significant variance in the timing of these questionnaires with some only applied post play and some applied pre and post play.

The general pattern of results thus far in the literature suggests positive findings between intrinsic motivation and learning and performance outcomes. However, little is still known about the mechanisms of the GBLEs that promote or hinder intrinsic motivation. There is not enough empirical evidence to state that it is the type of motivation that drives achievement because the level of achievement might drive the type of motivation as well (Lepper et al., 2005). For example, Tsai et al. (2016) indicated that players with different conceptual learning outcomes in GBLEs had significantly different flow experience while playing the game. Students with higher comprehension expressed higher levels of control and concentration than those with lower comprehension achievements. Individuals with higher abilities have higher flow experiences, but it is not known if the correlation between flow and performance arises simply because expertise leads to more flow, instead of flow fostering performance (Engeser & Rheinberg, 2008).

Also, research in traditional learning environments has revealed lower levels of intrinsic motivation for older students (Lepper et al., 2005). Yet, the results of the studies reviewed here have varied with regard to intrinsic motivation and developmental level. Thus, there is a need to examine how intrinsic motivation changes with age in GBLEs across various genres and domains using psychometrically sound and consistent measurement tools across studies.

There appears to be an emerging positive effect of competition between cooperative groups in GBLEs on intrinsic motivation. However, variation in the effects of other manipulated components within GBLEs on intrinsic motivation and learning appears to be less clear. This lack of clarity can be attributed, to some extent, to the variation in measurement and lack of definitional consistency across studies. Yet, a review of the literature indicates a significant need for more investigations examining the impact of specific GBLE components on intrinsic motivation.

5.7 Suggestions for Future Research

Understanding and assessing intrinsic motivation in GBLEs is at an early stage of development. Although studies presented herein have presented unique contributions to the field, there are common challenges observed in these studies in the way researchers have defined and measured intrinsic motivation. Recognizing this, we present a number of suggestions for researchers as they consider designs that will add to the existing literature. First, it is clear from the review above that definitional clarity is needed with regard to intrinsic motivation. Drawing from numerous definitions, many of which coming from SDT, we suggest that intrinsic motivation should be considered a higher-order construct within the context of GBLEs that is evidenced by players willfully participating in and enjoying a GBLE without extrinsic influences while being engaged in challenging learning activities. In GBLEs that emphasize or require coordination by multiple players' intrinsic motivation would also reflect relatedness on the part of the player.

Second, Habgood and Ainsworth (2011) argued that it is critical to consider the effect of adding learning content to an intrinsically motivating game rather than creating extrinsic games that provide gameplay as a reward for learning content. They assessed the intrinsic integration approach that depends on the ability of educational games to effectively harness the intrinsic motivation. The authors examined the learning gains of students (7–11 years old) who played either the intrinsic, extrinsic, or control variants of an educational math game called *Zombie Division.* They also compared time on task for the intrinsic and extrinsic variants of the game when students had free choice of which game to play. The results indicated that students learned more from the intrinsic version of the game under fixed time limits and spent seven times longer playing it in free-time situations. Such studies offer evidence for the value of an intrinsic approach but there is not enough evidence about the positive effects of this approach on learning. Therefore, similar studies should be conducted and replicated in various contexts for creating effective educational games that foster intrinsic motivation.

Third, given that measures of interest and engagement are often used as proxies for intrinsic motivation we would suggest studies to explicitly examine overlap and distinctions between such constructs. Similarly, constructs in GBLEs such as immersion appear to be used interchangeably with the term flow. This suggests that future research should investigate the definition and measurement of immersion due to such potential overlap with the construct of flow (Sweetser et al., 2017).

Fourth, a number of different approaches were found in the measurement of intrinsic motivation in the current studies. Researchers tend to adapt measures of previous research without questioning the assumptions of the theoretical framework, construct definition, and grain size of measurement that can range from a micro level (i.e., individual in the moment, task, and learning activity) to a macro level (e.g., group of learners in a class, course, school, or community) (see Sinatra, Heddy, & Lombardi, 2015). Therefore, researchers should consider designs that clearly define, measure, and analyze intrinsic motivation and use multiple measurements during gameplay such as think alouds, eye-tracking, or physiological data in addition to self-reports (Azevedo, 2015). To the best of our knowledge, previous studies have not tracked real-time changes in intrinsic motivation during gameplay, and how other constructs such as competence, autonomy, and relatedness are related to any changes in intrinsic motivation in GBLEs. Our recommendation would be to consider trace data approaches where assumptions can be drawn regarding intrinsic motivation during gameplay and be recorded in a stealth assessment approach (Shute, 2011; Ventura, Shute, & Zhao, 2013) that maintains the flow of gameplay.

Multidimensional triangulation methods (e.g., interviews, observations, and trace data) should be developed to enhance the overall validity of the research in the field of GBLEs (Hou, 2015). For example, further investigation of possible relationships between all aspects of flow measures and eye-tracking measures in various GBLE contexts could be beneficial as these measures could be essential indicators for the flow experience (Tsai et al., 2016). Given the complexity of intrinsic motivation as a construct, using a variety of methodologies for a deeper understanding of this construct will assist researchers in maximizing the ability to provide learning opportunities for all students (Phillips, Horstman, Vye, & Bransford, 2014). This includes researchers providing techniques and methodologies that capture the changes in cognitive, metacognitive, motivational, and emotional processes (Azevedo, 2014). We also suggest researchers use longer gameplay time with larger sample sizes when possible. Despite using multi-approach analysis, studies appeared to have very short gameplay with small sample sizes (e.g., Hou, 2015; Tsai et al., 2016). We believe that this is a common challenge due to time constraints with new developing technologies.

Five, studies in GBLEs have reported on the potential of games to increase intrinsic motivation and learning of challenging academic content but concrete empirical data to support or refute these theoretical claims is still missing (Annetta, Minogue, Holmes, & Cheng, 2009). Research that isolates the impact of components of GBLEs on intrinsic motivation is sorely needed. In order to move the field beyond the general understanding that intrinsic motivation is a positive construct to promote learning and performance, specific recommendations are needed for the construction of GBLEs. There is little information regarding the relationship between intrinsic and extrinsic motivation in GBLEs that has implications regarding the way we should assess intrinsic motivation (Guay, Vallerand, & Blanchard, 2000). Thus, future research should continue to examine extrinsic motivation in combination with intrinsic motivation (Guay et al., 2000) as well as the effects of extrinsic motivation on intrinsic motivation in gameplay.

Sixth, as a broader goal related to the facilitation of intrinsic motivation, game developers should attempt to create environments that facilitate self-regulated learning more broadly. This would include an emphasis on learners creating and tracking personal goals and learning to manage their strategy use, time, and reflect on their learning. Providing students with autonomy in the game to manage such goals promotes an environment to simultaneously encourage both self-regulation and intrinsic motivation. Within this context, another important contribution to the literature would be recommendations for game design from a developmental standpoint using cross-sectional and longitudinal designs. Existing literature in educational and developmental psychology might inform developmentally appropriate learning scaffolds and environmental designs.

References

Admiraal, W., Huizenga, J., Akkerman, S., & Ten Dam, G. (2011). The concept of flow in collaborative game-based learning. *Computers in Human Behavior, 27*, 1185–1194.

Annetta, L. A., Minogue, J., Holmes, S. Y., & Cheng, M. (2009). Investigating the impact of video games on high school students' engagement and learning about genetics. *Computers & Education, 53*, 74–85.

Azevedo, R. (2014). Multimedia learning of metacognitive strategies. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 647–672). Cambridge, UK: Cambridge University Press.

Azevedo, R. (2015). Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. *Educational Psychologist, 50*, 84–94.

Barzilai, S., & Blau, I. (2014). Scaffolding game-based learning: Impact on learning achievements, perceived learning, and game experiences. *Computers & Education, 70*, 65–79.

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*, 1.

Bouvier, P., Lavou, E., & Sehaba, K. (2014). Defining engagement and characterizing engaged-behaviors in digital gaming. *Simulation and Gaming, 45*, 491–507.

Boyle, E. A., Connolly, T. M., Hainey, T., & Boyle, J. M. (2012). Engagement in digital entertainment games: A systematic review. *Computers in Human Behavior, 28*, 771–780.

Brockmyer, J. H., Fox, C. M., Curtiss, K. A., McBroom, E., Burkhart, K. M., & Pidruzny, J. N. (2009). The development of the game engagement questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology, 45*, 624–634.

Burgers, C., Eden, A., van Engelenburg, M. D., & Buningh, S. (2015). How feedback boosts motivation and play in a brain-training game. *Computers in Human Behavior, 48*, 94–103.

Chen, C. H. (2018). The impacts of peer competition-based science gameplay on conceptual knowledge, intrinsic motivation, and learning behavioral patterns. *Educational Technology Research and Development, 67*(1), 179–198.

Chen, C. H., & Law, V. (2016). Scaffolding individual and collaborative game-based learning in learning performance and intrinsic motivation. *Computers in Human Behavior, 55*, 1201–1212.

Chen, C. H., Wang, K. C., & Lin, Y. H. (2015). The comparison of solitary and collaborative modes of game-based learning on students' science learning and motivation. *Journal of Educational Technology & Society, 18*(2), 237–248.

Chen, J. (2007). Flow in games (and everything else). *Communications of the ACM, 50*(4), 31–34.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research, 86*, 79–122.

Csikszentmihalyi, M. (1975). Play and intrinsic rewards. *Journal of Humanistic Psychology, 15*, 41–63.

Csikszentmihalyi, M., & Csikszentmihalyi, I. S. (Eds.). (1992). *Optimal experience: Psychological studies of flow in consciousness*. Cambridge, UK: Cambridge University Press.

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*(4), 227–268.

Engeser, S., & Rheinberg, F. (2008). Flow, performance and moderators of challenge-skill balance. *Motivation and Emotion, 32*(3), 158–172.

Erhel, S., & Jamet, E. (2013). Digital game-based learning: Impact of instructions and feedback on motivation and learning effectiveness. *Computers & Education, 67*, 156–167.

Eseryel, D., Law, V., Ifenthaler, D., Ge, X., & Miller, R. (2014). An investigation of the interrelationships between motivation, engagement, and complex problem solving in game-based learning. *Journal of Educational Technology & Society, 17*, 42–53.

Fang, X., Zhang, J., & Chan, S. S. (2013). Development of an instrument for studying flow in computer game play. *International Journal of Human Computer Interaction, 29*(7), 456–470.

Fu, F. L., Su, R. C., & Yu, S. C. (2009). EGameFlow: A scale to measure learners' enjoyment of e-learning games. *Computers & Education, 52*(1), 101–112.

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

Guay, F., Vallerand, R. J., & Blanchard, C. (2000). On the assessment of situational intrinsic and extrinsic motivation: The situational motivation scale (SIMS). *Motivation and Emotion, 24*(3), 175–213.

Habgood, M. J., & Ainsworth, S. E. (2011). Motivating children to learn effectively: Exploring the value of intrinsic integration in educational games. *The Journal of the Learning Sciences, 20*(2), 169–206.

Hamari, J., Shernoff, D. J., Rowe, E., Coller, B., Asbell-Clarke, J., & Edwards, T. (2016). Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Computers in Human Behavior, 54*, 170–179.

Hawlitschek, A., & Joeckel, S. (2017). Increasing the effectiveness of digital educational games: The effects of a learning instruction on students' learning, motivation and cognitive load. *Computers in Human Behavior, 72*, 79–86.

Hou, H. T. (2015). Integrating cluster and sequential analysis to explore learners' flow and behavioral patterns in a simulation game with situated-learning context for science courses: A video-based process exploration. *Computers in Human Behavior, 48*, 424–435.

Hsieh, Y., Lin, Y., & Hou, H. (2015). Exploring elementary-school students' engagement patterns in a game-based learning environment. *Journal of Educational Technology & Society, 18*, 336–348.

Huang, W. H., Huang, W. Y., & Tschopp, J. (2010). Sustaining iterative game playing processes in DGBL: The relationship between motivational processing and outcome processing. *Computers & Education, 55*, 789–797.

Huizenga, J., Admiraal, W., Akkerman, S., & Dam, G. T. (2009). Mobile game-based learning in secondary education: Engagement, motivation and learning in a mobile city game. *Journal of Computer Assisted Learning, 25*, 332–344.

Iacovides, I., Aczel, J., Scanlon, E., Taylor, J., & Woods, W. (2011). Motivation, engagement and learning through digital games. *International Journal of Virtual and Personal Learning Environments, 2*, 1–16.

Isen, A. M., & Reeve, J. (2005). The influence of positive affect on intrinsic and extrinsic motivation: Facilitating enjoyment of play, responsible work behavior, and self-control. *Motivation and Emotion, 29*(4), 295–323.

Ke, F. (2009). A qualitative meta-analysis of computer games as learning tools. In *Handbook of research on effective electronic gaming in education* (Vol. 1, pp. 1–32). Hershey, PA: IGI Global.

Kiili, K. (2006). Evaluations of an experiential gaming model. *An Interdisciplinary Journal on Humans in ICT Environments, 2*(2), 187–201.

Kivikangas, J. M., Chanel, G., Cowley, B., Ekman, I., Salminen, M., Järvelä, S., & Ravaja, N. (2011). A review of the use of psychophysiological methods in game research. *Journal of Gaming & Virtual Worlds, 3*(3), 181–199.

Lepper, M. R., Corpus, J. H., & Iyengar, S. S. (2005). Intrinsic and extrinsic motivational orientations in the classroom: Age differences and academic correlates. *Journal of Educational Psychology, 97*(2), 184.

Liao, C. W., Chen, C. H., & Shih, S. J. (2019). The interactivity of video and collaboration for learning achievement, intrinsic motivation, cognitive load, and behavior patterns in a digital game-based learning environment. *Computers & Education, 133*, 43–55.

Malone, T. W. (1980). *What makes things fun to learn? A study of intrinsically motivating computer games.* Doctoral dissertation, ProQuest Information & Learning.

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science, 5*(4), 333–369.

Malone, T. W., & Lepper, M. R. (1987). Making learning fun: A taxonomy of intrinsic motivations for learning. In R. E. Snow & M. J. Farr (Eds.), *Aptitude, learning, and instruction. Cognitive and affective process analyses* (Vol. 3, pp. 223–253). Hillsdale, NJ: Lawrence Erlbaum.

Moos, D. C., & Marroquin, E. (2010). Multimedia, hypermedia, and hypertext: Motivation considered and reconsidered. *Computers in Human Behavior, 26*, 265–276.
O'Brien, H. L., & Toms, E. G. (2010). The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology, 61*, 50–69.
Papastergiou, M. (2009). Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation. *Computers & Education, 52*, 1–12.
Peng, W., Lin, J. H., Pfeiffer, K. A., & Winn, B. (2012). Need satisfaction supportive game features as motivational determinants: An experimental study of a self-determination theory guided exergame. *Media Psychology, 15*, 175–196.
Phillips, R. S., Horstman, T., Vye, N., & Bransford, J. (2014). Engagement and games for learning expanding definitions and methodologies. *Simulation and Gaming, 45*, 548–568.
Przybylski, A. K., Rigby, C. S., & Ryan, R. M. (2010). A motivational model of video game engagement. *Review of General Psychology, 14*, 154.
Ronimus, M., Kujala, J., Tolvanen, A., & Lyytinen, H. (2014). Children's engagement during digital game-based learning of reading: The effects of time, rewards, and challenge. *Computers & Education, 71*, 237–246.
Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*, 68–78.
Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion, 30*, 344–360.
Sheldon, K. M., & Filak, V. (2008). Manipulating autonomy, competence, and relatedness support in a game-learning context: New evidence that all three needs matter. *British Journal of Social Psychology, 47*, 267–283.
Shernoff, D. J., Csikszentmihalyi, M., Shneider, B., & Shernoff, E. S. (2003). Student engagement in high school classrooms from the perspective of flow theory. *School Psychology Quarterly, 18*, 158–176.
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.
Sinatra, G. M., Heddy, B. C., & Lombardi, D. (2015). The challenges of defining and measuring student engagement in science. *Educational Psychologist, 50*(1), 1–13.
Sung, H. Y., Hwang, G. J., Lin, C. J., & Hong, T. W. (2017). Experiencing the analects of confucius: An experiential game-based learning approach to promoting students' motivation and conception of learning. *Computers & Education, 110*, 143–153.
Sung, H. Y., Hwang, G. J., & Yen, Y. F. (2015). Development of a contextual decision-making game for improving students' learning performance in a health education course. *Computers & Education, 82*, 179–190.
Sweetser, P., Johnson, D., Wyeth, P., Anwar, A., Meng, Y., & Ozdowska, A. (2017). GameFlow in different game genres and platforms. *Computers in Entertainment, 15*(3), 1.
Sweetser, P., & Wyeth, P. (2005). GameFlow: A model for evaluating player enjoyment in games. *Computers in Entertainment, 3*(3), 3.
Tamborini, R., Bowman, N. D., Eden, A., Grizzard, M., & Organ, A. (2010). Defining media enjoyment as the satisfaction of intrinsic needs. *Journal of Communication, 60*, 758–777.
Tsai, M. J., Huang, L. J., Hou, H. T., Hsu, C. Y., & Chiou, G. L. (2016). Visual behavior, flow and achievement in game-based learning. *Computers & Education, 98*, 115–129.
Tuzun, H., Yılmaz-Soylu, M., Karakuş, T., Inal, Y., & Kızılkaya, G. (2009). The effects of computer games on primary school students' achievement and motivation in geography learning. *Computers & Education, 52*, 68–77.
Ventura, M., Shute, V. J., & Zhao, W. (2013). The relationship between video game use and a performance-based measure of persistence. *Computers & Education, 60*, 52–58.

Vos, N., van der Meijden, H., & Denessen, E. (2011). Effects of constructing versus playing an educational game on student motivation and deep learning strategy use. *Computers & Education, 56*, 127–137.

Whitton, N. (2011). Game engagement theory and adult learning. *Simulation & Gaming, 42*, 596–609.

Wouters, P., van Nimwegen, C., Oostendorp, H., & van der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology, 105*, 249–265.

Part II
Emerging Methods and Practices

Chapter 6
Examining Designed Experiences: A Walkthrough for Understanding Video Games as Performance Assessments

Michael P. McCreery, P. G. Schrader, S. Kathleen Krach, Jeffrey R. Laferriere, Catherine A. Bacos, and Joseph P. Fiorentini

6.1 Introduction

Empirical investigations of video games follow a few primary approaches. Typically, they examine: (1) consequences of gaming (e.g., learning from games; De Freitas, 2018), (2) interactions with games (e.g., from a human–computer interaction perspective; Fortes Tondello et al., 2018), or (3) learning within games as a situated context (Jabbari & Eslami, 2019). Broadly, the majority of learning-related video game literature tends to fall into one of four general categories: intervention studies (Stefanidis, Psaltis, Apostolakis, Dimitropoulos, & Daras, 2019), addiction studies (Mancini, Imperato, & Sibilla, 2019), learning studies (Wouters, Van Nimwegen, Van Oostendorp, & Van der Spek, 2013), or social interaction research (McCreery, Vallett, & Clark, 2015).

Although the breadth of work associated with learning and video games continues to develop, there is a dearth of examples on how to extract complex, dynamic, and emergent data using video game contexts. Similarly, there are limited examples that outline strategies and tools for interpreting game-based data. As such, the main purpose of this paper is to outline one possible process to use the complex environ-

M. P. McCreery (✉) · P. G. Schrader · C. A. Bacos · J. P. Fiorentini
University of Nevada, Las Vegas, Las Vegas, NV, USA
e-mail: michael.mccreery@unlv.edu; pg.schrader@unlv.edu; bacosc@unlv.nevada.edu; fiorej2@unlv.nevada.edu

S. K. Krach
Florida State University, Tallahassee, FL, USA
e-mail: skrach@fsu.edu

J. R. Laferriere
Lebanon Valley College, Annville, PA, USA
e-mail: laferrie@lvc.edu

D. Ifenthaler, Y. J. Kim (eds.), *Game-Based Assessment Revisited*, Advances in Game-Based Learning, https://doi.org/10.1007/978-3-030-15569-8_6

ment of a video game as a data collection tool. Readers should expect to exact a greater understanding of how data captured from observing video gameplay can be used in conjunction with path analytic techniques to elucidate the process of learning. Fundamentally, this work exposes strategies to leverage existing off-the-shelf video games as contexts for performance assessment.

6.2 Performance Assessments

There has been substantive effort to evaluate performance in video games as spaces for experiential learning (i.e., how game experiences impact learning; Anetta, Minogue, Holmes, & Cheng, 2009; Harvianinen, Lainema, & Saarinen, 2014; Shaffer, Squire, Halverson, & Gee, 2005; Squire, 2011). However, less research has been conducted on leveraging video games as encapsulated, performance assessments (i.e., how interconnected gameplay experiences influence outcomes). At their core, performance assessments are grounded in the principle that learning occurs within a situated or sociocultural context (Wang, Shute, & Moore, 2015). From this perspective, learners develop mental representations (i.e., schemata, scripts) as they interact with the world. Subsequently, those representations are called upon as heuristics to aid in decision-making processes (Govaerts, Van Der Vleuten, Schuwirth, & Muijtjens, 2007). Accordingly, the best way to assess performance learning is to ask the learner to demonstrate higher-order thinking and apply their conceptual understanding of the world in novel situations (Shavelson, Baxter, & Gao, 1993).

Typically, performance assessments are designed in ways that position the learner to: (a) perform a goal-oriented exercise that demonstrates success on a summative task, and (b) demonstrate understanding of the process or steps associated with its successful completion (Shavelson et al., 1993). This dual-oriented emphasis (i.e., goal-oriented performance from a process-oriented lens) serves to reveal the connection between higher-order thinking and conceptual understanding in novel situations. Consequently, performance assessments differ substantially from most traditional assessments, particularly multiple-choice tests. For example, items on multiple choice tests are generally designed to be independent of one another; items can be arranged in any order, and success on one item does not influence the success on subsequent items (Yen, 1993).

In contrast, performance assessments are defined in terms of item interdependence. In most cases, a setting (e.g., narrative) is first established and learners must make decisions within that narrative (Yen, 1993). Each decision has predefined and intentional dependencies that are linked to previous choices and early choices have implications for subsequent decisions. For example, some decisions may expose new options or limit choices. As such, specific decisions may be examined formatively; while collectively, the sum of those activities can be examined in the context of summative outcomes to provide meaningful insight into the overall process, degree, and nature of learning (Shute, Leighton, Jang, & Chu, 2016).

6.3 Video Games and Assessment

For decades, researchers have asserted that video games are rich tools and environments for the study of learning and related mechanisms (de Freitas, 2018). However, in recent years this work has expanded its focus to include the examination of process-oriented data (Schrader, McCreery, & Vallett, 2017). From this perspective, games provide access to behavioral and learning data that are dynamic, emergent, and complex. Researchers have argued that these process-oriented data have great potential to yield insight into learning as it evolves through gameplay. For example, Vallett (2016) described the dynamic process of acting and adjusting behavior to the environment as situated learning via "soft failure" (e.g., dying and restarting a level). Here, gameplay experiences act as a performance tuning mechanism (Schrader et al., 2017; Vallett, 2016). Each interaction within the system provides information and a potential source of data. Players must discern what information is useful and adjust their behavior accordingly. Failure is inevitable and when it occurs, the situation provides the player an opportunity to reevaluate the usefulness of the information, problem solve, and reattempt the action (Schrader et al., 2017; Vallett, 2016). Collectively, these data provide evidence of patterns of behavior during the learning process. As assessments, games offer more than a mechanism to examine performance through outcomes. Games provide new opportunities for researchers to collect, analyze, and interpret data during these experiences (Schrader et al., 2017).

Although it is often difficult to capture process-oriented data, games regularly monitor interactions within the environments and commonly collect data on player performance (Shute, Ke, & Wang, 2017). While these data are typically used to provide feedback and cues for players, the same data may be captured and used by researchers to provide unique and additional insights into variables associated with processes (Schrader et al., 2017). It follows from this perspective that although *summative* evaluation of performance is useful for many questions, the development of a meaningful *formative* understanding of learning through systematic observation and analysis of behaviors within a video game (e.g., game's cues and player's actions) adds numerous options to researchers' repertoire (Schrader et al., 2017).

By leveraging games as performance assessments that capture process data (i.e., data that are complex, dynamic, and emerge over time), researchers can look beyond the gameplay as a singular or aggregated experience to be observed. This subtle, yet important, shift augments the research perspective in a fundamental way by moving the focus from assessments characterized by success or failure, to understanding how higher-order thinking and the learner's conceptual understanding of the world informs connected outcomes (Schrader et al., 2017; Shute et al., 2017). With respect to games that provide a finite number of choices, the game structure is similar to a nested multiple-choice decision tree or flowchart. In this example, each decision relies on the previous one, and taken as a whole, performance can be characterized by the path that player takes coupled with the outcome (e.g., Tic Tac Toe, Othello, or a Moral Choice game). As noted earlier, each gameplay decision is interdependent with other decisions. By extension, play serves as an opportunity to document and capture dynamic, in-game interactions, link those interactions to formative activities, and then examine the ways in which those activities influence the overall goal.

With these ideas in mind, and because games differ significantly in their structure, affordances, and capabilities, we first outline the factors involved with evaluating a game's suitability (Schrader & McCreery, 2012). In particular, we focus on games that function as complex systems and produce data that are aligned to a process-oriented perspective (Schrader & McCreery, 2012). Second, we establish a heuristic for identifying data and their coding. Third, we explore analytic techniques that are appropriate to process-oriented data. In this case, we describe path analysis and its potential to elucidate how player interactions are tied to learning as an emergent, dynamic process. Throughout, methods for capturing, coding, and analyzing within-game data are described pursuant to this goal.

6.4 Game Selection

Researchers have described various reasons for selecting the specific video game contexts they study. In some cases, the environments are constructed as part of broader work (e.g., Quest Atlantis, River City, or Whyville). In others, selection criteria and rationale focus on game popularity or interesting interactions within the system (see Schrader & McCreery, 2008). Whatever the reason, game selection is a vital component of the research process. The game governs the types of affordances that are available to players, shapes the research questions, informs the types of data that can be collected, and impacts researchers' choice of designs and methods. When a dual-oriented emphasis (i.e., goal-oriented performance from a process-oriented lens) is adopted, game selection is even more important.

In general, all players' choices within games can be represented or mapped in some manner. For example, actions within open-world games, although vast and overwhelming, can be observed as classes, categories, or groups of actions that are based on the constraints and affordances of the game being investigated. By contrast, player decisions within moral-choice games (i.e., *The Deed*) are finite and can be mapped more easily. When represented visually, the decision structure is similar to a flowchart, in which each fork represents a choice or interaction within the game. Similar to a performance assessment, each fork provides the player with an opportunity to select an optimal or a suboptimal solution (i.e., correct or incorrect choice). As a result, these actions serve as isolated error checks, as well as a more holistic performance assessment that is readily quantified and analyzed. In this way, the format of the game provides an ideal platform to evaluate gameplay performance methodology; specifically, concrete data that are specific to the player's decision-making processes at every stage of gameplay.

In most games, the structures, models, algorithms, and rules within these systems are implicit. As a result, the deconstruction of the game model begins with an inductive process associated with extensive play or game experience (Schrader, Deniz, & Keilty, 2016). Essentially, researchers are encouraged to observe the various options for action and the constraints on action, particularly as they relate to the agency of: (a) players, (b) developers, and (c) researchers. Although there may be some overlap, the agency for players is often different than the agency for developers or researchers. For example, the ability to access command line input may be available to developers, but unavailable to players because they are intended to rely more heavily on visual stimuli. Collectively, player and developer affordances inform everything from the type of questions that are appropriate to opportunities for data collection. It should be noted that this process is focused on the potential for action and the constraints imposed on the system rather than the intentions behind either. For these reasons, the deconstruction of the game model is both reasonable and necessary; it provides a means to evaluate key design characteristics and affordances (e.g., narrative and gameplay mechanics) in relation to research suitability. This typically happens prior to game selection, but certainly before any empirical study commences.

Often, environments are selected because they are popular and/or have a set of features that give rise to interesting studies or player interactions. This means that research frequently involves commercial and publicly available software. Unfortunately, researchers do not usually have access to the design principles, guidelines, or gameplay diagrams. Similarly, it is very difficult to capture clickstream data, process data, or the "under-the-hood" mechanics due to the proprietary nature of commercial games. For researchers, this is a common scenario and often requires a labor-intensive scheme to extract and code data from the system. In this case, researchers identified, catalogued, and mapped all available actions within the game. This is a necessary step in quantifying key data for analysis.

6.5 Selecting the Deed

In the current example, *The Deed* (Grab the Games, 2015) was selected because of its structure, compelling story and plot, and alignment with guidelines for performance assessments (see Shute et al., 2017). The process of selecting *The Deed* followed the same approach identified above. Members of the research team identified the game as a potential candidate for research based on reviews and game descriptions. Subsequently, they played the game multiple times with an intent to identify the key elements of agency in the game based on what players might be able to accomplish through their experience, what developers intended, and how those two perspectives might inform research. Briefly, *The Deed* is a moral-choice role-playing murder mystery video game in which players' in-game decisions are limited in ways that are like a choose-your-own-adventure novel. There is a compelling social narrative that contextualizes a complex, puzzle-oriented game that focuses on the players' ability to reverse traditional moral roles. Unlike many other murder-mystery games, the objective of *The Deed* is to commit the act of murder (i.e., "the deed") and secure the family inheritance, rather than solve a crime that has been committed. The plot involves murdering the main character's own sister, framing another character for the murder, and ensuring that the main character avoids conviction for the crime. The plot helps shape players' decisions and social interactions, all of which result in a finite number of outcomes. More importantly, the social interactions with characters in the game allow players to unravel the clues to the social puzzle they are attempting to solve (e.g., interacting with characters, and the various weapon and evidence choices).

Similar to a play, the narrative of *The Deed* can be divided into five experiences: The Introduction and Four Acts. These acts include: (1) the homecoming (2) the dinner (3) the deed, and (4) the murder investigation and verdict. At the start of the game, the player has an opportunity to read the Introduction. This is the first learning opportunity for the player. If the player chooses to read the Introduction, they receive critical information that includes how to experience the game narrative, the importance of weapon and evidence selection (i.e., formative activities), and how planting evidence will impact the outcome (i.e., the summative outcome). Act One immediately follows the Introduction. Throughout this act, the player is given numerous learning opportunities to interact with characters (i.e., maid, butler, mother, father, and sister) and objects (i.e., weapons, evidence items, and story flashbacks). These interactions are intended to help players gain critical information to better develop problem-solving strategies. Moreover, they inform a set of formative tasks, including the successful (or not) selection of a weapon and an item of evidence that will be used to commit the deed and scapegoat another character for the murder. The player is given the choice to engage in these learning opportunities or to pass on them. However, in order for the player to move on to the second act, two items must be selected (i.e., a weapon and piece of evidence [correct response], two weapons, or two pieces of evidence [incorrect response]).

Act Two consists of a dinner celebrating the father's birthday. The player is seated at a table while interacting with other characters through a series of response options to statements made during the dinner conversation. Act Three is when the deed is committed; during this act, gameplay includes the formative tasks of successfully planting the evidence selected and using the weapon selected in Act One. The player has the option to forgo planting evidence and advance to committing the deed. However, not planting evidence is the only option if the player decided not to select an item of evidence during Act One (i.e., selected two weapons). Conversely, if the player decided not to select a weapon in Act One (i.e., selected two items of evidence), the only option is to commit the deed using the character's bare hands. Finally, in Act Four the murder investigation takes place. The player faces an investigator who has been called to the house. During the interview with the investigator, the player is questioned in relation to their prior decisions. In order to achieve a successful summative outcome (i.e., not going to prison), the player must succeed at each of the formative tasks presented throughout the narrative.

Ultimately, *The Deed* was determined to: (1) be a contextualized experience (i.e., social narrative); (2) provide clear linkages between choices (i.e., formative activities); and (3) be a goal-oriented exercise (i.e., summative outcome). In total, this game can take up to an hour to complete. For the purpose of research and assessment, this short time period is crucial (see Schrader et al., 2017). It may be unreasonable to use a game where players have different levels of expertise (McCreery, Schrader, & Krach, 2011), or that are overly time consuming given the purpose of the assessment (Kline, 2005). Collectively, these characteristics, evident in *The Deed*, provided researchers with access to, and the ability to assess, transactional learning experiences during gameplay in a situation that meets the added constraints (e.g., time, setting, replicability) that researchers often impose on design. In other words, learning experiences within *The Deed* are grounded in the interplay among the learner (i.e., player), context (i.e., narrative), and content (i.e., plot) (Moore, 1993).

Essentially, the game selected for this study was reverse engineered to understand the behind-the-scenes game mechanics that afford the range of player actions and outcomes in the game. Because *The Deed* involved a finite number of choices, the act of defining game elements and choices was somewhat straightforward. The selection and deconstruction process resulted in a *data dictionary* and *behavioral observation protocol* through which all gameplay data could be collected and analyzed.

6.6 Creating a Data Dictionary

Once the researcher has played the game, consumed other details and media, and deconstructed its mechanics, the next step is to define pertinent game elements. In some cases, this means observing general trends of players' interactions. For example, McCreery et al. (2015) created a matrix of observable behaviors that was based on Whiteside's model of social presence (Whiteside & Garrett Dikkers, 2012). The researchers then addressed questions related to players' interactions within a complex, dynamic, and emergent game (i.e., *World of Warcraft*) through cataloging observed behaviors in the game. By contrast to the open-endedness of the *World of Warcraft*, as well as many other games, *The Deed* includes a finite number of choices. Although there is no set pattern or pre-scripted path through the game, researchers were able to identify and define all game content. As a result, each opportunity for action and all player interactions were able to be tracked and analyzed. In this case, a detailed inventory of actions and interactions was appropriate because of the specific type of game originally selected. Below are the suggested steps of a game deconstruction process:

1. Identify all potential outcomes: go to prison (failure); get away with murder but no inheritance (partial success); get away with murder and gain inheritance (full success).
2. Identify the formative activities that must be accomplished in order to achieve a successful outcome: weapon selection, evidence selection, evidence planting.
3. Identify broad categories of in-game affordances that players can interact with in order to gain information necessary for problem-solving: non-player characters (i.e., computer controlled), weapons, evidence, flashback objects (e.g., painting on a wall that when interacted with provides narrative clues).
4. Identify all individual in-game affordances within each broad category (i.e., each character; weapon; piece of evidence; and flashback object).

The sum of all this information resulted in a *data dictionary*. In this example, a data dictionary outlined and defined key concepts, terms, ideas, and behaviors that were known to exist in the game. The data dictionary was created to provide the entire research team with consistent and shared understanding of game elements, features, mechanics, and play. Further, the data dictionary allowed the team to organize and categorize each of the game elements based on the constructs being analyzed and the variables being measured.

6.7 The Behavioral Observation Protocol and Coding Data

Once the essential elements of a game are defined and, in this case, categorized in a dictionary, the next step involves creating a resource for coding. For this example, a behavioral observation protocol was developed that included an array of important, observable player exhibited behaviors (i.e., it happened or it didn't) in order to limit qualitative inference. These behaviors were organized in ways that address the research question and its underlying theoretical framework. Moreover, whether researchers are mapping the game space in its entirety or a targeted set of behaviors (see McCreery, Krach, Schrader, & Boone, 2012 for an example), a behavioral

Evidence Interations

UID	E-LP	E-LT	E-MD	E-UG	Evidence Selected	Evidence Planted	Total
000	0	2	1	1	E-LT	0	4

LP (Love Poem); LT (Leather Tawse); MD (Mother's Diary); UG(Undergarments)

Character Interactions

UID	Maid	Butler	Mother	Father	Sister	Total
000	2	1	2	1	2	8

Weapon Interactions

UID	W-BH	W-BR	W-CS	W-FS	W-GL	W-KN	W-RO	W-RP	W-Q	W-SG	Weapon Selected	Weapon Used	Total
000	0	0	0	1	0	0	0	1	0	1	W-GL	W-GL	3

BH (Bare Hands); BR (Broom Handle); CS (Candlestick); FS (Fencing Sword); GL (Shard of Glass); KN (Knife); RO (Rope); Q (Pool Cue); SG (Shotgun)

Flashbacks

UID	F-CH-MO	F-GC-FA	F-MI-SI	Total
000	0	1	0	1

Trigger Object - Character Involved, CH-MO (Chair-Mother); GC-FA (Class Cabinet-Father); MI-SI (Mirror-Sister)

Fig. 6.1 Behavioral observation protocol example

observation protocol provides boundary conditions on the behavior that must be recorded and those that are not pertinent to the questions being answered (Alevizos, DeRisi, Liberman, Eckman, & Callahan, 1978; Milne, 2015).

The development of a behavioral observation protocol is an applied psychological approach to data collection that in the context of a video game entails two major steps. First, researchers begin by translating the elements of the data dictionary into a spreadsheet(s) that will become a comprehensive record of relevant player behaviors. This spreadsheet becomes a scorecard on which to record (i.e., tally) all of the observable behaviors, formative activities, and summative outcomes for each player. Behaviors must be operationally defined (e.g., specific, quantifiable, observable, concrete action) in order to ensure content validity and interrater reliability (Tapp, Wehby, & Ellis, 1995). Second, the protocol template is then generated for each player and distributed to the coders. The template then serves as a checklist for each coder to observe and record player behavior. For example, in Fig. 6.1, four types of interactions (i.e., evidence, character, weapon, and flashback) as defined during the creation of the data dictionary were translated into the behavioral observation protocol. Additionally, more specific interactions associated with interaction type (e.g., E-LP = evidence, love poem) are also defined. The coder can then record every time a player (represented by UID or user identification in the example) interacts with that specific element of the game.

The behavioral observation protocol was created to account for each of the possible interactions in *The Deed*. In Act One, the following player behaviors were recorded based on elements defined in the data dictionary: watching the introduction, dialogue with characters, story flashbacks viewed (i.e., objects in the story setting that when selected trigger a story flashback revealing more information about the other characters), weapons viewed and selected, and items of evidence viewed and selected. In Act Two, the dialogue with characters during dinner is coded in the same format as conversations in Act One. The dialogue checklist for the coder provides a listing of all the character statements and response choices to those statements. While viewing the video recording of the player's gameplay, the coder checks a box indicating the character interaction (e.g., spoke with the mother) and the response selected among the possible options listed for that character interaction (e.g., response choice 1, 2, or 3).

In Act Three, coders used a checklist to mark whether the player planted evidence selected in Act One, where the evidence was planted, and finally, what weapon was used to commit the murder. In Act Four, coders used a checklist to indicate responses to the crime investigator's interview questions. A checklist was also provided to coders to indicate one of the following outcomes: (1) the player was convicted of murder and sent to prison, (2) the player was not convicted of murder, or (3) another character was convicted of the murder because of the evidence planted against them, and the player received the inheritance.

6.8 Analytics of Gameplay

Once all the data from the player's gameplay is recorded, additional spreadsheets can be created for each of the constructs and related variables being measured as defined in the data dictionary. Further, because the nature of the data is a count (i.e., it happened or it didn't) interrater agreement in its true form, consistency of subject ratings is not needed (McHugh, 2012). However, for the sake of accuracy interrater data should be collected. In the present example, the coded spreadsheets for *The Deed* noted each interaction (exogenous variables) with weapons, story flashbacks, characters, and evidence items. The coded spreadsheet also noted the successful completion of each linked outcome (endogenous variables) across the game. Specifically, the variables coded as formative outcomes included: successful selection of a weapon and evidence item (Item Selection); successful planting of the evidence (Evidence Planted); and finally, the summative outcome, successfully get away with murder (Successful Outcome).

6.9 Analytic Techniques to Understand Player Experience

Using this process, data that are extracted from observations of players' behavior within *The Deed* are dynamic, emergent, and complex. It is common practice in low-dimensional, independent systems to test for significance using techniques like, t-test, ANOVA, MANOVA, etc. By contrast, complex systems involve increasing degrees of emergence and higher levels of dimensionality; this ilk of analyses is not very informative or useful. Fortunately, there exists a variety of analytic techniques that have the potential to expose patterns in data extracted from video games. For example, time series techniques, analysis of spline equations, structural equation modeling, and path analysis have been used with this class of data. It should be noted that each approach has distinct assumptions and each address different types of questions. For more details, please refer to Little, Bovaird, and Slegers (2006).

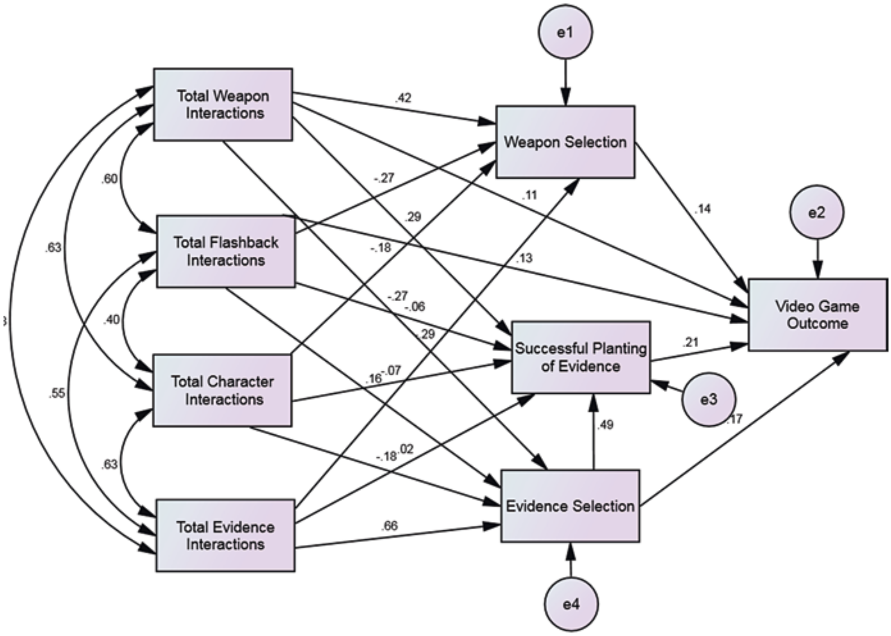


Fig. 6.2 Example path model

In this example, data were coded based on an event-dependent sample (as opposed to a time-dependent sample). Researchers employed path analysis to demonstrate causal effects among constructs in the game model: knowledge interaction, formative activities, and the summative outcomes. This form of analysis allows the researchers to link in-game *observable* information activities (emphasis added) directly with both formative and summative outcomes to better understand the process of learning. This process yielded a viable model (see Fig. 6.2) based on the relationships between the game constructs. While the details for this study are presented elsewhere (see McCreery, Laferriere, Bacos, & Krach, 2018), what should be noted is that the model illustrates that player outcomes are specifically related to the information acquired through interaction in the game space. For example, the more a player interacts with the available evidence (i.e., Total Evidence Interaction), the better is the understanding they appear to have in terms of the required Evidence Selection necessary to win the game. Alternatively, as a player increases their interaction with weapons (i.e., Total Weapon Interaction), the more likely those interactions become a distractor in terms of Evidence Selection necessary to win the game.

6.10 Discussion and Implications

The current work demonstrates the potential for video games to serve as unique and useful data-collection methods. By following the steps outlined in this chapter, researchers can extract data from complex contexts, in which players' choices can be represented or mapped. In the most general terms, researchers should plan carefully when deciding on the appropriate game to choose, how the game context allows for data collection of constructs of interest, and how the data can be collected in a psychometrically sound manner. Researchers are encouraged to plan for data collection in games from multiple lenses, perspectives, and levels. This includes whether it is appropriate to capture behavioral data. Moreover, if behavioral data are deemed appropriate, examine whether it is feasible to map the game space (e.g., *The Deed*) or does emergent gameplay (e.g., *World of Warcraft*) require a more targeted approach. Answers to these questions are critical as they will provide insight into the underlying mechanics and encapsulating contexts of games, and promote an increased understanding for the purpose of hypothesis generation, study design, data collection, data coding, and analytic approaches.

The example employed in this chapter (i.e., *The Deed*) is best characterized as a moral-choice game. By design, players are forced to make decisions in an attempt to achieve the game's main objective. From a limited point of view, the game is a finite collection of mappable choices that are either beneficial (right) or not (wrong). From this perspective, *The Deed* is structured in the same way as any performance assessment including: a contextualized narrative, goal-oriented summative outcome, and clearly linked formative activities. Moreover, unlike traditional multiple-choice tests, where each item is independent of one another and evaluated individually, in choice-based games, each decision is necessarily dependent upon the previous response. This suggests that there is an opportunity to examine choices at a discreet, individual level and also collectively as a whole. As a result, path analysis is the logical procedure to examine performance in these systems when overall performance, defined here to be the sum of all items is dependent upon one another.

Using this logic, information can be presented as a hint to aid the player or as distractor to lead them astray. Further, some choices could be considered correct answers (e.g., Evidence Selection), which are conducive to increased success. By contrast, distractor or error choices correspond with diminished success (e.g., the longer you examine your weapons choices, or Total Weapon Interactions, the less likely you are to experience success at the game). Ultimately, designers of *The Deed* presented information in three key ways: (a) there is information that is critical to success (e.g., information gained from interacting with pieces of evidence predicts the selection of evidence); (b) there is information that contributes to the atmosphere or narrative, but is not germane to the solution (e.g., interactions with flashback objects do not influence the selection of evidence); (c), there is information that is intended to distract and test your problem-solving ability.

Collectively, the manner in which the information is presented to the player and the heuristics that must be employed shift the focus of the experience away from a recall task to a situated performance assessment. Moreover, the fundamental structure of choice-based games and this process approach to capturing data, raise exciting possibilities for new forms of assessment. Future assessments could be designed to capture process data, rather than after the fact as presented here. There are several significant benefits to such a design: (a) it would provide researchers with a clearer understanding of how design elements impact the assessment (e.g., usability and psychometrics); (b) integrated data capture tools would limit resource expenditures (e.g., time coding data); and (c) provide a clearer manner in which to evaluate learning process discrepancies between actual and target learning.

Although the first two points are obviously important, the last one warrants additional discussion. Since the days of Dewey (1899), researchers and theorist alike have argued the importance of understanding learning as a process rather than solely an outcome. It is within the process that one can tease out misunderstanding, ineffectual problem-solving strategies, and misplaced heuristics. Game-based performance assessments may provide new opportunities to better understand how these issues arise. Specifically, a players' individual process model can be evaluated against the successful solution(s) in order to better understand where additional help should be given. This not only provides both teacher and learner with a more detailed understanding of where a problem(s) has emerged, but also discussion points to better understand both the *how and why* (emphasis added) choices were made.

References

Alevizos, P., DeRisi, W., Liberman, R., Eckman, T., & Callahan, E. (1978). The behavior observation instrument: A method of direct observation for program evaluation. *Journal of Applied Behavior Analysis, 11*(2), 243–257. https://doi.org/10.1901/jaba.1978.11-243

Anetta, L. A., Minogue, J., Holmes, S. Y., & Cheng, M. (2009). Investigating the impact of video games on high school students' engagement and learning about genetics. *Computers & Education, 53*, 74–85.

de Freitas, S. (2018). Are games effective learning tools? A review of educational games. *Journal of Educational Technology & Society, 21*(2), 74–84. Retrieved from http://www.jstor.org/stable/26388380

Dewey, J. (1899). *The school and society*. Chicago, IL: University of Chicago Press.

Fortes Tondello, G., Valtchanov, D., Reetz, A., Wehbe, R. R., Orji, R., & Nacke, L. E. (2018). Towards a trait model of video game preferences. *International Journal of Human–Computer Interaction, 34*(8), 732–748.

Govaerts, M. J. B., Van der Vleuten, C. P. M., Schuwirth, L. W. T., & Muijtjens, A. M. M. (2007). Broadening perspectives on clinical performance assessment: Rethinking the nature of in-training assessment. *Advances in Health Sciences Education, 12*(2), 239–260. https://doi.org/10.1007/s10459-006-9043-1

Grab the Games. (2015). *The deed* [Computer software]. Retrieved from http://store.steampowered.com/app/420740/The_Deed/

Harvianinen, J. T., Lainema, T., & Saarinen, E. (2014). Player-reported impediments to game-based learning. *Transactions of the Digital Games Research Association, 1*(2), 1–11. Retrieved June 14, 2018, from http://todigra.org/index.php/todigra/article/view/14/21

Jabbari, N., & Eslami, Z. R. (2019). Second language learning in the context of massively multiplayer online games: A scoping review. *ReCALL, 31*(1), 92–113.

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, CA: Sage.

Little, T. D., Bovaird, J. A., & Slegers, D. W. (2006). Methods for the analysis of change. In D. K. Mroczek & T. D. Little (Eds.), *Handbook of personality development* (pp. 181–211). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Mancini, T., Imperato, C., & Sibilla, F. (2019). Does avatar's character and emotional bond expose to gaming addiction? Two studies on virtual self-discrepancy, avatar identification and gaming addiction in massively multiplayer online role-playing game players. *Computers in Human Behavior, 92*, 297–305.

McCreery, M. P., Krach, S. K., Schrader, P. G., & Boone, R. (2012). Defining the virtual self: Personality, behavior and the psychology of embodiment. *Computers in Human Behavior, 28*(3), 976–983. https://doi.org/10.1016/j.chb.2011.12.019

McCreery, M. P., Laferriere, J., Bacos, C. A., & Krach, S. K. (2018). Video games as a model for assessing learning behaviors. In *American Educational Research Association Annual Conference*, New York, NY.

McCreery, M. P., Schrader, P. G., & Krach, S. K. (2011). Navigating massively multiplayer online games: Evaluating 21st century skills for learning within virtual environments. *Journal of Educational Computing Research, 44*(4), 473–493. https://doi.org/10.2190/EC.44.4.f

McCreery, M. P., Vallett, D., & Clark, C. (2015). Social interaction in a virtual environment: Examining sociospatial interactivity and social presence using behavioral analytics. *Computers in Human Behavior, 51*, 203–206. https://doi.org/10.1016/j.chb.2015.04.044

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. Retrieved from https://hrcak.srce.hr/89395

Milne, D. (Ed.). (2015). *Evaluating mental health practice (psychology revivals): Methods and applications*. London, UK: Routledge.

Moore, M. G. (1993). Theory of transactional distance. In *Theoretical principles of distance education* (Vol. 1, pp. 22–38). London, UK: Routledge.

Schrader, P. G., Deniz, H., & Keilty, J. (2016). Breaking SPORE: Building instructional value in science education using a commercial, off-the shelf game. *Journal of Learning and Teaching in Digital Age, 1*(1), 32–46. Retrieved from http://joltida.org/index.php/joltida/article/view/6/54

Schrader, P. G., & McCreery, M. (2008). The acquisition of skill and expertise in massively multiplayer online games. *Educational Technology Research & Development, 56*(5–6), 557–574.

Schrader, P. G., & McCreery, M. (2012). Are all games the same? Examining three frameworks for assessing learning from, with, and in games. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 11–28). New York, NY: Springer.

Schrader, P. G., McCreery, M. P., & Vallett, D. (2017). Performance in situ. In Y. K. Baek (Ed.), *Game-based learning: Theory, strategies and performance outcomes* (pp. 421–438). Hauppauge, NY: Nova Science Publishers.

Shaffer, D. W., Squire, K. R., Halverson, R., & Gee, J. P. (2005). Video games and the future of learning. *Phi Delta Kappan, 87*(2), 104–111. https://doi.org/10.1177/003172170508700205

Shavelson, R., Baxter, G., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215–232. Retrieved from http://www.jstor.org/stable/1435044

Shute, V. J., Ke, F., & Wang, L. (2017). Assessment and adaptation in games. In P. Wouters & H. van Oostendorp (Eds.), *Instructional techniques to facilitate learning and motivation of serious games*. Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-39298-1_4

Shute, V. J., Leighton, J. P., Jang, E. E., & Chu, M. W. (2016). Advances in the science of assessment. *Educational Assessment, 21*(1), 34–59.

Squire, K. D. (2011). *Video games and learning: Teaching and participatory culture in the digital age*. New York, NY: Teachers College Press.

Stefanidis, K., Psaltis, A., Apostolakis, K. C., Dimitropoulos, K., & Daras, P. (2019). Learning prosocial skills through multiadaptive games: A case study. *Journal of Computers in Education, 6*, 167–190.

Tapp, J., Wehby, J., & Ellis, D. (1995). A multiple option observation system for experimental studies: MOOSES. *Behavior Research Methods, Instruments, & Computers, 27*(1), 25–31.

Vallett, D. (2016). Soft failure in educational and learning games. In R. Lamb & D. McMahon (Eds.), *Educational and learning games: New research*. New York, NY: Nova Science.

Wang, L., Shute, V., & Moore, G. R. (2015). Lessons learned and best practices of stealth assessment. *International Journal of Gaming and Computer-Mediated Simulations, 7*(4), 66–87.

Whiteside, A. L., & Garrett Dikkers, A. (2012). Using the Social Presence Model to maximize interactions in online environments. In K. S. Amant & S. Kelsey (Eds.), *Computer-mediated communication across cultures: International interactions in online environments*. Hershey, PA: IGI Global.

Wouters, P., Van Nimwegen, C., Van Oostendorp, H., & Van der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology, 105*(2), 249–265. https://doi.org/10.1037/a0031311

Yen, W. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Chapter 7
Press Play! How Immersive Environments Support Problem-Solving Skills and Productive Failure

Benjamin Emihovich, Logan Arrington, and Xinhao Xu

7.1 Introduction

Over the past decade, education researchers have explored how video games and immersive environments can support learning and assessment known as game-based learning (GBL). While well-designed video games are engaging and fun, there are challenges in producing valid and reliable assessment measures in games without disrupting the flow of the gameplay experience (Van Eck, Shute, & Rieber, 2017). In addition, there are also challenges in being able to produce valid and reliable assessments that ensure accuracy between what is being measured, and what is intended to be measured in a study. However, the challenge of addressing confounding constructs and ensuring construct validity can be alleviated by using an existing assessment framework in this field of research.

One possibility is using an evidence-centered design (ECD) framework that allows researchers to make valid inferences about the types of competencies (problem-solving processes) learners acquire during gameplay and the behaviors that provide evidence to validate claims made about the competencies (Mislevy, Steinberg, & Almond, 2003; Shute, Hansen, & Almond, 2008). ECD-based assessments are valid for gaming research since players are active learners and learning

B. Emihovich (✉)
University of Michigan-Flint, Flint, MI, USA
e-mail: bemihovi@umich.edu

L. Arrington
University of West Georgia, Carrollton, GA, USA
e-mail: logana@westga.edu

X. Xu
University of Missouri, Columbia, MO, USA
e-mail: xuxinhao@missouri.edu

D. Ifenthaler, Y. J. Kim (eds.), *Game-Based Assessment Revisited*, Advances in Game-Based Learning, https://doi.org/10.1007/978-3-030-15569-8_7

through gameplay is situated in an authentic context (Shute & Emihovich, 2018). Video gameplay can produce rich data on how learners demonstrate motivation, persistence, and spatial ability, which are important skills that are not suitable for traditional assessment measures. One example of a game-based assessment (GBA) is stealth assessment, where assessments are embedded directly into the game environment without disrupting gameplay for the learner (Shute, 2011). This type of assessment can also be adapted to address assessment of failure to learn during gameplay.

The idea of productive failure stems from the thought that there are four outcomes associated with any learning and performance activity (Kapur, 2016; Kapur & Rummel, 2012). These outcomes are productive failure, productive success, unproductive failure, and unproductive success. Failure or success refers to short-term outcomes by the learner, usually through initial performance. Whether an item is productive or unproductive is based upon whether meaningful long-term learning takes place. Some researchers have indicated that failure in an initial activity leads to better learning in the long term, if learners have an opportunity to receive guidance or feedback afterward (Kapur, 2016). There are various methods to trigger a productive failure outcome; however, the most traditional is to provide students with a complex ill-structured problem prior to instruction on a topic. They then consolidate their knowledge through some form of feedback or instruction. If productive failure outcomes are considered superior to productive success outcomes, as argued by Kapur (2016), then there is great import for verifying these outcomes in games. Games naturally lend themselves to productive failure outcomes.

In addition to producing valid and reliable assessments in GBL, GBAs need to address challenges in immersive virtual reality (VR) environments. VR technologies elevated the nature and possibilities of GBL, bringing unique research opportunities in instruction, training, and assessment. Pivotal components in a VR environment include a shared space, inhabitants/avatars, interactions with peers and the environment, and perceptions, experiences, and interpretations of the users/players. In VR environments, educators and trainers can design gamified scenarios for their teaching and training purposes in settings that are either impossible to replicate or too costly in the natural world.

In this chapter, we review the relevant literature on immersive environments and authentic learning that takes place during gameplay and can support the development of problem-solving skills. We also address how ECD can be adapted to fit emerging methods and practices in GBA, such as stealth assessment of problem-solving skills and challenges with measuring productive failure in GBL. The chapter will conclude with a discussion on how to address challenges of producing valid and reliable assessments in immersive environments for learning, with implications for future research.

7.1.1 *Authentic Learning and Immersive Gameplay*

Immersive environments can provide interactive learning experiences that are grounded in sound learning principles such as feedback, rewards, and authentic problem-solving scenarios to foster longer-term learning (Shute, Rahimi, & Emihovich, 2018). Gee (2005) argues that well-designed video games provide players with meaningful interactions in immersive environments where they can explore issues of identity, culture, politics, and values, which are not usually experienced by players outside of the game environment. Immersion enhances learning in the following ways: allowing multiple perspectives for players to interact with content, providing situated learning experiences, and engendering transfer (Dede, 2009). Dede (2005) suggests that immersive environments (IEs) support situated learning with knowledge distributed across a community among novices and experts, fluency in multiple media, and authentic problem-solving scenarios that engage multiplayer interactions among players and artificial characters that differ in knowledge, skills, and abilities. Multiple perspectives in immersive learning allow for egocentric and exocentric frames of reference meant to support motivation and promote abstract insights from a distance (Dede, 2009). The immersive gameplay associated with well-designed video games to promote cognitive competencies like problem-solving skills is grounded in situated learning (Van Eck et al., 2017). Situated learning takes place in the same context in which it is applied, and learning is a social process whereby knowledge is co-constructed among members in a community (Kirk & MacPhail, 2002; Lave & Wenger, 1991).

The immersive environments discussed in this chapter emphasize learning as an active social process where knowledge is co-constructed and distributed. Constructivism states that effective learning occurs when learners explore, collaborate, and interact with tools, resources, the environment, and people (Vygotsky, 1978). Situated learning views cognition as a process that occurs within each activity, context, and culture in situ. During gameplay in IEs, the learner dictates the pace, and personalizes the learning process by actively participating in an authentic environment. As players progress through a game, they assimilate and accommodate new knowledge structures by encountering and defeating progressively more difficult problem-solving scenarios. These gameplay scenarios often require players to use tools and resources efficiently and effectively: a facet of problem-solving skill (Shute, Ke, & Wang, 2017).

7.1.2 *Problem-Solving Skills and Stealth Assessment*

Problem solving is a cognitive process that requires planning, lateral thinking, and reflection to find a solution that is not known to the problem solver (Mayer & Wittrock, 2006). Problem-solving skills are important for lifelong development, but there is a gap in problem-solving skills acquired through formal learning settings and recent college graduates' preparedness when solving problems. Employers in the public and private sectors report that only 38% of recent college graduates that are hired can analyze and solve complex problems in the workplace environment (Hart Associates, 2018). Schools tend to focus on instruction that features well-designed problem-solving scenarios, whereby there is a knowable solution and solution pathway to solve the problem (Jonassen, 1997). More meaningful types of problem-solving scenarios engender critical thinking and promote transfer by requiring students to address ill-defined or ill-structured problems that often have no clear or knowable solution (Jonassen, 2000). Unlike the well-structured problems that students face in formal learning settings, well-designed games provide students with challenging scenarios in immersive gameplay that includes ongoing feedback for the players to hone their problem-solving skills over time (Van Eck et al., 2017).

Scholars who support GBL argue that problem-solving skills are a benefit of video gameplay through situated learning (Eseryel, Law, Ifenthaler, Ge, & Miller, 2014; Gee, 2008). While some researchers have attempted to design specific games to promote problem-solving skills (Van Eck, Hung, Bowman, & Love, 2009), there are various commercial video games on the market, *Portal 2* and *Plants* vs. *Zombies 2*, which can support the development of problem-solving skills (Shute, Ventura, & Ke, 2015; Shute, Wang, Greiff, Zhao, & Moore, 2016). These immersive environments can offer valuable assessment by providing students with repeated practice of problem-solving scenarios during gameplay. This requires players to analyze givens and constraints, which are facets of problem-solving skill (Shute & Emihovich, 2018). The challenge is being able to assess learners' problem-solving skills without interrupting the gameplay experience.

Recent studies in GBL have indicated learning gains in engagement, gameplay, and enjoyment of gameplay in addition to self-regulated learning and problem solving (Fong, Jenson, & Hebert, 2018; Taub, Azevedo, Bradbury, Millar, & Lester, 2018). Yet, the aforementioned researchers in each study acknowledge difficulty in assessing student learning from gameplay as a study limitation. Scholars agree there are challenges associated with assessing confounding constructs of gameplay and problem-solving skills in game-based research (DiCerbo, Shute, & Kim, 2017; Shute & Emihovich, 2018).

One way this problem can be alleviated is by using a GBA framework that is grounded in ECD. Stealth assessment is a framework that embeds assessments directly within immersive environments by: (1) defining claims made about targeted competencies; (2) linking evidence to problem solving during gameplay to validate claims; and (3) defining the tasks that generate data to elicit performance (Shute, 2011). This framework allows researchers to make valid inferences about student performance during gameplay without causing a disruption of the gameplay experience that can hinder the benefits of immersion. As an example, Shute et al. (2017) implemented stealth assessments in *Use Your Brainz* (a modified version of the original game *Plants* vs. *Zombies 2*). In this game, stealth assessments were woven directly into gameplay to assess middle-school students' problem-solving skills. During gameplay, players were tasked with defending their gardens against zombies by planting flowers that repel the zombie onslaught. Each player generated resources by planting sunflowers on the map and this in turn produced sun power, allowing players to summon plants that can defend the garden. Some of the plants that could defend against zombies included the pea shooter, which fires rapidly at zombies; the snow pea, which slows zombies from advancing toward the garden; and the walnut, which acts as a barrier that zombies must eat before reaching the player.

As research participants play video games, their performance is captured in data logs where behavioral indicators that elicit application of targeted competencies update assumptions made about the competency model. The researchers developed a competency model that defines problem-solving-skills during gameplay and behavioral indicators that provide evidence for each facet of problem-solving skill. The model included four facets of problem-solving skill, based on extensive review of the literature, hours of their own gameplay, and viewing expert solutions on various social media platforms. The four problem-solving facets in the model included: (1) analyzing givens and constraints; (2) planning a solution pathway; (3) using tools effectively and efficiently when implementing solutions; and (4) monitoring and reflecting progress (Shute et al., 2017). In addition to the competency model, the researchers also established criterion of behavioral indicators through gameplay that provided evidence for each facet of problem-solving skill. The behavioral indicators included, for example, planting snow peas behind walnuts, and planting sunflowers in the back of the map.

The problem-solving model was implemented in the game using Bayesian networks. The results of this analysis from the experiment generated data from the competency model, which reflects changes in what students know and can achieve in immersive environments. The ECD framework approach of stealth assessment connects assessment tasks during gameplay with claims made about student competencies to validate arguments made about student performance on targeted competencies during gameplay (DiCerbo et al., 2017). Shute et al. (2017) were able to validate stealth assessment by collecting data from students after 3 h of gameplay including two external problem-solving measures (MicroDYN and Raven's Progressive Matrices). Results demonstrated that stealth assessment estimates of problem-solving skills were significantly correlated with both problem-solving measures, helping establish construct validity of the assessment. For further examples and descriptions of scoring and Bayes Nets (BNs) in stealth assessment, see Shute et al. (2015, 2016, 2017). In the next section, we discuss some of the challenges with assessing productive failure in GBL. Under certain conditions, productive failure allows learners to struggle, persist, and even fail at problems that are ill-structured, but with the long-term goal of helping learners developing lateral thinking to solve authentic problems (Abrahamson & Kapur, 2018).

7.2 Productive Failure

Well-designed games can provide scenarios where students can fail. Additionally, data on their failures (i.e., their solutions to the problems faced) can be instantly collected by the game, whereas an application of strategies to produce this outcome in other learning environments results in data that cannot be collected as quickly. Thus, it is important for game-based environments to assess learners' initial efforts and help identify prominent methods of consolidation through this assessment. Embedded within well-designed games are authentic learning activities based on real-world contexts when designers, instructors, and/or learners are restricted by logistical limitations. Additionally, well-designed games in virtual immersive environments can act as a sandbox for learners to explore and encounter initial shortcomings in performance in order to grow over time. However, through feedback provided in the game's internal mechanisms or other types of methods (e.g., trial and error, help seeking) learners can enhance their long-term learning. This outcome is known as productive failure.

As mentioned above, productive failure stems from the four potential outcomes at the intersection of short-term performance and long-term learning (i.e., productive failure, productive success, unproductive failure, and unproductive success). Productive failure refers to an instance where learners encounter a failure in short-term performance, which leads to a more meaningful long-term learning experience. As an instructional strategy, productive failure is broken down into two phases, *exploration* and *consolidation* (Kapur & Bielaczyc, 2012). In exploration, learners face a challenging problem that elicits various opportunities for deep exploration of the problem and offers the opportunity for learners to create multiple solutions. The problem that is used in this phase should be within the learners' grasp (i.e., not frustratingly difficult) but still complex. Within most studies on this topic, ill-structured problems are used (Kapur, 2008). Ill-structured problems lack clear solutions, present excess or insufficient information for developing a solution, or have multiple processes for developing a solution (Jonassen, 1997). The most important component for productive failure is that the problem must allow learners to generate multiple solutions to the problem (Kapur, 2016; Kapur & Bielaczyc, 2012).

The content of the problem does not have to be overly complicated, but the problem needs to be ill-structured. For example, several the studies have investigated math concepts, specifically variance (e.g., Kapur, 2012; Loibl & Rummel, 2014b). In addition to the design of the problem being ill-structured, the problem should build upon learners' prior knowledge (Kapur & Bielaczyc, 2012). However, researchers debate the specifics of the needed prior knowledge. Originally, Kapur and Bielaczyc (2012) argued that the required prior knowledge was twofold, including content knowledge and knowledge of solving similar problems. However, in a later study, Toh and Kapur (2017) found that providing learners with specific micro-level instruction related to the content of the problem did not improve student learning. They did find that these students were capable of generating more solution attempts to the problem than their counterparts who did not receive the instruction. In addition to the design of the problem, many researchers have traditionally considered the exploration phase of productive failure as a collaborative problem-solving opportunity (Kapur, 2008; Kapur & Bielaczyc, 2012). The majority of productive failure research has examined the exploration phase as a collaborative effort (e.g., Kapur, 2008, 2009; Kapur & Kinzer, 2009; Loibl & Rummel, 2014a, 2014b; Westermann & Rummel, 2012). A number of the studies have investigated the exploration phase as an individualized problem-solving effort (e.g., Kapur, 2014, 2015; Mazziotti, Loibl, & Rummel, 2015). Mazziotti et al. (2015) aimed to determine if collaboration affected learning by comparing students solving problems in groups and individually. They found no significant difference between students collaboratively or individually solving problems.

A final consideration in the design of the exploration phase is the learning environment. Kapur and Bielaczyc (2012) identified the need for an environment that not only welcomes failure but also encourages it. While there is not one clear effective design prescription for creating the exploration phase, the intent of all approaches remains the same. The exploration phase should prime the learners to receive instruction in the subsequent phase, consolidation.

At the conclusion of the learners' problem-solving attempts (i.e., exploration), they must consolidate the knowledge they generated throughout this phase. This consolidation experience can come in many forms, but it should directly address elements of the overarching problem that the learners attempted to solve during the previous phase (Kapur & Bielaczyc, 2012). In most cases, teacher-led instruction in some form is used to help the learners refine their knowledge generated solving problems. Some researchers have investigated the focus and content of the instruction. Loibl and Rummel (2014b) explored whether instruction focusing on contrasting solutions (i.e., typically generated and the canonical solutions) to the problem used in the exploration phase would be a more effective consolidation experience than general instruction on the topic. They found that the former led to a much higher conceptual understanding on the topic than the latter approach. Another component that must be included in this phase is the opportunity for the learners to engage with the material (Kapur & Bielaczyc, 2012).

Lastly, the consolidation phase should continue the similar atmosphere introduced in the exploration phase (i.e., a safe place to fail). The emphasis during this phase is on how the solutions generated in the previous phase relate to the overall concept or solution to the problem and not that the learners made errors (Kapur & Bielaczyc, 2012). These two phases combine to create productive failure learning experiences. The design of each of these phases must complement the other as each has meaning in productive failure. Thus, in games, certain aspects should be assessed during each of these phases. Below, we briefly highlight some important characteristics that fit within game design and assessment for each of these phases.

7.2.1 What Do We Want to Assess During the Exploration Phase?

The exploration phase of productive failure is present in most game design. Players are presented with a problem that they must solve. Depending on the complexity of the problem and the available manipulatives within the game, the players are allowed to approach a problem from multiple avenues. In these multiple avenues, the players are allowed to generate multiple solutions to the problem. In addition, learners can attempt to solve the problem in a safe, low-consequence environment.

Typically, in productive failure studies there are a number of aspects assessed during the exploration phase. The most common is the number of solutions generated (also referred to as representations and solution methods). Various studies on productive failure have examined the impact of the quantity (Kapur, 2010, 2012, 2014, 2015; Kapur & Bielaczyc, 2012) and quality (Loibl & Rummel, 2014a). In most studies, there was a positive relationship between the number of solutions generated and learners' knowledge gains (Kapur, 2012, 2014, 2015; Kapur & Bielaczyc, 2012). This component is easily measured in game-based approaches as the tool itself can log these solutions. Additionally, researchers have investigated the impact of the problem-solving process on learner's cognitive load. Unsurprisingly, when learners are generating multiple solutions to a problem, they report higher cognitive load (Glogger-Frey, Fleischer, Grüny, Kappich, & Renkl, 2015; Kapur, 2013, 2014) during the exploration phase. However, the research is inconclusive as the higher cognitive load did not influence the learning gains in some research (e.g., Kapur, 2013, 2014), while in others there was a negative effect with higher cognitive load (Glogger-Frey et al., 2015). Additionally, affective variables are measured in the exploration phase. These variables include engagement and confidence. Typically, levels of engagement remain similar across groups during exploration (Glogger-Frey et al., 2015; Kapur, 2012, 2013, 2014). Learner's confidence can be low during the initial problem-solving attempts; however, the results on whether or not this impacts learning is inconclusive. The cognitive load imposed by the task and these affective aspects taken together are valuable variables to consider as learners solve problems in game-based environments.

7.2.2 What Do We Want to Assess During the Consolidation Phase?

In many game-based environments, as players make multiple attempts to solve the problem, some version of feedback is presented or offered to the learners. This feedback can come in various formats. In productive failure, the most common method of consolidation is delivered via teacher-led instruction. The content of the instruction may be generic or focused specifically on the most correct solution to the problem presented during exploration (Loibl & Rummel, 2014b). An important distinction should be made here in that in many games, feedback can occur too soon based on the design before learners have enough of an opportunity to explore other solutions. The success of a productive failure approach is another indicator that feedback could be delayed or based on the number of attempts in these environments.

Typically, affective variables are measured during, or in response to, the consolidation phase. The learners' engagement during consolidation, or their satisfaction with the consolidating experience, is measured. From a GBA perspective, the learners' incorporation of feedback can be easily measured as due to the instantaneous nature of assessment. The assessment of learning traditionally takes place during or after this phase. In most productive failure literature, assessment has included comprehension questions and solving of more structured problems. However, GBA methods allow for a more instant assessment of learning.

7.2.3 Productive Failure in GBL

While most of the literature in productive failure has focused on traditional educational contexts (i.e., online learning or face-to-face learning in a structured environment), some recent literature has incorporated the idea of productive failure within explaining the outcomes of their research. Anderson, Dalsen, Kumar, Berland, and Steinkuehler (2018) found that middle-school students who encountered more failures before succeeding were likely to learn more on the topic of virology. Additionally, these failures prompted discourse among students. While the researchers did not intend to design a productive failure experience within the game, their study is an indicator of how a game can capture generated solutions during the exploration phase. While the authors did not identify a consolidation phase within the game, the learners' discourse generated by their failed attempts fulfilled this role. In a similar investigation, Jagust, Boticki, and So (2018) found that more incorrect attempts by students within the game led to better learning outcomes in arithmetic. However, in this case the learners had received instruction prior to their explorations in the game.

Whereas the previous researchers (i.e., Anderson et al., 2018; Jagust et al., 2018) were identifying post hoc occurrences of productive failure, Gauthier and Jenkinson (2018) differentiated their approach by using productive failure as a desirable design element. They used the term productive negativity to represent a gameplay loop (i.e., a re-exploration within a certain parameter of the game, or a productive failure experience). Their qualitative analysis of a game versus simulated environment identified three design components that could affect game design and assessment of productive failure experiences. First, the authors found that restricting components of the game before introducing others limits their future explorations. However, the explorations can still be productive. Second, the authors found that incorporating more mechanics in one iteration led to more exploration within that loop. Lastly, they found that the key to integrating productively negative experiences was to integrate more mandatory variables/elements into the game than unnecessary ones. Overall, these design components explain how games can be designed to elicit productive failure experiences, thus allowing for the assessment of the productively negative (i.e., problem-solving iterations) experiences.

Scholars are beginning to notice the applications of productive failure for game-based learning. Due to the authentic challenging experiences games can provide, the opportunity to fail in a safe environment, the immediacy of data collection, and the control of feedback, there is a clear benefit to considering these types of outcomes within assessment. Additionally, there is an added benefit of assessing productive failure as it presents authentic problems in authentic environments, such as virtual reality enabled environments. Advancements in recent funding of VR technology has led to exciting developments in the field of training, research, and instruction (Shute et al., 2018).

7.3 Virtual Reality and Assessment

Computer games may take place in an immersive VR environment. VR technologies have elevated the nature and possibilities of learning games to another level, bringing unique research opportunities in instruction, training, and assessment. Major components in VR environments for education (VRE[2]) include a shared space, inhabitants/avatars, interactions with peers and the environment, and perceptions, experiences, and interpretations of the users/players. In VR environments, educators and trainers can design gamified scenarios for their teaching and training purposes in settings that are either impossible to replicate or too costly in the natural world. With the fast advancement of VR technologies in recent decades, researchers and practitioners have been applying VR in various instructional and training settings, such as medical training, professional simulations, and school education (e.g., Andersen, Konge, & Sørensen, 2018; Chang & Weiner, 2016; Cho et al., 2013; Ke, Lee, & Xu, 2016; Leder, Horlitz, Puschmann, Wittstock, & Schütz, 2019; Nagendran, Gurusamy, Aggarwal, Loizidou, & Davidson, 2013; Smith & Hamilton, 2015; Sugden et al., 2012; Tiffany & Hoglund, 2016). The infinite research possibilities that serious games in VR can provide also invite challenges in the areas of assessment and evaluation being studied.

7.3.1 Challenges with Assessment in VRE^2

In the context of this chapter, we argue that assessment in VRE^2 consists of two aspects: (1) to assess affective and learning outcomes of VRE^2 to exam benefits or drawbacks that it brings to participants; and (2) to assess VRE^2 itself to evaluate the gaming/playing elements and how the VRE^2 designs fit educational goals. Most existing studies in VRE^2 concentrate on the actual outcome of the affective domain and learning effects of the participants. While acknowledging the promising educational benefits that VR can bring, some researchers also placed doubts on whether VRE^2 could actually deliver learning effects in favor of content knowledge acquisition and application (Hew & Cheung, 2010). Recent VRE^2 brings onboard the head-mounted displays (HMD), for example, Oculus Rift and HTC Vive, to offer participants more immersive in-world experience. However, recently some researchers have found that other than psychomotor, visual, and special skills acquisition, such HMD-enabled VRE^2 do not necessarily lead to advantages over traditional instructional methods, and in some situations, might bring obstacles to learning task accomplishment (Jensen & Konradsen, 2018; Leder et al., 2019; Richards & Taylor, 2015). Possible reasons are technical challenges in such VRE^2, added cognitive load, distractions of fancy gaming experience, and even cybersickness. To minimize, if not resolve, the influence of such obstacles, it is not enough to apply only post hoc assessment to the VRE^2. It will be crucial to implement assessment while designing and playing/operating the VRE^2. We advocate the following directions and challenges for VRE^2 assessment.

VRE^2 features a dynamic system with numbers of elements related to usability, playability, and learning integration. Generally, usability concerns the interface, control mechanism, and technology used. Playability often relates to game challenge, task, enjoyment, and rewards. Learning could be integrated to any elements of usability, playability, and certainly the VR gameplay scenarios. Modern VRE^2 features multi-thread design and offers dynamic scenarios. Participants do not normally follow a linear approach in the gameplay but will interact with the system dynamically. Their learning processes, emotions, gameplay time, and accomplishments are echoed in their in situ behaviors while interacting with the usability and playability elements.

VRE^2 can be a sound platform for assessment itself. Clarke-Midura and Dede (2010) pointed out that conventional means of assessment, like multiple choice tests, could not fully reveal the learning of inquiry skills for students. The researchers further argued that a virtual environment can evaluate scientific inquiry skills of students through gamified and simulated activities in the virtual world. Compared with conventional tests, such carefully designed activities were more controllable and achievable in a virtual environment, and the actions and movements of a student were easily captured with the help of computer systems for further analysis (Clarke-Midura & Dede, 2010). Given the recent development of consumer-level VR devices that may bring more immersive experience to users, researchers have been studying and applying VRE^2-enabled technologies for assessment across disciplines (McGrath et al., 2018; Passig, Tzuriel, & Eshel-Kedmi, 2016).

7.3.2 Analysis and Measurement in VRE^2

Not every element can be measured directly. Such elements without direct measurement are reflected and externalized by game designs, technology affordances, and user performances. There may also be objective measures and subjective judgment for each element. For example, for game usability, once a VRE^2 is designed, the layouts and control features are normally fixed. However, different users may have their own experience interacting with the features. How to assess the game usability may vary depending on what we are interested to know. If we want to assess the human–computer interaction characteristics, objective measures like the game logs faithfully reflect game controls, time stamps, routes of exploring, or even participants' physiological information. In VRE^2 training, science lab and military combat for example, physiological information of electroencephalogram, haptic feedback, heart rate, blood oxygen saturation, blood pressure, and breathing behaviors is collected through wearable devices (e.g., Makransky, Terkildsen, & Mayer, 2019; McGregor, Bonnis, Stanfield, & Stanfield, 2017). On the other hand, if we are interested in user experience regarding usability, the assessment instruments of surveys, interviews, and focus groups are sound choices. Researchers apply both qualitative and quantitative analysis methods to data collected.

The latest trend is to analyze the data in an ad hoc and real-time manner, especially for the quantitative data collected instantly while the VR environment is running. This can be seen with the ECD approach of stealth assessment (Shute et al., 2016, 2017). With the advancement of big data and the increases in computing power, some educational researchers and scholars utilize approaches in artificial intelligence, pattern recognition, and machine learning to analyze the data collected (e.g., McGregor et al., 2017; Stanica, Dascalu, Bodea, & Moldoveanu, 2018). Creating a VRE^2 for job interview practice, Stanica et al. (2018) implemented chatbots with artificial intelligence, facial detection techniques, and semantic analysis while the mock-interview was running. Such real-time assessment may help the VRE^2 system to accommodate each individual participant with a personalized training road map and to detect their instant emotions to some extent. It may not only elevate individual learning experience, but also maximize the learning outcomes because each participant's VRE^2 experience is uniquely tailored to its best.

For learning outcomes like knowledge acquisition from VRE², most researchers apply the same assessment as that of the traditional instructional methods (normally as control groups), for example, paper-pencil tests, or online questionnaires. While acknowledging the effectiveness of such assessment, we also call for creative assessment that fits the intervention. For example, in some VRE² in which embodied features are implemented, participants can use body movements or gestures to interact with the VR learning scenarios. Since such body movements are part of the learning modality (Macedonia & von Kriegstein, 2012; Xu & Ke, 2014), it may be "unfair" to exclude the embodied part in the learning acquisition assessment. Assessment more closely aligned with the approach to acquiring content knowledge may reveal different learning outcome (Johnson-Glenberg & Megowan-Romanowicz, 2017; Nathan & Walkington, 2017). In a recent study in which some participants utilized embodied interactions as their major modality to learn in a VRE², a gesture-based test signaled results that were in favor of those participants who mainly used gestures in the virtual game compared with those who utilized fewer gestures when learning (Johnson-Glenberg & Megowan-Romanowicz, 2017). The study results imply that the format of assessment may play an important role in measurement within the context of VRE², and researchers and practitioners are encouraged to design assessments that will accommodate the actual learning experience of the users.

7.4 Future Implications for Game-Based Assessment

In this chapter, we have discussed how immersive environments for learning support the development of cognitive competencies such as problem-solving skills and promote the generation and exploration of representations and solutions methods for solving novel problems through productive failure. We also presented a framework adapted from the work of Shute (2011) on stealth assessment for assessing cognitive competencies that are grounded in ECD, as shown in Fig. 7.1. Stealth assessment addresses challenges of confounding constructs in game-based research and provides continuous streams of data to assess player performance without impeding the gameplay experience. In addition, the challenge of developing a competency model is vital to the validity of any GBA, including mapping behavioral indicators during gameplay to target competencies. ECD combined with stealth assessment can be a GBA framework that may guide future research in virtual and game-based immersive environments for assessing cognitive and noncognitive competencies.

Immersive technologies can now be used as an economical resource in formal educational settings. Scholars can explore the effects of immersive technologies on learning and cognition as well as address issues on equity and access to immersive technologies in underrepresented groups and communities. One example may be

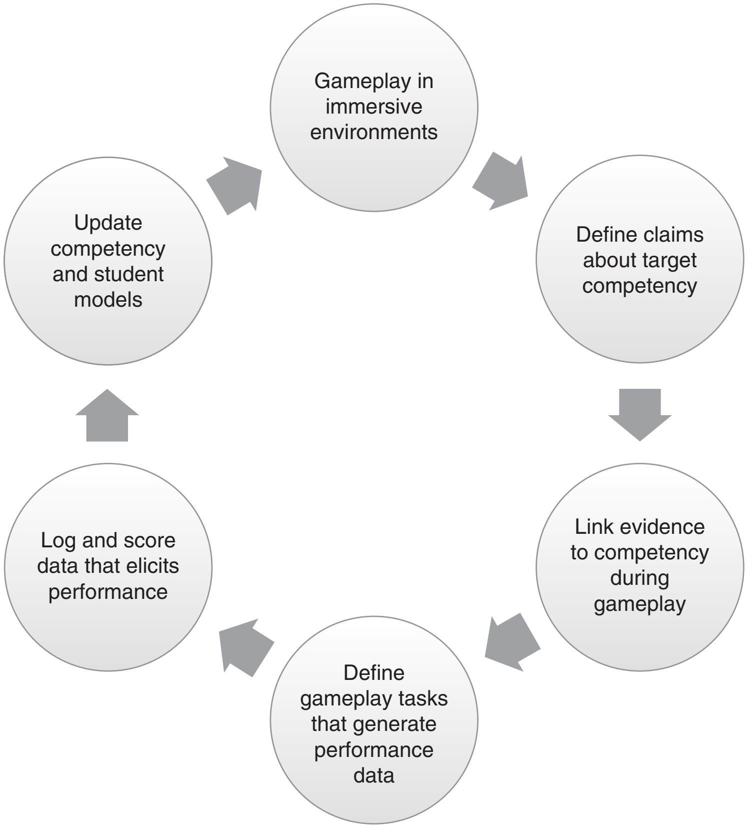


Fig. 7.1 Stealth assessment in immersive environments

using VRE[2] to help students with disabilities learn life transition skills. While VR technologies have existed in various sectors and domains including media, cinema, art, and the military, the application of VR in educational settings has been challenging given the cost and limited accessibility. However, mainstream and educational VR experiences are now possible with products such as the Oculus Go and Oculus Quest, which are self-contained VR systems that can be used without any additional computer hardware. In addition, a recent report (Adams Becker, Freeman, Giesinger Hall, Cummins, & Yuhnke, 2016) indicates substantial investments in immersive VR experiences can be expected to benefit the education sector in the near future. A similar report by Goldman Sachs predicted that immersive technologies as an industry can project to an $80 bn market by 2025 (Bellini et al., 2016). Given these trends, several technology-driven companies (e.g., Facebook, Google, and Samsung) have competed for investing, designing, and developing immersive technologies to provide mainstream VR immersive experiences (Brown & Green, 2016). Moreover, these companies realize the potential benefits of immersive environments for learning just as game-based scholars have demonstrated that video gameplay can improve learners' problem-solving skills by interacting with novel problem-solving scenarios (e.g., Shute et al., 2015).

Problem solving is an integral part of game design. Through design, games also serve as an excellent tool to assess learners' problem-solving skills. In addition to these problem-solving skills, designers can easily build-in components that borrow from the instructional approach of productive failure to elicit productive failure outcomes (i.e., shortcomings in their initial attempts at solving problems, which lead to successful long-term learning). By utilizing game design elements that elicit these outcomes and by measuring the aspects within each phase (e.g., within exploration measuring the learners' problem-solving attempts and within consolidation measuring the learners' ability to consolidate based on the feedback type), game designers can hopefully provide more meaningful long-term learning experiences for their learners. During each of these phases, games can measure explicitly and implicitly certain variables that could moderate learners overarching success for achieving these types of outcomes. Additionally, researchers can use games as an environment to test the efficacy of this approach. The aforementioned GBAs also apply to VR environments for education. The unique features of varied forms of VRs, researchers and practitioners are encouraged to tailor types of assessment to accommodate learners' learning process in the virtual world. It will be more equitable to assess the learners in the ways they apply content knowledge.

References

Abrahamson, D., & Kapur, M. (2018). Reinventing discovery learning: A field-wide research program. *Instructional Science, 46*(1), 1–10.

Adams Becker, S., Freeman, A., Giesinger Hall, C., Cummins, M., & Yuhnke, B. (2016). *Horizon report: 2016 K-12 edition*. Austin, TX: New Media Consortium.

Andersen, S. A. W., Konge, L., & Sørensen, M. S. (2018). The effect of distributed virtual reality simulation training on cognitive load during subsequent dissection training. *Medical Teacher, 40*, 684–689.

Anderson, C. G., Dalsen, J., Kumar, V., Berland, M., & Steinkuehler, C. (2018). Failing up: How failure in a game environment promotes learning through discourse. *Thinking Skills and Creativity, 30*, 135–144.

Bellini, H., Chen, W., Sugiyama, M., Shin, M., Alam, S., & Takayama, D. (2016). *Profiles in innovation: Virtual & augmented reality*. New York, NY: Goldman Sachs Group, Inc.

Brown, A., & Green, T. (2016). Virtual reality: Low-cost tools and resources for the classroom. *TechTrends, 60*(5), 517–519. https://doi.org/10.1007/s11528-016-0102-z

Chang, T. P., & Weiner, D. (2016). Screen-based simulation and virtual reality for pediatric emergency medicine. *Clinical Pediatric Emergency Medicine, 17*(3), 224–230.

Cho, J. S., Hahn, K. Y., Kwak, J. M., Kim, J., Baek, S. J., Shin, J. W., & Kim, S. H. (2013). Virtual reality training improves da Vinci performance: A prospective trial. *Journal of Laparoendoscopic & Advanced Surgical Techniques, 23*(12), 992–998.

Clarke-Midura, J., & Dede, C. (2010). Assessment, technology, and change. *Journal of Research on Technology in Education, 42*(3), 309–328.

Dede, C. (2005). Planning for neomillennial learning styles. *Educause Quarterly, 28*(1), 7–12.

Dede, C. (2009). Immersive interfaces for engagement and learning. *Science, 323*(66), 66–69.

DiCerbo, K., Shute, V. J., & Kim, Y. J. (2017). The future of assessment in technology rich environments: Psychometric considerations. In J. M. Spector, B. Lockee, & M. Childress (Eds.), *Learning, design, and technology: An international compendium of theory, research, practice, and policy* (pp. 1–21). New York, NY: Springer.

Eseryel, D., Law, V., Ifenthaler, D., Ge, X., & Miller, R. (2014). An investigation of the interrelationships between motivation, engagement, and complex problem solving in game-based learning. *Educational Technology & Society, 17*(1), 42–53.

Fong, K., Jenson, J., & Hebert, C. (2018). Challenges with measuring learning through digital gameplay in K-12 classrooms. *Media and Communication, 6*(2), 112–125.

Gauthier, A., & Jenkinson, J. (2018). Designing productively negative experiences with serious game mechanics: Qualitative analysis of gameplay and game design in a randomized trial. *Computers & Education, 127*, 66–89.

Gee, J. P. (2005). Learning by design: Good video games as learning machines. *E-Learning and Digital Media, 2*(1), 5–16.

Gee, J. P. (2008). Learning and games. In K. Salen (Ed.), *The ecology of games: Connecting youth, games, and learning* (pp. 21–40). Cambridge, MA: MIT.

Glogger-Frey, I., Fleischer, C., Grüny, L., Kappich, J., & Renkl, A. (2015). Inventing a solution and studying a worked solution prepare differently for learning from direct instruction. *Learning and Instruction, 39*, 72–87. https://doi.org/10.1016/j.learninstruc.2015.05.001

Hart Research Associates. (2018). *Fulfilling the American dream: Liberal education and the future of work*. Washington, DC: Association of American Colleges and Universities.

Hew, K. F., & Cheung, W. S. (2010). Use of three-dimensional (3-D) immersive virtual worlds in K-12 and higher education settings: A review of the research. *British Journal of Educational Technology, 41*(1), 33–55.

Jagust, T., Boticki, I., & So, H. (2018). Examining competitive, collaborative and adaptive gamification in young learners' math learning. *Computers & Education, 125*, 444–457.

Jensen, L., & Konradsen, F. (2018). A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies, 23*(4), 1515–1529.

Johnson-Glenberg, M. C., & Megowan-Romanowicz, C. (2017). Embodied science and mixed reality: How gesture and motion capture affect physics education. *Cognitive Research: Principles and Implications, 2*(1), 24.

Jonassen, D. H. (1997). Instructional design models for well-structured and III-structured problem-solving learning outcomes. *Educational Technology Research and Development, 45*(1), 65–94.

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development, 48*(4), 63–85.

Kapur, M. (2008). Productive failure. *Cognition and Instruction, 26*(3), 379–424. https://doi.org/10.1111/cogs.12107

Kapur, M. (2009). Productive failure in mathematical problem solving. *Instructional Science, 38*(6), 523–550. https://doi.org/10.1007/s11251-009-9093-x

Kapur, M. (2010). A further study of productive failure in mathematical problem solving: Unpacking the design components. *Instructional Science, 39*(4), 561–579. https://doi.org/10.1007/s11251-010-9144-3

Kapur, M. (2012). Productive failure in learning the concept of variance. *Instructional Science, 40*(4), 651–672. https://doi.org/10.1007/s11251-012-9209-6

Kapur, M. (2013). Comparing learning from productive failure and vicarious failure. *Journal of the Learning Sciences, 23*(4), 651–677. https://doi.org/10.1080/10508406.2013.819000

Kapur, M. (2014). Productive failure in learning math. *Cognitive Science, 38*(5), 1008–1022. https://doi.org/10.1111/cogs.12107

Kapur, M. (2015). The preparatory effects of problem solving versus problem posing on learning from instruction. *Learning and Instruction, 39*, 23–31. https://doi.org/10.1016/j.learninstruc.2015.05.00

Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist, 51*(2), 289–299. https://doi.org/10.1080/00461520.2016.1155457

Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences, 21*(1), 45–83. https://doi.org/10.1080/10508406.2011.591717

Kapur, M., & Kinzer, C. K. (2009). Productive failure in CSCL groups. *International Journal of Computer-Supported Collaborative Learning, 4*(1), 21–46. https://doi.org/10.1007/s11412-008-9059-z

Kapur, M., & Rummel, N. (2012). Productive failure in learning from generation and invention activities. *Instructional Science, 40*(4), 645–650. https://doi.org/10.1007/s11251-012-9235-4

Ke, F., Lee, S., & Xu, X. (2016). Teaching training in a mixed-reality integrated learning environment. *Computers in Human Behavior, 62*, 212–220.

Kirk, D., & MacPhail, A. (2002). Teaching games for understanding and situated learning: Rethinking the Bunker-Thorpe model. *Journal of Teaching in Physical Education, 21*, 117–192.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, MA: Cambridge University Press.

Leder, J., Horlitz, T., Puschmann, P., Wittstock, V., & Schütz, A. (2019). Comparing immersive virtual reality and powerpoint as methods for delivering safety training: Impacts on risk perception, learning, and decision making. *Safety Science, 111*, 271–286.

Loibl, K., & Rummel, N. (2014a). The impact of guidance during problem-solving prior to instruction on students' inventions and learning outcomes. *Instructional Science, 42*(3), 305–326. https://doi.org/10.1007/s11251-013-9282-5

Loibl, K., & Rummel, N. (2014b). Knowing what you don't know makes failure productive. *Learning and Instruction, 34*, 74–85. https://doi.org/10.1016/j.learninstruc.2014.08.004

Macedonia, M., & von Kriegstein, K. (2012). Gestures enhance foreign language learning. *Biolinguistics, 6*(3–4), 393–416.

Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction, 60*, 225–236.

Mayer, R., & Wittrock, M. (2006). Problem solving. In P. Alexander & P. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 287–303). Mahwah, NJ: Erlbaum Publishers.

Mazziotti, C., Loibl, K., & Rummel, N. (2015). Collaborative or individual learning within productive failure. Does the social form of learning make a difference? In O. Lindwall, P. Häkkinen, T. Koschman, P. Tchounikine, & S. Ludvigsen (Eds.), *Exploring the material conditions of learning: The computer supported collaborative learning (CSCL) conference 2015* (Vol. 2, pp. 570–575). Gothenburg, Sweden: ISLS.

McGrath, J. L., Taekman, J. M., Dev, P., Danforth, D. R., Mohan, D., Kman, N., … Talbot, T. B. (2018). Using virtual reality simulation environments to assess competence for emergency medicine learners. *Academic Emergency Medicine, 25*(2), 186–195.

McGregor, C., Bonnis, B., Stanfield, B., & Stanfield, M. (2017). Integrating Big Data analytics, virtual reality, and ARAIG to support resilience assessment and development in tactical training. In *2017 IEEE 5th International Conference on Serious Games and Applications for Health* (pp. 1–7). IEEE.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62.

Nagendran, M., Gurusamy, K. S., Aggarwal, R., Loizidou, M., & Davidson, B. R. (2013). Virtual reality training for surgical trainees in laparoscopic surgery. *Cochrane Database of Systematic Reviews*, (8), CD006575.

Nathan, M. J., & Walkington, C. (2017). Grounded and embodied mathematical cognition: Promoting mathematical insight and proof using action and language. *Cognitive Research: Principles and Implications, 2*(1), 9.

Passig, D., Tzuriel, D., & Eshel-Kedmi, G. (2016). Improving children's cognitive modifiability by dynamic assessment in 3D Immersive Virtual Reality environments. *Computers & Education, 95*, 296–308.

Richards, D., & Taylor, M. (2015). A Comparison of learning gains when using a 2D simulation tool versus a 3D virtual world: An experiment to find the right representation involving the Marginal Value Theorem. *Computers & Education, 86*, 157–171.

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., & Emihovich, B. (2018). Assessing problem-solving skills in immersive environments. In J. Voogt, G. Knezek, R. Christensen, & K.-W. Lai (Eds.), *International handbook of information technology in primary and secondary education*. Cham, Switzerland: Springer.

Shute, V. J., Hansen, E. G., & Almond, R. G. (2008). You can't fatten a hog by weighing it—Or can you? Evaluating an assessment for learning system called ACED. *International Journal of Artificial Intelligence in Education, 18*(4), 289–316.

Shute, V. J., Ke, F., & Wang, L. (2017). Assessment and adaptation in games. In P. Wouters & H. van Oostendorp (Eds.), *Instructional techniques to facilitate learning and motivation of serious games* (pp. 59–78). New York, NY: Springer.

Shute, V. J., Rahimi, S., & Emihovich, B. (2018). Assessment for learning in immersive environments. In D. Lui, C. Dede, R. Huang, & J. Richards (Eds.), *Virtual, augmented, and mixed realities in education* (pp. 71–89). Heidelberg, Germany: Springer.

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., Wang, L., Greiff, S., Zhao, W., & Moore, G. R. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior, 63*, 106–117.

Smith, P. C., & Hamilton, B. K. (2015). The effects of virtual reality simulation as a teaching strategy for skills preparation in nursing students. *Clinical Simulation in Nursing, 11*(1), 52–58.

Stanica, I., Dascalu, M. I., Bodea, C. N., & Moldoveanu, A. D. B. (2018). VR job interview simulator: Where virtual reality meets artificial intelligence for education. In *2018 Zooming Innovation in Consumer Technologies Conference (ZINC)* (pp. 9–12). IEEE.

Sugden, C., Aggarwal, R., Banerjee, A., Haycock, A., Thomas-Gibson, S., Williams, C. B., & Darzi, A. (2012). The development of a virtual reality training curriculum for colonoscopy. *Annals of Surgery, 256*(1), 188–192.

Taub, M., Azevedo, R., Bradbury, A. E., Millar, G. C., & Lester, J. (2018). Using sequence mining to reveal the efficiency in scientific reasoning during STEM learning with a game-based learning environment. *Learning and Instruction, 54*, 93–103.

Tiffany, J. M., & Hoglund, B. A. (2016). Using virtual simulation to teach inclusivity: A case study. *Clinical Simulation in Nursing, 12*(4), 115–122.

Toh, P. L. L., & Kapur, M. (2017). Is having more prerequisite knowledge better for learning from productive failure? *Instructional Science, 45*(3), 377–394. https://doi.org/10.1007/s11251-016-9402-0

Van Eck, R. N., Hung, W., Bowman, R., & Love, S. (2009). 21st century game design: A model and prototype for promoting scientific problem solving. In *Proceedings of the Twelfth IASTED International Conference on Computers and Advanced Technology in Education: Globalization of Education Through Advanced Technology*. Calgary, Canada. ACTA Press.

Van Eck, R. N., Shute, V. J., & Rieber, L. P. (2017). Leveling up: Game design research and practice for instructional designers. In R. Reiser & J. Dempsey (Eds.), *Trends and issues in instructional design and technology* (4th ed., pp. 227–285). New York, NY: Pearson.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Westermann, K., & Rummel, N. (2012). Delaying instruction: Evidence from a study in a university relearning setting. *Instructional Science, 40*(4), 673–689. https://doi.org/10.1007/s11251-012-9207-8

Xu, X., & Ke, F. (2014). From psychomotor to 'motorpsycho': Learning through gestures with body sensory technologies. *Educational Technology Research and Development, 62*(6), 711–741.

Chapter 8
New Perspectives on Game-Based Assessment with Process Data and Physiological Signals

Steve Nebel and Manuel Ninaus

8.1 Introduction

The increasing acceptance and adoption of digital educational games (Boyle et al., 2016) as a tool for assessment and learning is often in contrast to traditional educational settings, which still often rely on written final exams as their primary source of evaluation. However, digital educational games can provide so much more. One crucial aspect of learning in and with digital educational games is the potential acquisition of numerous interaction data that can inform about the ongoing learning process. The question, though, is how can we maximally benefit from the data provided by the learners or the learning environment, respectively, to enable personalization and improve learning processes and assessment. The answer to this question becomes even more complex when considering the recent advances in sensor technology, which allow for (neuro) physiological measurements during learning (Schneider, Börner, van Rosmalen, & Specht, 2015), thus providing deeper insight into the underlying processes. Consequently, in this chapter, we analyze the existing literature on the use of behavioral process, as well as physiological data in game-based assessment (GBA) and learning (GBL). In addition to the assessment of cognitive processes, we specifically focus on motivational and emotional processes, as learning is not merely a cognitive process but is essentially influenced by emotions and motivation (e.g., Howard-Jones & Jay, 2016; Pekrun, 2011; Wise, 2004). Therefore, we first give a brief applied example of how process data can enrich

Steve Nebel and Manuel Ninaus shared first-authorship.

S. Nebel
Technische Universität Chemnitz, Chemnitz, Germany
e-mail: steve.nebel@phil.tu-chemnitz.de

M. Ninaus (✉)
Leibniz-Institut für Wissensmedien, Tübingen, Germany
e-mail: m.ninaus@iwm-tuebingen.de

D. Ifenthaler, Y. J. Kim (eds.), *Game-Based Assessment Revisited*, Advances in Game-Based Learning, https://doi.org/10.1007/978-3-030-15569-8_8

assessment beyond mere win/lose or correct/wrong classifications, and we then focus specifically on the use of behavioral process data for GBA on the one hand and physiological data on the other. Subsequently, we analyze current technological trends in GBA, which might facilitate the integration of these data. A combination of both behavioral process and physiological data is discussed next, after which we provide guidelines on how researchers as well as nonexperts alike can benefit from this approach. Finally, we end by summarizing the implications of our perspective.

8.2 A Post-Game Score Is Not Enough

Most game systems assessing learner skills rely on pre- or post-game measures (Smith, Blackmore, & Nesbitt, 2015). However, such data might only be a rough approximation regarding play experience, performance, or processes within the user. To illustrate this argument and to provide a first glimpse into the importance of process and physiological data, a game-related example will be introduced. An example from the field of eSports is particularly suitable, as measurements and comparisons of performance are integral elements of competitive sports, and eSports are naturally connected to the field of video games. For instance, within Counter Strike: Global Offensive (Valve Software & Hidden Path Entertainment, 2012), two teams play against each other, until one of them manages to win 16 individual rounds (for the sake of simplicity, overtime rules and different game modes and matches that include more than one map will not be discussed here). A round consists of five players on each team, either aiming to plant a "bomb" or trying to prevent this. If a team is completely defeated (i.e., each player is "dead") or fails to achieve the objective, the opposing team scores a point. From the perspective of traditional GBA, the final score (e.g., 16:8) with at least 16 individual measurements (i.e., the outcome of each round) could be suitable to describe each team's performance. However, as multiple players are involved and each round might take several minutes, it seems evident that this metric can provide only a rough approximation. Additionally, this score is only interpretable relative to the skill level of another team. The nature of the game, the number of necessary skills, and the complexity of involved processes cannot be adequately assessed by such a simple measurement. Thus, if researchers try to gain better insights, the scope of the involved measures has to shift toward assessing more individual data during the process. By doing so, they can investigate if one team is technically better (e.g., regarding their aiming) but lacks other skills (e.g., tactical expertise). The desire to gather deeper insights and explanations is not exclusive to eSports (e.g., batting average within baseball); however, digital media such as video games offer the advantage that many necessary data streams are digitized (e.g., user inputs, communication, and behavioral data). In addition, user experience research or media and instructional psychology develop further physiological measures that could extend the insights into the individual player. Concurrently, the necessary hardware becomes less intrusive and expensive. For instance, a hardware specialist released a gaming mouse that measures heart rate and galvanic skin response (Mionix, 2018). This could be used during a Counter Strike match to assess whether a shot was more likely missed because of insufficient skill (process data of the mouse input) or because of a stress reaction (physiological data), while both measures offer much more validity than the final score. In a similar way, behavioral process data, as well as physiological data, can inform and support GBA by providing deeper insights into the cognitive, emotional, and motivational processes of the learner.

8.3 Process Data

Many investigations regarding central concepts in media psychology and psychological game research require deeper investigations using process data. By investigating such datasets, researchers target the acquisition of detailed information over the course of a specific action, in contrast to a singular event or an aggregated final measure. Typically, such data gains importance if the addressed psychological state is not stable or binary and post hoc questionnaires cannot be utilized to reconstruct the individual's psychological state during a specific time frame. For example, Peter Vorderer and his colleagues (2003, p. 4) define spatial presence (e.g., while playing a video game) as "a binary experience, during which perceived self-location and realization of action possibilities are connected to a mediated spatial environment; mental capacities are bound by the mediated environment instead of reality; and these conditions can be enhanced by different sensory input and action feedbacks but does not necessarily rely on them." However, the concept is measured with questionnaires such as the MEC-Spatial Presence Questionnaire (Vorderer et al., 2004). Among other scales, the questionnaire addresses the core concept of spatial presence as two 8-item scales, ranging from one to five. Thus, the individual and temporal phenomenon of perceiving spatial presence is projected onto a numerical interval. This might be suitable to estimate the frequency of perceived spatial presence but partially contradicts the initial concept. For instance, a researcher could estimate that a participant reporting a score of three might have perceived spatial presence more often than a participant reporting a score of two. However, if the researchers could acquire process data (e.g., a binary spatial presence value and timestamps), they could gain much more detailed insights into the formation of spatial presence or similar phenomena. For example, did a specific event or media feature shift the perception of spatial presence? This procedural perspective is embraced within qualitative approaches (Lamnek & Krell, 2016), but within quantitative experiments such data are often missing or challenging to acquire. Additionally, the lack of process information increases the difficulty of investigating temporal cause-and-effect patterns and contributes to the current deficiency of detailed data-driven process models within the field of (media) psychology. Similarly, researchers in the field of educational psychology are aware of this issue and approaches of real-time processing data are considered (Alexander, 2018). Moreover, Mayer (2018) argues that one of the central areas of educational psychology is the investigation of learning *processes,* further emphasizing the importance of this perspective. In sum, it is not surprising that the interest in process data has emerged within the fields of GBL and GBA as researchers utilize methods originating from the intersection of educational and media psychology.

In respect to the focus of this chapter on cognitive, emotional, and motivational aspects, several other examples highlight the crucial importance of process data. For example, within the widely acknowledged theoretical framework of the Cognitive Load Theory (e.g., Kalyuga & Singh, 2016; Paas & Sweller, 2014; Sweller, 1994), the concept of cognitive load is described as the individual cognitive strain imposed on the space-limited working memory. The nature of this load, extraneous (i.e., learning irrelevant) or intrinsic (i.e., learning-relevant), and its specific value fluctuates during the course of a learning task (Fig. 8.1).

However, specific and time-dependent values are rarely accessed. More frequently, overall scores measuring the more or less specific types of load are used

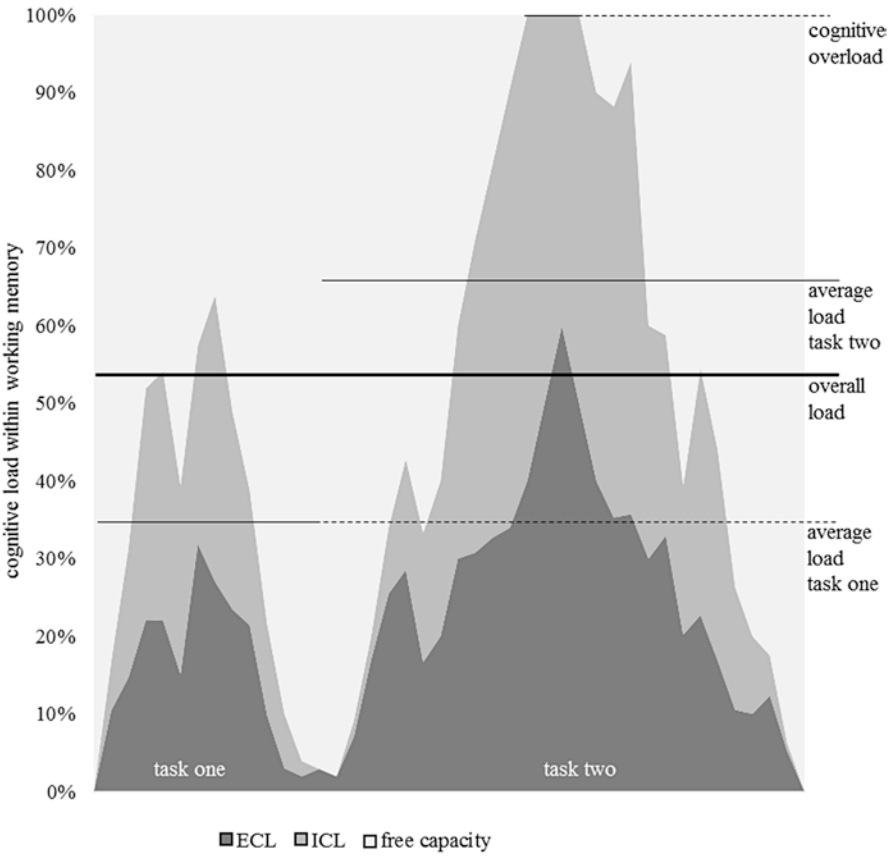


Fig. 8.1 The course of cognitive load during learning, taken from Nebel (2017)

(e.g., Eysink et al., 2009). Although frequently validated and sufficient for various scientific questions, such aggregated data raise challenges regarding the interpretation of mechanisms within information *processing* systems. Therefore, new approaches attempt to utilize other measures, such as speech analysis (e.g., Wirzberger, Herms, Esmaeili Bijarsari, Rey, & Eibl, 2017), to gather deeper insights into the underlying processes. Such measures could be especially useful within the gaming context, as speech-analysis is a noninvasive method of data collection that could accompany the natural communication included in many social games.

Second, the topic of emotions within the learning process is currently evoking interest from educational psychologists, especially as evidence provided by quantitative experimental approaches (e.g., Schneider, Nebel, & Rey, 2016) and new perspectives within theoretical frameworks (e.g., the Integrated Cognitive Affective Model of Learning with Multimedia; see Plass & Kaplan, 2016) has highlighted the potential influences on the learner. However, similar to cognitive strains, emotions and related concepts within educational settings are frequently assessed with standardized post hoc questionnaires (e.g., Achievement Emotions Questionnaire; see Pekrun, Goetz, Frenzel, Barchfeld, & Perry, 2011) or by nonstandardized tests (e.g., asking how much *fun* the participant had; see Nebel, Schneider, Schledjewski, & Rey, 2017). However, emotions and affect do not only change on the macro level of a lifetime (e.g., Charles, Reynolds, & Gatz, 2001). Instead, they might be affected or even intentionally manipulated by specific games or gameplay mechanics (Granic, Lobel, & Engels, 2014). Therefore, a continuous measurement of the whole emotional *process* might be crucial for understanding the impact of individual game elements on the learner. For example, the learner might report positive feelings in a post hoc questionnaire, but a rather depressing experience might have triggered a crucial learning experience. Such detailed measures might be essential, especially if complex emotional experiences or *serious* issues are addressed within educational video games (e.g., Charles University & Czech Academy of Sciences, 2017).

Finally, motivational concepts are of interest while reflecting the potential and challenges within process data. Most prominently, the concept of *flow* (Csikszentmihalyi, 1990) is investigated as an important motivational state during the use of video games. Similar to the previously introduced example of spatial presence, a state of flow is a binary state. If currently not experienced, other states are assumed, such as anxiety or boredom. During the process of playing a video game, players might perceive these different states within different segments of the game. However, within educational video game research, the concept of flow is frequently assessed rather unspecific with observations (e.g., Admiraal, Huizenga, Akkerman, & Ten Dam, 2011) or post hoc questionnaires (e.g., Hamari et al., 2016). Within the related field of difficulty research, researchers seek to systematically vary game elements to induce specific perceptions (e.g., the perception of a difficult or easy game; see Nebel, Schneider, Beege, & Rey, 2017) or manipulate game characteristics linearly (Lomas et al., 2017) to gain insights into the complete range of potential states. Although potentially more valid, even these approaches do not provide sufficient insights into the whole process of emerging motivational states or their fluctuations during the course of gameplay. However, such process data might be crucial as game designers start to adapt game properties while the players are playing (e.g., Xue, Wu, Kolen, Aghdaie, & Zaman, 2017) and similar approaches start to emerge within educational games. For example, the Adaptation and Assessment (TwoA) component (Nyamsuren, Van der Vegt, & Westera, 2017) was developed to constantly assess players' skill levels and the task difficulty. Nevertheless, such *stealth assessment algorithms* include constant user variables. For instance, the algorithm tries to match the individual skill level and task difficulty to achieve a constant predicted success rate of 75% (Klinkenberg, Straatemeier, & van der Maas, 2011). Extensive use of process date might elaborate such constants or, if included throughout the implementation, lead to adaptive systems supporting the whole learning process.

8.4 Physiological Data

In recent years, interest in utilizing physiological data to provide better and more personalized support for learning, education, and assessment has increased (for a review, see Schneider et al., 2015). However, this has proven to be a nontrivial endeavor. Regardless of advances in (stealth assessment) algorithms and learning analytics methods, it remains a challenge to infer learning-relevant user states from the huge amounts of data provided from physiological sensors (e.g., Wu, Huang, & Hwang, 2016 for the case of identifying emotional states). Nevertheless, building on the strengths of process data, physiological data offer further advantages. First, cognitive, emotional, and motivational states can be assessed continuously during learning/playing without probing or disturbing users. Second, in contrast to post hoc questionnaires, they provide an objective measure of subjective experiences. Third, physiological data are not affected by inaccurate recall and memory, such as recency (Freeman, Avons, Pearson, & IJsselsteijn, 1999), as compared to post hoc questionnaires. Fourth, physiological responses are usually involuntary and, thus, are not subject to manipulation as written or verbal responses are. Consequently, physiological data provide unique possibilities for getting access to fine-grained internal changes by allowing a deeper and more direct insight into cognitive, emotional, and motivational processes.

Indeed, the most direct types of physiological measurement used to assess internal and learning-relevant states are neurofunctional measures, such as electroencephalography (EEG), functional near-infrared spectroscopy (fNIRS), and functional magnetic resonance imaging (fMRI). For obvious reasons, fMRI is not applicable to real-life-like learning scenarios. However, this method offers ways of improving our understanding of cognitive, emotional, and motivational processes during gameplay (for a review, see Ninaus et al., 2014). For instance, Klasen, Weber, Kircher, Mathiak, and Mathiak (2012) identified flow-specific neural activation patterns during free play of a first-person shooter. In another study (Baumgartner et al., 2008), researchers observed highly specific neural networks modulating the experience of spatial presence in virtual environments. Both of these studies provide important insights into the underlying mechanisms of flow and presence. Hence, they might also contribute to learning about fostering these subjective experiences as well as facilitate the identification of related behavioral and physiological data metrics.

EEG, as one of the most often used neurofunctional methods during gaming, and fNIRS are commonly favored for assessing learning-relevant user states due to their lightweight, noninvasive, and rather unobtrusive natures (Ninaus, Kober, Friedrich, Dunwell, et al., 2014). In numerous EEG studies, neuronal responses to instantaneous game events or during prolonged game sessions were recorded to detect varying levels of cognitive workload (e.g., Allison & Polich, 2008), engagement (e.g., Pugnetti et al., 1996), mental effort (e.g., Nacke, Grimshaw, & Lindley, 2010; Pellouchoud, Smith, McEvoy, & Gevins, 1999; Salminen & Ravaja, 2008), and cognitive demand (e.g., Salminen & Ravaja, 2007), as well as the emotional states of the user (e.g., Liu et al., 2010). However, even more elusive states, such as flow (e.g., Berta, Bellotti, de Gloria, Pranantha, & Schatten, 2013) and presence (e.g., Baumgartner, Valko, Esslen, & Jäncke, 2006; Kober & Neuper, 2012) were investigated using EEG. For instance, Kober and Neuper (2012) used event-related potentials and a sophisticated experimental paradigm to successfully differentiate between individuals in a high or low presence condition. Importantly, their neurofunctional results demonstrated a tight link between the feeling of being present in a virtual learning environment and attention.

The use of fNIRS as a method for human–computer-interaction (HCI) and GBL research has become rather popular in recent years (e.g., Solovey et al., 2012). It is a new, noninvasive optical neuroimaging technique that was first introduced in 1992 and utilizes hemodynamic changes in the blood of the human brain (Ferrari & Quaresima, 2012). It is argued that fNIRS is better suited for realistic (learning) settings compared to other neurofunctional methods because it is cheaper, more portable, and more robust to noise (but, see also Strait, Canning, & Scheutz, 2013 for a more realistic view). Consequently, fNIRS has been used in various scenarios to investigate learning-relevant states, such as players' skill levels (Hattahara, Fujii, Nagae, Kazai, & Katayose, 2008) or game difficulty (Girouard et al., 2009). For instance, Witte et al. (Witte, Ninaus, Kober, Neuper, & Wood, 2015; see also Ninaus, Kober, Friedrich, Neuper, & Wood, 2014) identified distinct neurofunctional activation patterns during gameplaying, successfully differentiating among states of active learning, application of knowledge, and no learning. These identifiable distinct neurofunctional patterns have particular appeal in GBL and GBA, as they allow for developing adaptive and personalized systems that can utilize brain activation data to adapt learning content.

Neurofunctional measures are indispensable tools for investigating cognitive, emotional, and motivational processes and their underlying mechanisms. However, as of today, they share common downsides for GBA: limited applicability in real-life learning scenarios and the inability of being used as stealth assessment methods. Advances in sensor technology might change this in the future. Importantly, though, other physiological signals, such as electrodermal activity (EDA) and heart rate, which might not be as direct and informational, allow for detecting cognitive, emotional, and motivation processes in more realistic learning scenarios.

The recent availability of wearable technologies extends the possibilities of measuring learning-relevant states in real-life-like learning environments in a nonintrusive manner. However, educational psychology and practice lag behind the potential of wearable technologies (Bower & Sturman, 2015). Nevertheless, numerous studies have demonstrated the high value of physiological sensor data to infer cognitive, emotional, and motivational states (for a review, see Schneider et al., 2015). For instance, in a recent study (Nourbakhsh, Chen, Wang, & Calvo, 2017), learners' galvanic skin response (GSR) was utilized for detecting different levels of cognitive load. In particular, certain GSR features were used to classify four levels of cognitive load with up to 80% accuracy during two arithmetic learning tasks. In another example, Xiao and Wang (2016) used changes in heart rate to identify disengagement of learners during a massive open online course (MOOC). The researchers then used this information to prompt messages (e.g., "Please Pay Attention!") when learners disengaged from the learning content, thus increasing learning gains by 20%. Most importantly, this adaptive system was particularly effective for the bottom performers, who improved their performance by 42%. Educational games research might greatly benefit from such solutions to study the impact of individual game elements on learners' engagement levels, consequently fostering learning outcomes.

The undeniable interdependency between emotions and learning (e.g., Ninaus, Moeller, McMullen, & Kiili, 2017; Plass & Kaplan, 2016) suggests a further important opportunity for physiological sensor usage in GBA (for an overview, see Novak & Johnson, 2012). Several studies employ a range of different sensor technologies to determine emotional states (e.g., Mandryk & Atkins, 2007). One popular approach is to record participants' faces to manually or automatically classify different emotional facial expressions (e.g., Guo & Dyer, 2005; Littlewort, Bartlett, Salamanca, & Reilly, 2011). Many of these approaches are based on using facial action units from the Facial Action Coding System (e.g., Cohn, Ambadar, & Ekman, 2007). In this context, emotions displayed in users' faces can also be determined by using facial electromyography (EMG) to measure the activity of facial muscles (Mandryk, Atkins, & Inkpen, 2006). Moreover, facial expressions can be used to differentiate between correct and incorrect trials in problem-solving tasks (Littlewort et al., 2011), providing a direct bridge between performance assessment and emotional facial expressions.

The detection of motivational states, such as flow (Csikszentmihalyi, 1990; for a review on flow in GBL, see Perttula, Kiili, Lindstedt, & Tuomi, 2017), can also benefit from physiological data (for an overview of physiological correlates of flow, see Peifer, 2012). For instance, Kivikangas (2006) employed EMG and identified flow to be associated with increased positive valence and decreased negative valence. On the other hand, Keller, Bless, Blomann, and Kleinböhl found reduced heart rate to be related to a state of flow indicating enhanced mental workload.

Neurophysiological and physiological measures offer key benefits for GBA. The rather unobtrusive detection of cognitive, emotional, and motivational states allows not only for fine-grained analysis of these states, but they also have the potential to be used for personalization and adaptation. However, as this field is still in its early stages, real-life GBL and GBA applications are underrepresented. One possibility for increasing widespread adoption and, consequently, rigorous empirical investigations might come from recent technological trends, for example, user-friendly game engines, which offer dedicated soft- and hardware support to integrate (neuro) physiological data into games.

8.5 Technological Developments Within GBA

Three relevant technological developments can be identified that might change the practice of GBA. First, *game engines* (i.e., middleware for the creation of game-related visual, auditory, network, interface, or simulation components) have evolved tremendously. Aside from its spectacular visual improvements (e.g., three-dimensional realism), such software has become widespread and more approachable. Second, simplified forms of game creation and programming, typically for educational purposes, can be harnessed as a powerful tool for *teachers as game developers*, subsequently increasing the access to the game mechanics. Third, games focusing on *user created content* and companies trying to foster the educational value of their games increase their potential regarding assessment, as well.

The first aspect of developments within game engines can be exemplified with the *Unreal Engine 4* (Epic Games, 2018). After a short phase of subscription-based monetization, the engine was released without initial costs in 2015. The engine is established within the industry and has been used for in-company products (e.g., Fortnite, Epic Games, 2017), by other professional studios (e.g., Hellblade: Senua's Sacrifice, Ninja Theory, 2017) and within experimental GBL research (e.g., Nebel, Beege, Schneider, & Rey, 2016; Nebel, Schneider, Beege, & Rey, 2017). Several strategic decisions have enhanced the potential of the engine regarding GBA. First of all, the developers integrated *Blueprints*, a form of visual coding (Fig. 8.2).

Thus, novice users can control game elements without the need to learn C++, the actual foundation of the engine. Furthermore, they can understand the game mechanisms better, as they can visually follow the flow of information while using the debugging tools. Altogether, this enables users to gain access to process data or to write simple code segments recording the desired information. For example, they can easily write code segments to record player movement (and save them to text files) or develop invisible objects that trigger predefined measurements during gameplay. In addition to the fully compiled engine release, the entire source code is frequently published on GitHub (GitHub, Inc., 2018). As a result, users can access every element of the engine. Although this might not be essential for the novice user, this open policy empowers experts to modify existing features and to add further features to the engine. For example, enthusiasts have created software

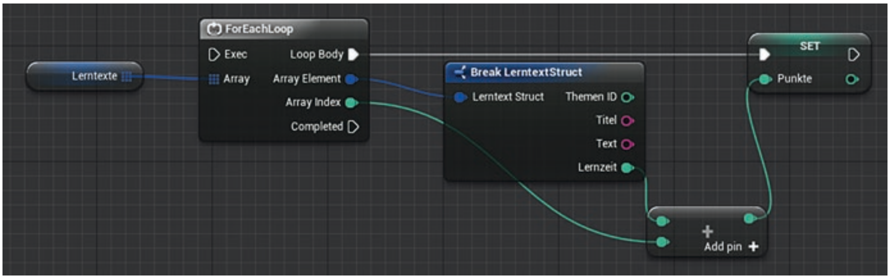


Fig. 8.2 A simple for-loop

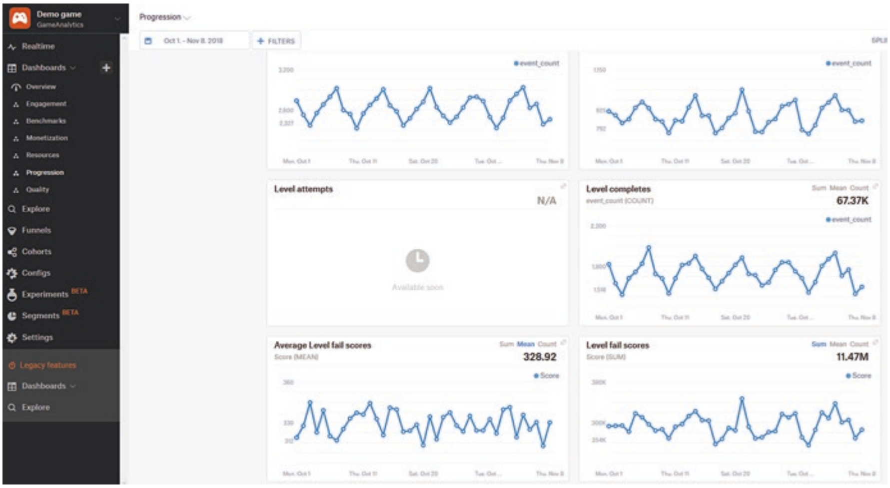


Fig. 8.3 Online interface of GameAnalytics

enabling connections to hardware necessary to measure physiological data (e.g., Thauros-Clan, 2016). The exchange of such modifications between developers and their potential customers is further supported with an embedded marketplace. For instance, in addition to the existing game analytics methods within the standard version of the engine, code plug-ins distributed through this interface might provide additional and often more accessible tools. For example, the GameAnalytics plug-in (GameAnalytics, 2016) provides dashboards with custom aggregated data visualizations (Fig. 8.3).

The second technological trend involves the new forms of simplified game creation and the empowerment of instructors as game designers. This is exemplified by the *Scratch* programming language (Maloney, Peppler, Kafai, Resnick, & Rusk, 2008; Maloney, Resnick, Rusk, Silverman, & Eastmond, 2010) and similar programs (*MIT App Inventor* from the Massachusetts Institute of Technology, 2012 or *Kodu* from Microsoft Research, 2009). Typically, such approaches are used within an educational context. Pupils can be taught the basics of programming with visual coding, utilizing a pedagogical optimized set of well-defined code segments (*blocks*). This approach can be used to create animations, interactive sceneries, and even small video games. As the target audience consists typically of learners themselves, even inexperienced instructors can use the programming language to create their own pedagogical games. Thus, in contrast to off-the-shelf games, teachers gain access to the information within the game. This could be used to gather process data (e.g., such data can be exported as .txt files). Naturally, the technological barriers are much stricter than within professional game engines, limiting the current potential. This should be considered especially regarding complex physiological measurements. However, the language allows external extensions (Dasgupta, Clements, Idlbi, Willis-Ford, & Resnick, 2015), for example, connections to external hardware or other JavaScript programs (some of them can be found within the ScratchX project: http://scratchx.org). As the interest in this form of educational tool grows, the functionalities are constantly expanded. For example, the latest version of *Scratch* (2.0) includes advanced tools such as video sensing.

Finally, some games rely on *user-generated content* (that could be of an educational nature) and provide the necessary tools within the game environment itself. Most prominently, *Minecraft* has been used within educational settings (for a detailed review, see Nebel, Schneider, & Rey, 2016). The approachable methods of content creation and its openness for modifications led to multifaceted applications and remarkable renown with a whole generation of teachers and learners. This amplified the spread of educational modifications or game versions, such as *ComputercraftEDU* (Ratcliffe, 2017) or *MinecraftEDU* (Microsoft, 2018). The latter has been taken over and re-released as a stand-alone by the Microsoft Corporation, after the company bought *Minecraft* as well, further emphasizing the renown of such video games. Within these games or modifications, players (i.e., learners or teachers) can create small educational game-like segments or environments for others to play and to learn with. This user-friendly process enables broad access to process data. For example, in an experiment by Marklund, Backlund, and Johannesson (2013), different texture packs were used to track the players' individual contributions. Additionally, with the conception and observation of such problem-solving tasks, deep stealth assessment of individual skills might be possible. For example, Shute and Wang (2016) illustrated how to assess the skill of *creativity* within a two-dimensional physics game. Similarly, researchers and teachers could create tasks within *Minecraft* for the purpose of stealth assessment without in-depth programming knowledge.

These three technological developments might shape the future of GBA. However, the development is not entirely linear, and new challenges arise. For example, strict monetization with impractical licensing models or closed software without open source projects might dampen future developments. Nonetheless, the outlined examples highlight the tremendous potential for widespread collection of process and physiological data. In this vein, Perez-Colado and his colleagues (Perez-Colado, Alonso-Fernandez, Freire, Martinez-Ortiz, & Fernandez-Manjon, 2018, p. 9) argue, that "the adoption of learning analytics can greatly benefit from its direct integration into game authoring tools that simplify costs and knowledge required for its application."

8.6 Merging Process and Physiological Data

Although the use of GBL is increasing and is being further facilitated by new technological developments, critics have raised issues regarding the effectiveness of GBL (e.g., Mayer, 2015; Shute & Ventura, 2013). It seems that many studies on GBL suffer from conceptual, theoretical, and methodological issues that undermine the value of GBL to foster learning, transfer of knowledge, and problem solving. Recent research, for instance, has indicated that, in specific scenarios, increased enjoyment or "fun" in GBA might be achieved by game mechanics, thus limiting the reliability of the assessment (Greipl, Ninaus, Bauer, Kiili, & Moeller, 2018). As a result, it has been argued that GBL and GBA can be improved by using theory-driven approaches and interdisciplinary approaches and methods to examine cognitive, emotional, and motivational processes during gameplay to better understand and comprehend their interaction, providing more than "simple" pre-post-test measures and self-reports (e.g., Mayer, 2014; Taub et al., 2017). New technological developments, as discussed above, allow for acquiring more data to infer learning-relevant user states. In this context, deducing cognitive, motivational, and emotional states from processes or physiological data alone have received considerable attention. However, results are often associated with varying degrees of uncertainty. A combination of both behavioral process data and physiological data should increase detection rates of user states by complementing each other, allowing a deeper and more direct look into cognitive, emotional, and motivational processes (e.g., Azevedo et al., 2013; Taub et al., 2017) at any given moment in time. In the long term, this should facilitate and advance theory building efforts in GBA.

The combination of multiple data channels is already well reflected in many studies employing physiological data to detect cognitive, emotional, and motivational user states (for a review, see Schneider et al., 2015). This is not only due to the rise of affordable wearable technologies but also to the necessity for reliable detection. Single physiological data channels only provide indicators for certain user states and, thus, provide limited information. Various multiple physiological data channels, which point in the same direction in terms of presence or absence of a learning-relevant user state, for example, flow, can improve detection rates and decrease uncertainty. For instance, intense positive and negative emotions are not that well distinguishable using facial expressions alone but are better distinguished by other body cues (Aviezer, Trope, & Todorov, 2012). Consequently, many studies employ several different sensors concurrently to identify emotional states of users with better detection rates (e.g., Mandryk & Atkins, 2007; Selvaraj, Murugappan, Wan, & Yaacob, 2013). Similarly, the detection of motivational states, such as flow (Csikszentmihalyi, 1990), benefits from a multimodal classification approach using physiological data. Using facial EMG activity and electrodermal activity (EDA), flow was found to be associated with positive valence and increased arousal (Nacke & Lindley, 2008), a result that is also usually found by conventional post hoc questionnaires (e.g., Kiili, Lindstedt, & Ninaus, 2018).

Studies combining physiological and behavioral data are much rarer, particularly in GBA. In one of these rare studies, eye tracking, in-game assessments, and conventional log files provided a comprehensive look into the cognitive and metacognitive processes underlying successful learning with a GBL environment (Taub et al., 2017). Specifically, the authors combined eye tracking data with specific in-game behaviors (i.e., conversations with non-player characters, collecting items, usage of in-game worksheets, and reading of in-game notes or books, respectively) to assess how well players performed on in-game assessments. In order to not be overwhelmed by this "flood" of data but rather benefit from it, the authors employed multilevel modeling (Raudenbush & Bryk, 2002), allowing fine-grained analyses of user states. Importantly, their analyses revealed that analysis of single data channels does not reveal the full picture of how users interacted with and learned within the GBL environment. In particular, the authors identified different results when including many data channels as compared to the analysis of main effects of single data channels.

What becomes evident from such studies is that the combination of behavioral processes and physiological data does not only yield benefits but, in fact, is also necessary to assess learning-relevant user states in complex or realistic learning scenarios, including—but not limited to—GBL environments. Moreover, sophisticated analytical techniques are required to deal with data from multiple sources. Consequently, while there is huge potential for widespread collection of behavioral processes and physiological data, relevant stakeholders will also need appropriate tools to analyze these large amounts of data.

8.7 Suggestions for the Future of Process and Physiological Data Within GBA

After highlighting the emerging changes within the field, it is essential to enumerate the future challenges that will need to be addressed. Derived from the previous argumentation, we propose 10 suggestions that should accompany and support future use and development of processes and physiological data within GBA. First, the question of how scientists can identify crucial patterns within large amounts of data needs to be discussed. For this, we postulate (1) that more case studies with an emphasis on data analysis are needed. The number of applications of GBA suggests, however, that the aspect of data evaluation has not yet been adequately validated. Additionally, (2) open science needs to be embraced within the field of GBA. Comprehensible and transparent datasets should become the norm. Thus, researchers can comprehend existing approaches and refine their own methods of data analysis prior to their implementation. This might also lead to (3) the development and/or consequent application of analytical methods that extend descriptive statistics. Through this, not only mistakes (e.g., the alpha or type I error, i.e., assuming an effect whereas there is actually none) might be prevented, but also important information (e.g., effect sizes) could be observed more frequently. Furthermore, methods such as *Bayesian statistics* (Box & Tiao, 2011) might be applicable, especially as open data should support the creation of essential prior distributions. Alternatively, methods such as neural networks (Rey & Wender, 2011) might support the evaluation of potentially overwhelming datasets. However, this should not interfere with a further (4) standardization of data analysis within GBA. In order to create comparable information and to advance the field as a whole, researchers need to find common ground for the desired measures. Derived from this suggestion, we postulate that (5) the scientific standardization should result in clear theory-driven guidance and user manuals explaining how to interpret individual measures. This might be essential to prevent misinterpretations, to provide essential reference frames, and to foster access to these insights for nonexperts. Furthermore, widely published practical recommendations are important, as not all measures might be applicable in every scenario (e.g., movements might interfere with GSR measures). Additionally, a (6) trade-off between accuracy and approachability has to be determined. This applies to the software (e.g., increased noise reduction) and to the hardware that are used (e.g., cheaper measures with lower resolution). Alternatively, (7) research regarding defining patterns has to be intensified (e.g., Kang, Liu, & Qu, 2017). For example, within the introduced example of *Counter Strike*, not each time span of a match might be equally interesting. However, a reliable classification is needed to determine which moment might be prototypical or especially noteworthy. This research could contribute to the suggested trade-off, as the relevant data material might be significantly reduced. Overall, (8) a strong focus on teachers as an important target audience is needed, especially in the light of complex data collection and evaluation that might result from the perspective of process data and physiological measures. If the tools cannot find their way into the classroom, they lack external validity. However, first approaches have been made (e.g., Calvo-Morata, Alonso-Fernández, Freire, Martínez-Ortiz, & Fernández-Manjón, 2018) and with the increasing attention focused on GBA, a potential market for user-friendly backend solutions might emerge. Finally, we postulate (9) that, in order to increase the potential of GBA in combination with process and physiological data, learning with educational video games itself needs to be supported much more. Conclusions from game data might be more rewarded and pursued after GBL has become accepted and widespread. This might need political involvement, but more importantly, it will require (10) training for teachers so they can use GBL and GBA effectively.

8.8 Conclusions

In this chapter, we focused on the use of the behavioral process, physiological data, and their combination in GBL and GBA. We outlined the benefits of investigating cognitive, emotional, and motivational states on a process and more fine-grained level with these data as compared to simple pre- or post-game measures. One of the major benefits of utilizing these kinds of data is the ability to continuously monitor learning-relevant processes during play without probing the learners. On the one hand, behavioral process and physiological data can be utilized to inform theoretical models of GBL and GBA and fundamentally advance the field. On the other hand, these data allow for real-time assessment and personalization to foster learning and increase the validity of GBA beyond a simple final score. For instance, a deeper understanding of the processes going on during learning might be particularly beneficial for underperforming individuals adapting the learning system to their specific needs (e.g., Xiao & Wang, 2016). We also tried to point out its current limitations, for instance, the lack of multichannel studies combining behavioral and physiological process data. However, recent technological trends might help to increase adoption of multichannel studies as it becomes easier to create games as well as to integrate multichannel data acquisition and processing.

Overall, we foresee that GBA and GBL can benefit greatly by using different types of data from various sources not only to improve assessment and personalization but also to foster maturation of the whole GBL field by addressing the conceptual, theoretical, and methodological issues that are currently undermining the value of GBL. For this to happen, however, we believe several steps need to be taken, for instance, standardization of data analysis, as well as the development of approachable tools for the target audience to handle and interpret complex data.

The use of behavioral process and physiological data for GBA is by no means a trivial endeavor, and adoption of these methods is still limited, but the potential of these data is undoubtedly high. At the same time, other relevant issues need to be considered more explicitly in the future. The acquisition of numerous rather personal data requires careful considerations regarding ethics and data protection, as well as the involvement of relevant stakeholders (e.g., Drachsler & Greller, 2016). One also needs to decide whether a comprehensive multichannel setup is always necessary as this is most often related to higher costs (e.g., equipment and effort). Future studies are needed to determine which specific GBA scenarios might benefit more from such an approach than others. Much work remains to be done to maximally benefit from combining behavioral process and physiological data in GBA. We hope the perspective of this chapter will also be beneficial for educators interested in understanding the potential of behavioral process and physiological data for GBA, which, in turn, might increase adoption of educational games in the classroom because we have demonstrated that multichannel data offer many advantages over conventional pre- and post-game measures, final scores, or grades.

Acknowledgments The current research was funded by the Leibniz-Competition Fund (SAW-2016-IWM-3) and the Leibniz-WissenschaftsCampus "Cognitive Interfaces" (MWK-WCT TP12) supporting Manuel Ninaus.

References

Admiraal, W., Huizenga, J., Akkerman, S., & Ten Dam, G. (2011). The concept of flow in collaborative game-based learning. *Computers in Human Behavior, 27*(3), 1185–1194. https://doi.org/10.1016/j.chb.2010.12.013

Alexander, P. A. (2018). Past as prologue: Educational psychology's legacy and progeny. *Journal of Educational Psychology, 110*(2), 147–162. https://doi.org/10.1037/edu0000200

Allison, B. Z., & Polich, J. (2008). Workload assessment of computer gaming using a single-stimulus event-related potential paradigm. *Biological Psychology, 77*(3), 277–283. https://doi.org/10.1016/j.biopsycho.2007.10.014

Aviezer, H., Trope, Y., & Todorov, A. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science, 338*(6111), 1225–1229. https://doi.org/10.1126/science.1224313

Azevedo, R., Harley, J., Trevors, G., Duffy, M., Feyzi-Behnagh, R., Bouchet, F., & Landis, R. (2013). Using trace data to examine the complex roles of cognitive, metacognitive, and emotional self-regulatory processes during learning with multi-agent systems. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (Vol. 28, pp. 427–449). New York, NY: Springer. https://doi.org/10.1007/978-1-4419-5546-3_28

Baumgartner, T., Speck, D., Wettstein, D., Masnari, O., Beeli, G., & Jäncke, L. (2008). Feeling present in arousing virtual reality worlds: Prefrontal brain regions differentially orchestrate presence experience in adults and children. *Frontiers in Human Neuroscience, 2*, 1–12. https://doi.org/10.3389/neuro.09.008.2008

Baumgartner, T., Valko, L., Esslen, M., & Jäncke, L. (2006). Neural correlate of spatial presence in an arousing and noninteractive virtual reality: An EEG and psychophysiology study. *Cyberpsychology & Behavior, 9*(1), 30–45. https://doi.org/10.1089/cpb.2006.9.30

Berta, R., Bellotti, F., De Gloria, A., Pranantha, D., & Schatten, C. (2013). Electroencephalogram and physiological signal analysis for assessing flow in games. *IEEE Transactions on Computational Intelligence and AI in Games, 5*(2), 164–175. https://doi.org/10.1109/TCIAIG.2013.2260340

Bower, M., & Sturman, D. (2015). What are the educational affordances of wearable technologies? *Computers & Education, 88*, 343–353. https://doi.org/10.1016/j.compedu.2015.07.013

Box, G. E. P., & Tiao, G. C. (2011). *Bayesian inference in statistical analysis*. New York, NY: Wiley.

Boyle, E. A., Hainey, T., Connolly, T. M., Gray, G., Earp, J., Ott, M., … Pereira, J. (2016). An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. *Computers & Education, 94*, 178–192. https://doi.org/10.1016/j.compedu.2015.11.003

Calvo-Morata, A., Alonso-Fernández, C., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2018). Making understandable game learning analytics for teachers. In G. Hancke, M. Spaniol, K. Osathanunkul, S. Unankard, & R. Klamma (Eds.), *Advances in Web-Based Learning – ICWL 2018* (pp. 112–121). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-96565-9_11

Charles, S. T., Reynolds, C. A., & Gatz, M. (2001). Age-related differences and change in positive and negative affect over 23 years. *Journal of Personality and Social Psychology, 80*(1), 136–151. https://doi.org/10.1037/0022-3514.80.1.136

Charles University, Czech Academy of Sciences. (2017). *Attentat 1942* [Computer Software]. Prague, Czechoslovakia: Author.
Cohn, J. F., Ambadar, Z., & Ekman, P. (2007). Observer-based measurement of facial expression with the Facial Action Coding System. In J. A. Coan & J. J. B. Allen (Eds.), *Handbook of emotion elicitation and assessment. Series in affective science* (pp. 203–221). New York, NY: Oxford University Press.
Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York, NY: Harper Perennial.
Dasgupta, S., Clements, S. M., Idlbi, A. Y., Willis-Ford, C., & Resnick, M. (2015). Extending Scratch: New pathways into programming. In *2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (pp. 165–169). Atlanta, GA: IEEE. https://doi.org/10.1109/VLHCC.2015.7357212
Drachsler, H., & Greller, W. (2016). Privacy and analytics: It's a delicate issue a checklist for trusted learning analytics. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16* (pp. 89–98). New York, NY: ACM. https://doi.org/10.1145/2883851.2883893
Epic Games. (2017). *Fortnite* [Computer Software]. Cary, NC: Author.
Epic Games. (2018). *Unreal Engine (Version 4)* [Computer Software]. Cary, NC: Author.
Eysink, T. H., de Jong, T., Berthold, K., Kolloffel, B., Opfermann, M., & Wouters, P. (2009). Learner performance in multimedia learning arrangements: An analysis across instructional approaches. *American Educational Research Journal, 46*(4), 1107–1149. https://doi.org/10.3102/0002831209340235
Ferrari, M., & Quaresima, V. (2012). A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *NeuroImage, 63*(2), 921–935. https://doi.org/10.1016/j.neuroimage.2012.03.049
Freeman, J., Avons, S. E., Pearson, D. E., & IJsselsteijn, W. A. (1999). Effects of sensory information and prior experience on direct subjective ratings of presence. *Presence: Teleoperators & Virtual Environments, 8*(1), 1–13. https://doi.org/10.1162/105474699566017
GameAnalytics. (2016). *GameAnalytics* [Computer Software]. Copenhagen, Denmark: Author.
Girouard, A., Solovey, E. T., Hirshfield, L. M., Chauncey, K., Jacob, R. J. K., Sassaroli, A., … Jacob, R. J. K. (2009). Distinguishing difficulty levels with non-invasive brain activity measurements. In T. Gross, J. Gulliksen, P. Kotzé, L. Oestreicher, & P. Palanque (Eds.), *Human-Computer Interaction – INTERACT 2009* (pp. 440–452). Heidelberg, Germany: Springer. https://doi.org/10.1007/978-3-642-03655-2_50
GitHub. (2018). *GitHub* [Computer Software]. San Francisco, CA: Author.
Granic, I., Lobel, A., & Engels, R. C. (2014). The benefits of playing video games. *American Psychologist, 69*(1), 66–78. https://doi.org/10.1037/a0034857
Greipl, S., Ninaus, M., Bauer, D., Kiili, K., & Moeller, K. (2018). A fun-accuracy trade-off in game-based learning. In M. Gentile, M. Allegra, & H. Söbke (Eds.), *International Conference on Games and Learning Alliance – Lecture Notes in Computer Science* (pp. 167–177). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-11548-7_16
Guo, G., & Dyer, C. R. (2005). Learning from examples in the small sample case: Face expression recognition. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics, 35*(3), 477–488. https://doi.org/10.1109/TSMCB.2005.846658
Hamari, J., Shernoff, D. J., Rowe, E., Coller, B., Asbell-Clarke, J., & Edwards, T. (2016). Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Computers in Human Behavior, 54*, 170–179. https://doi.org/10.1016/j.chb.2015.07.045
Hattahara, S., Fujii, N., Nagae, S., Kazai, K., & Katayose, H. (2008). Brain activity during playing video game correlates with player level. In *Proceedings of the 2008 International Conference on Advances in Computer Entertainment Technology* (pp. 360–363). New York, NY: ACM. https://doi.org/10.1145/1501750.1501835

Howard-Jones, P., & Jay, T. (2016). Reward, learning and games. *Current Opinion in Behavioral Sciences, 10*, 65–72. https://doi.org/10.1016/j.cobeha.2016.04.015

Kalyuga, S., & Singh, A. M. (2016). Rethinking the boundaries of cognitive load theory in complex learning. *Educational Psychology Review, 28*(4), 831–852. https://doi.org/10.1007/s10648-015-9352-0

Kang, J., Liu, M., & Qu, W. (2017). Using gameplay data to examine learning behavior patterns in a serious game. *Computers in Human Behavior, 72*, 757–770. https://doi.org/10.1016/j.chb.2016.09.062

Kiili, K., Lindstedt, A., & Ninaus, M. (2018). Exploring characteristics of students' emotions, flow and motivation in a math game competition. In J. Koivisto & J. Hamari (Eds.), *Proceedings of the 2nd International GamiFIN Conference* (pp. 20–29). Pori, Finland: CEUR Workshop Proceedings.

Kivikangas, J. M. (2006). *Psychophysiology of flow experience: An explorative study* (Master's thesis). Retrieved from http://urn.fi/URN:NBN:fi-fe20061271

Klasen, M., Weber, R., Kircher, T. T. J., Mathiak, K. A., & Mathiak, K. (2012). Neural contributions to flow experience during video game playing. *Social Cognitive and Affective Neuroscience, 7*(4), 485–495. https://doi.org/10.1093/scan/nsr021

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education, 57*(2), 1813–1824. https://doi.org/10.1016/j.compedu.2011.02.003

Kober, S. E., & Neuper, C. (2012). Using auditory event-related EEG potentials to assess presence in virtual reality. *International Journal of Human-Computer Studies, 70*(9), 577–587. https://doi.org/10.1016/j.ijhcs.2012.03.004

Lamnek, S., & Krell, C. (2016) *Qualitative Sozialforschung* [*Qualitative social research*]. Weinheim, Germany: Beltz Verlagsgruppe.

Littlewort, G. C., Bartlett, M. S., Salamanca, L. P., & Reilly, J. (2011). Automated measurement of children's facial expressions during problem solving tasks. *Face and Gesture, 2011*, 30–35. https://doi.org/10.1109/FG.2011.5771418

Liu, Y., Kosmadoudi, Z., Sung, R. C. W., Lim, T., Louchart, S., & Ritchie, J. (2010). Capture user emotions during computer- aided design. In *Proceedings of the Integrated Design and Manufacturing in Mechanical Engineering (IDMME) and Virtual Conference* (pp. 2–4).

Lomas, J. D., Koedinger, K., Patel, N., Shodhan, S., Poonwala, N., & Forlizzi, J. L. (2017). Is difficulty overrated? The effects of choice, novelty and suspense on intrinsic motivation in educational games. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 1028–1039). New York, NY: ACM. https://doi.org/10.1145/3025453.3025638

Maloney, J., Peppler, K., Kafai, Y. B., Resnick, M., & Rusk, N. (2008). Programming by choice: Urban youth learning programming with scratch. In *Proceedings of the 39th SIGCSE Technical Symposium on Computer Science Education* (pp. 367–371). Portland, OR: ACM. https://doi.org/10.1145/1352135.1352260

Maloney, J., Resnick, M., Rusk, N., Silverman, B., & Eastmond, E. (2010). The scratch programming language and environment. *ACM Transactions on Computing Education, 10*(4), 1–15. https://doi.org/10.1145/1868358.1868363

Mandryk, R. L., & Atkins, M. S. (2007). A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International Journal of Human-Computer Studies, 65*(4), 329–347. https://doi.org/10.1016/j.ijhcs.2006.11.011

Mandryk, R. L., Atkins, M. S., & Inkpen, K. M. (2006). A continuous and objective evaluation of emotional experience with interactive play environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems CHI 06* (pp. 1027–1036). New York, NY: ACM. https://doi.org/10.1145/1124772.1124926

Marklund, B. B., Backlund, P., & Johannesson, M. (2013). Children's collaboration in emergent game environments. In *Proceedings of the 8th International Conference on the Foundations of Digital Games* (pp. 306–313). New York, NY: ACM.

Massachusetts Institute of Technology. (2012). *MIT App Inventor* [Computer Software]. Cambridge, MA: Author.

Mayer, R. E. (2014). *Computer games for learning: An evidence-based approach*. Cambridge, MA: MIT.

Mayer, R. E. (2015). On the need for research evidence to guide the design of computer games for learning. *Educational Psychologist, 50*(4), 349–353. https://doi.org/10.1080/00461520.2015.1133307

Mayer, R. E. (2018). Educational psychology's past and future contributions to the science of learning, science of instruction, and science of assessment. *Journal of Educational Psychology, 110*(2), 174–179. https://doi.org/10.1037/edu0000195

Microsoft. (2009). *Kodu* [Computer Software]. Redmond, WA: Author.

Mionix. (2018). *Naos QG* [Apparatus and Software]. Växjö, Sweden: Author.

Mojang. (2018). *MinecraftEdu* [Computer Software]. Stockholm, Sweden: Author.

Nacke, L., & Lindley, C. A. (2008). Flow and immersion in first-person shooters. In *Proceedings of the 2008 Conference on Future Play Research, Play, Share - Future Play '08* (pp. 81–88). New York, NY: ACM. https://doi.org/10.1145/1496984.1496998

Nacke, L. E., Grimshaw, M. N., & Lindley, C. A. (2010). More than a feeling: Measurement of sonic user experience and psychophysiology in a first-person shooter game. *Interacting with Computers, 22*(5), 336–343. https://doi.org/10.1016/j.intcom.2010.04.005

Nebel, S. (2017). *Investigating the mechanisms of competition within educational video games - Die Mechanismen des Wettbewerbs in digitalen Lernspielen* (Doctoral dissertation). https://doi.org/10.13140/RG.2.2.28445.41440.

Nebel, S., Beege, M., Schneider, S., & Rey, G. D. (2016). The higher the score, the higher the learning outcome? Heterogeneous impacts of leaderboards and choice within educational videogames. *Computers in Human Behavior, 65*, 391–401. https://doi.org/10.1016/j.chb.2016.08.042

Nebel, S., Schneider, S., Beege, M., & Rey, G. D. (2017). Leaderboards within educational videogames: The impact of difficulty, effort and gameplay. *Computers & Education, 113*, 28–41.

Nebel, S., Schneider, S., & Rey, G. D. (2016). Mining learning and crafting scientific experiments: A literature review on the use of Minecraft in education and research. *Journal of Educational Technology & Society, 19*(2), 355–366.

Nebel, S., Schneider, S., Schledjewski, J., & Rey, G. D. (2017). Goal-setting in educational video games: Comparing goal-setting theory and the goal-free effect. *Simulation & Gaming, 48*(1), 98–130. https://doi.org/10.1177/1046878116680869

Ninaus, M., Kober, S. E., Friedrich, E. V. C., Dunwell, I., De Freitas, S., Arnab, S., … Neuper, C. (2014). Neurophysiological methods for monitoring brain activity in serious games and virtual environments: A review. *International Journal of Technology Enhanced Learning, 6*(1), 78–103. https://doi.org/10.1504/IJTEL.2014.060022

Ninaus, M., Kober, S. E., Friedrich, E. V. C., Neuper, C., & Wood, G. (2014). The potential use of neurophysiological signals for learning analytics. In *2014 6th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)* (pp. 1–5). Valletta, Malta: IEEE. https://doi.org/10.1109/VS-Games.2014.7012169

Ninaus, M., Moeller, K., McMullen, J., & Kiili, K. (2017). Acceptance of game-based learning and intrinsic motivation as predictors for learning success and flow experience. *International Journal of Serious Games, 4*(3), 15–30. https://doi.org/10.17083/ijsg.v4i3.176

Ninja Theory. (2017). *Hellblade: Senua's Sacrifice* [Computer Software]. Cambridge, UK: Author.

Nourbakhsh, N., Chen, F., Wang, Y., & Calvo, R. A. (2017). Detecting users' cognitive load by galvanic skin response with affective interference. *ACM Transactions on Interactive Intelligent Systems, 7*(3), 1–20. https://doi.org/10.1145/2960413

Novak, E., & Johnson, T. E. (2012). Assessment of student's emotions in game-based learning. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning* (pp. 379–399). New York, NY: Springer. https://doi.org/10.1007/978-1-4614-3546-4_19

Nyamsuren, E., Van der Vegt, W., & Westera, W. (2017). Automated adaptation and assessment in serious games: A portable tool for supporting learning. In M. Winands, H. van den Herik, & W. Kosters (Eds.), *Advances in computer games* (pp. 201–212). Cham, Switzerland: Springer.

Paas, F., & Sweller, J. (2014). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 27–42). Cambridge, MA: Cambridge University Press.

Peifer, C. (2012). Psychophysiological correlates of flow-experience. In S. Engeser (Ed.), *Advances in flow research* (pp. 139–164). New York, NY: Springer. https://doi.org/10.1007/978-1-4614-2359-1

Pekrun, R. (2011). Emotions as drivers of learning and cognitive development. In R. A. Calvo & S. K. D'Mello (Eds.), *New perspectives on affect and learning technologies* (pp. 23–39). New York, NY: Springer. https://doi.org/10.1007/978-1-4419-9625-1

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology, 36*(1), 36–48. https://doi.org/10.1016/j.cedpsych.2010.10.002

Pellouchoud, E., Smith, M. E., McEvoy, L., & Gevins, A. (1999). Mental effort-related EEG modulation during video-game play: Comparison between juvenile subjects with epilepsy and normal control subjects. *Epilepsia, 40*(s4), 38–43. https://doi.org/10.1111/j.1528-1157.1999.tb00905.x

Perez-Colado, I., Alonso-Fernandez, C., Freire, M., Martinez-Ortiz, I., & Fernandez-Manjon, B. (2018). Game learning analytics is not informagic! In *2018 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1729–1737). Tenerife, Spain: IEEE. https://doi.org/10.1109/EDUCON.2018.8363443

Perttula, A., Kiili, K., Lindstedt, A., & Tuomi, P. (2017). Flow experience in game based learning – A systematic literature review. *International Journal of Serious Games, 4*(1). https://doi.org/10.17083/ijsg.v4i1.151

Plass, J. L., & Kaplan, U. (2016). Emotional design in digital media for learning. In S. Tettegah & M. Gartmeier (Eds.), *Emotions, technology, design, and learning* (pp. 131–161). New York, NY: Elsevier. https://doi.org/10.1016/B978-0-12-801856-9.00007-4

Pugnetti, L., Mendozzi, L., Barbieri, E., Rose, F. D., Attree, E. A., & Barberi, E. (1996). Nervous system correlates of virtual reality experience. In P. M. Sharkey (Ed.), *Proceedings of the First European Conference on Disability, Virtual Reality and Associated Technology* (pp. 239–246). Maidenhead, UK: The University of Reading.

Ratcliffe, D. (2017). *ComputercraftEdu* [Computer software]. Cambridge, UK: Author.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Rey, G. D., & Wender, K. F. (2011). *Neuronale Netze: eine Einführung in die Grundlagen, Anwendungen und Datenauswertung*. Bern, Switzerland: Huber.

Salminen, M., & Ravaja, N. (2007). Oscillatory brain responses evoked by video game events: The case of super monkey ball 2. *Cyberpsychology & Behavior, 10*(3), 330–338. https://doi.org/10.1089/cpb.2006.9947

Salminen, M., & Ravaja, N. (2008). Increased oscillatory theta activation evoked by violent digital game events. *Neuroscience Letters, 435*(1), 69–72. https://doi.org/10.1016/j.neulet.2008.02.009

Schneider, J., Börner, D., Van Rosmalen, P., & Specht, M. (2015). Augmenting the senses: A review on sensor-based learning support. *Sensors, 15*(2), 4097–4133. https://doi.org/10.3390/s150204097

Schneider, S., Nebel, S., & Rey, G. D. (2016). Decorative pictures and emotional design in multimedia learning. *Learning and Instruction, 44*, 65–73. https://doi.org/10.1016/j.learninstruc.2016.03.002

Selvaraj, J., Murugappan, M., Wan, K., & Yaacob, S. (2013). Classification of emotional states from electrocardiogram signals: A non-linear approach based on hurst. *Biomedical Engineering Online, 12*(1), 44. https://doi.org/10.1186/1475-925X-12-44

Shute, V., & Wang, L. (2016). Assessing and supporting hard-to-measure constructs in video games. In A. A. Rupp & J. P. Leighton (Eds.), *The Wiley handbook of cognition and assessment* (pp. 535–562). Hoboken, NJ: Wiley.
Shute, V. J., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. Cambridge, MA: MIT.
Smith, S. P., Blackmore, K., & Nesbitt, K. (2015). A meta-analysis of data collection in serious games research. In C. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics* (pp. 31–55). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-05834-4_2
Solovey, E., Schermerhorn, P., Scheutz, M., Sassaroli, A., Fantini, S., & Jacob, R. (2012). Brainput: Enhancing interactive systems with streaming fnirs brain input. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems - CHI '12* (pp. 2193–2202). New York, NY: ACM. https://doi.org/10.1145/2207676.2208372
Strait, M., Canning, C., & Scheutz, M. (2013). Limitations of NIRS-based BCI for realistic applications in human-computer interaction. In *Proceedings of the Fifth International Brain-Computer Interface Meeting* (pp. 2–3). Graz, Austria: Graz University of Technology Publishing House. https://doi.org/10.3217/978-3-85125-260-6-2
Sweller, J. (1994). Cognitive load theory, learning difficulty and instructional design. *Learning and Instruction, 4*(4), 295–312. https://doi.org/10.1016/0959-4752(94)90003-5
Taub, M., Mudrick, N. V., Azevedo, R., Millar, G. C., Rowe, J., & Lester, J. (2017). Using multichannel data with multi-level modeling to assess in-game performance during gameplay with Crystal Island. *Computers in Human Behavior, 76*, 641–655. https://doi.org/10.1016/j.chb.2017.01.038
Thauros-Clan. (2016). *Brain computer interface plugin* [Computer Software]. Author.
Valve Corporation. (2012). *Counter Strike: Global Offensive* [Computer Software]. Bellevue, WA: Author.
Vorderer, P., Wirth, W., Gouveia, F. R., Biocca, F., Saari, T., Jäncke, F., … Jäncke, P. (2004). *MEC Spatial Presence Questionnaire (MECSPQ): Short documentation and instructions for application.* Report to the European Community, Project Presence: MEC (IST-2001-37661). Retrieved from https://www.researchgate.net/publication/318531435_MEC_spatial_presence_questionnaire_MEC-SPQ_Short_documentation_and_instructions_for_application
Vorderer, P., Wirth, W., Saari, T., Gouveia, F. R., Biocca, F., Jäncke, F., … Jäncke, P. (2003). *Constructing presence: Towards a two-level model of the formation of Spatial Presence*. Unpublished report to the European Community, Project Presence: MEC (IST-2001-37661). Hannover, Munich, Helsinki, Porto, Zurich.
Wirzberger, M., Herms, R., Esmaeili Bijarsari, S., Rey, G. D., & Eibl, M. (2017). Influences of cognitive load on learning performance, speech and physiological parameters in a dual-task setting. In *Poster session presented at the meeting of the 20th Conference of the European Society for Cognitive Psychology*, Potsdam, Germany.
Wise, R. A. (2004). Dopamine, learning and motivation. *Nature Reviews Neuroscience, 5*(6), 483–494. https://doi.org/10.1038/nrn1406
Witte, M., Ninaus, M., Kober, S. E. S. E., Neuper, C., & Wood, G. (2015). Neuronal correlates of cognitive control during gaming revealed by near-infrared spectroscopy. *PLoS One, 10*(8), e0134816. https://doi.org/10.1371/journal.pone.0134816
Wu, C. H., Huang, Y. M., & Hwang, J. P. (2016). Review of affective computing in education/learning: Trends and challenges. *British Journal of Educational Technology, 47*(6), 1304–1323. https://doi.org/10.1111/bjet.12324
Xiao, X., & Wang, J. (2016). Context and cognitive state triggered interventions for mobile MOOC learning. In *ICMI '16: Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 378–385). New York, NY: ACM. https://doi.org/10.1145/2993148.2993177
Xue, S., Wu, M., Kolen, J., Aghdaie, N., & Zaman, K. A. (2017). Dynamic difficulty adjustment for maximized engagement in digital games. In *WWW '17 Companion: Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 465–471). Geneva, Switzerland: International World Wide Web Conferences Steering Committee and Republic and Canton of Geneva. https://doi.org/10.1145/3041021.3054170

Chapter 9
A Provisional Framework for Multimodal Evaluation: Establishing Serious Games Quality Label for Use in Training and Talent Development

Wee Hoe Tan and Ivan Boo

9.1 Introduction

The idea of using games for serious purposes was proposed by Abt in 1970s, in which he delineated serious games from simulations. All serious games simulate something from the real world, but not all simulations are games because the outcomes of simulations are predetermined results, as opposed to winning or losing outcome in games (Abt, 1970). However, it was the Serious Games Initiative that reintroduced the notion of serious games (Serious Games Society, 2008), where Sawyer and Smith (2008) presented the taxonomy of serious games, providing an overview of the status quo of the serious games industry in the last decade. Then the potential of games was exploited extensively in various fields of study, namely education (Gloria, Bellotti, Berta, & Lavagnino, 2014), military (US Army, 2017), business (Popescu, Romero, & Usart, 2013), and healthcare (Wattanasoontorn, Boada, Garcia, & Sbert, 2013).

9.1.1 Evaluation of Custom-Made and Commercial Off-the-Shelf Serious Games

When setting up the Serious Games Institute at the Coventry University in mid-2000, de Freitas (2006) predicted widespread use of game technologies in formal contexts through a serious games movement. Nonetheless, the potential of games has been

W. H. Tan (✉)
Faculty of Art, Computing and Creative Industry, Sultan Idris Education University, Tanjong Malim, Perak, Malaysia
e-mail: whtan@fskik.upsi.edu.my

I. Boo
Serious Games Association, Singapore, Singapore
e-mail: ivan@seriousgamesasia.com

D. Ifenthaler, Y. J. Kim (eds.), *Game-Based Assessment Revisited*, Advances in Game-Based Learning, https://doi.org/10.1007/978-3-030-15569-8_9

justified for having positive effects, especially when its use was meant for engaging targeted players through the use of flow channel (Csikzentmihalyi, 2008). In the game industry, the concept of the flow channel is commonly used when balancing learning curve to ensure that challenges in the game are neither too difficult nor too easy in relation to the knowledge and skills accumulated by players over time (Adams, 2013).

The balancing is particularly important for serious games since learning outcomes in the game world should have been constructively aligned with transferrable knowledge and skills in real life. In this sense, the extent to which the learning contents are aligned to the game elements should be validated by subject matter experts when evaluating the quality of serious games. The validation could be done by verifying the alignment between every goal in a serious game and the three components of an intended learning outcome, i.e., (1) an observable behavior or action taken by a player or game avatar when attaining the outcome, (2) conditions or rules which limit a player or game avatar in the game world when attaining the outcome, and (3) the degree to which a player or game avatar attains the outcome (Tan, 2015). However, this content validation approach is only feasible if the serious games are custom-made for a specific purpose, assuming that correction or revision can be done after the validation. For non-custom-made serious games, such as commercial off-the-shelf games, it might not be economic to revise or modify the games before applying in a supposed context. In this case, academics are still arguing for a unified approach to evaluating such serious games (Brady, Devitt, & Jameson, 2016).

The evaluation approach of ready-made serious games, alongside the content validation methods for custom-made serious games, should be integrated into a comprehensive evaluation framework. This framework needs to cover the selection, organization, and presentation of serious games, in which expert knowledge would be required to determine their quality for specific purposes. This evaluation framework can be used by game developers as a guideline to follow in the process of producing quality serious games. At the same time, the framework can also guide trainers and subject matter experts to get themselves prepared when using serious games as training materials or managing serious games training programs.

Formal evaluation of serious games can justify the return on investment (Falstein, 2007); thus the importance of a reliable evaluation framework cannot be overstressed, especially in selecting serious games for health (Dutch Game Garden, 2016). Without a framework for validating and evaluating games for use in the healthcare and wellness industry, there will be different standards or, worse, no standard to assure the safety of use on patients, the credential of medical professionals, and the risk-free hospital conditions. The quality assurance of serious games is essential to hospitals that embraced the practice of game-based treatment or prevention of illness.

Standardized evaluation of serious games is also important for training or developing talents in the healthcare industry. For instance, a standardized evaluation in training using games is crucial to avoid the Dunning-Kruger effect (Dunning, Johnson, Ehrlinger, & Kruger, 2003). According to Dunning et al. (2003), after winning in the game world, low-ability learners may suffer from illusory superiority, while high-ability trainees may underestimate their relative competence in the real world. To assure trainees are aware of the Dunning-Kruger effect and subsequently hinder the effect, a multimodal serious games evaluation framework is proposed in this chapter for the healthcare and wellness industry.

9.1.2 Serious Games for Training and Talent Development

During the last decade, the educational simulation used in the healthcare industry has arguably matured (McDougall, 2015). Simulation no longer represents a novelty to clinical education (Aebersold, 2018). Rather, it is now a vetted standard educational practice for nursing and medical education at various levels (Jones, Passos-Neto, & Braghiroli, 2015). Manikin-based simulation laboratories are now ubiquitous features in all areas of clinical nurse education (Cant & Cooper, 2017). In particular, this chapter regards simulation-based education as a template that provides a starting point for educators in the healthcare industry to integrate innovative technology into training and talent development. Such innovative technology could be digital games, mobile apps, virtual reality (VR), augmented reality (AR), or other cutting-edge computing applications that can be aligned to simulation-based education in order to achieve specific training outcomes.

Recent advances in technology have led to an explosion of sophisticated multimedia and digital educational content. Although high-fidelity manikin-based simulators remain relatively expensive (Lin, Cheng, Hecker, Grant, & Currie, 2018), delivery platforms that support complex digital environments, games, and mobile apps are increasingly more accessible and affordable to students, faculty, and researchers. Compared to the last two decades, the integration of manikin-based simulators with innovative technologies is becoming common and ubiquitous. Trainers who have embraced game-based learning, mobile apps, and virtual reality are in a better position to advance clinical curricula. This in turn better prepares trainees for modern and future practice in the healthcare industry.

9.1.3 Serious Games Quality Labelling

With more games coming in the market, it is crucial to establish a serious games evaluation framework with clear and transparent guidelines, which will help in tagging serious games with a quality label. The quality labelling will serve as a guide, not only for users who conduct or manage game-based training but also for serious games developers, particularly in Asia Pacific which generated the highest revenue in the global serious games market in 2016 (Sonawane, 2017). As a user, hospital corporate buyers would need to know if the accuracy of game content had been verified. It would be detrimental if a game is providing inaccurate information when used as a training tool as it would have a direct impact on patient care and safety.

The proposed quality label is a tag that will enable users to identify an overall standard established by the producers of the game. Games are tagged within a use-case category, based on a set of established criteria. In contrast to a statement claimed by a game developer or a service provider, an independent third party assigns a quality label to a serious game that meets the established criteria. Eventually, the quality-labelling mechanism would provide accurate and impartial information about the quality of serious games to consumers. The impact of the serious games quality labelling is manifested at two levels, i.e., the industry level and the individual level.

The first level applies to the serious games industry, where there is a need to normalize the process of developing serious games for easy deployment to target users and for sustaining the scalability of deployment. The normalized development process would enable serious games to be an emerging technology for training professionals in the healthcare industry. With a growing and sustaining demand for serious games that offer "a realistic environment for training and development activities in the defense, education, healthcare and others on employee engagement solutions" (Sonawane, 2017), there will be an interest to set up companies or studios that can survive or even prosper by focusing on serious games business. This will further help in promoting greater research and development (R&D) interest in the healthcare and training industry. As the industry expands, there will be more serious games titles in the market, so the quality labelling of serious games will give assurance to hospital corporate buyers when they are purchasing specific game titles with appropriate quality.

The second level applies to individuals who take part in the quality-labelling practice. Essentially, the process of acquiring the skills to label serious games using the quality-labelling mechanism should neither be lengthy nor financially costly. Thus, the quality-labelling practice should be straight forward and affordable because this can motivate serious games developers, instructional designers, and subject matter experts to have their games labelled for quality. Individuals who intend to evaluate serious games should be certified in their respective fields, endorsing their own domain expertise. The certifications should also reflect their competence in a standardized serious games development process. To initiate the practice, these individuals can become members or associate members of the Serious Games Association, where they can work jointly with the instructional designer community and the healthcare domain expert community toward establishing the framework and creating the quality label. Figure 9.1 depicts how two levels of serious games quality labelling interact with each other through the proposed evaluation framework. Individuals who are from serious games developer community, instructional designer community, and healthcare domain expert community create quality labels to establish the framework. Then, they participate in expert evaluation and scientific validation to label games submitted by the game industry in a neutral games repository. Eventually, target users may acquire serious games in the repository by referring to quality labels.

9.2 A Provisional Multimodal Evaluation Framework

A provisional evaluation framework for serious games should encompass both validation and evaluation. The validation can be started as early as the ideation or conceptualization stage of serious game design by examining the quality of mapping between intended outcomes and six structural elements of the game (Tan, 2015). For off-the-shelf games, validation can be done to reuse or repurpose a game for new contexts or specific intent that is different from its original design goal. In practice, evaluation of serious games may be carried out when a game has been used either as a prototype in a laboratory or as a game-based intervention in the field testing (Barnum, 2011).

The key function of both the validation and the evaluation of serious games in the evaluation framework is to assure a correct alignment between the intended out-

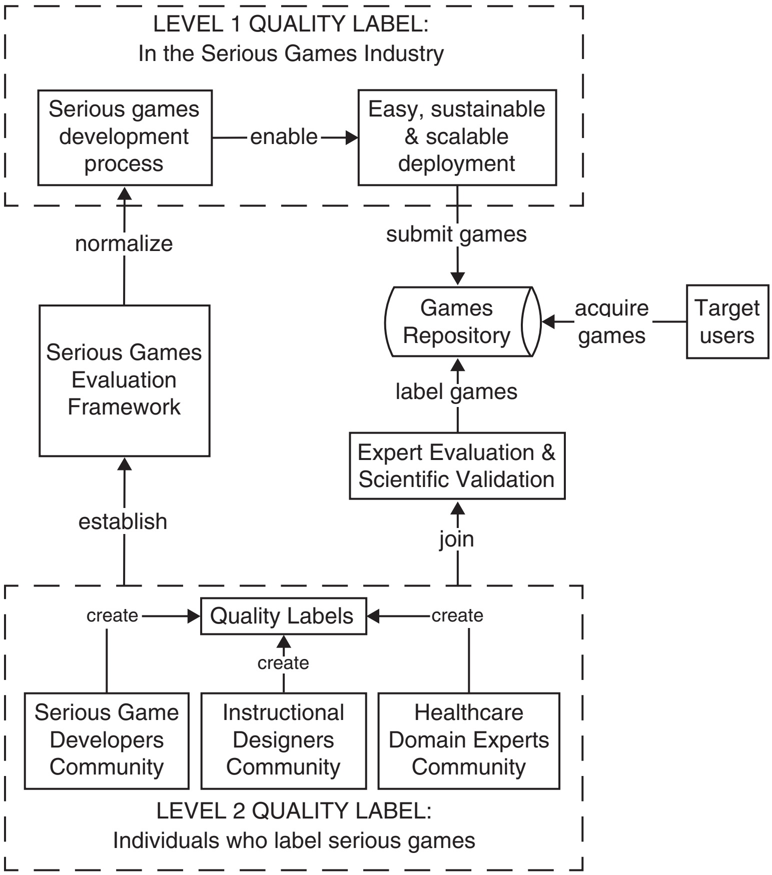


Fig. 9.1 How different communities work together to establish the serious games evaluation framework and label games

comes and the core mechanics of serious games (Biggs & Tang, 2007). This, in turn, will allow experts in the healthcare and wellness industry to provide constructive suggestions on how to repurpose, deploy, and implement evaluated serious games in healthcare contexts. The suggestions may also include serious games evaluation strategies for inter-player and intra-player performance when playing specific games. Since the framework is structured based on established guiding principles, these principles can be referred by serious games developers in the creative industry.

9.2.1 Adopting the Star Rating System of DSSH

The Serious Games Association has adopted the 5-star rating system (see Table 9.1) used by the Dutch Society for Simulation in Healthcare (DSSH) and has developed a transparent evaluation framework with an associated quality label for medical serious games (Doyen, Mert, Meijer, Nijboer, & Dankbaar, 2018). The quality label was based on their game evaluation criteria which contain 62 items in 5 categories of information: game description, rationale, functionality, validity, and data protection.

All game entries submitted for evaluation to obtain quality labelling will be evaluated by a committee that is formed by the Serious Games Association. Each committee could be different because specific expertise might be needed from different subject matter experts, instructional designers, and technologists. Once the committee achieves consensus about the evaluation, a jury report is created, and the applicants are informed of the rating, with narrative information and feedback about their game.

Even if the game is not awarded any stars, feedback will be provided. Applicants are encouraged to further improve their game, in order to achieve a higher level of validity, and obtain a better star rating. If the applicant wishes to apply for a higher rating, a new application must be submitted, and the process is repeated, ideally until the highest quality, e.g., a 5-star rating can be achieved. A provisional evaluation framework for serious games was developed and shown in Fig. 9.2.

9.2.2 Necessary Game Elements

In the DSSH star rating, the first criterion is to ensure that an entry that gets into the evaluation is actually a serious game, that is, “an interactive digital application, characterized by a storyline, a clear goal or objective, that is suitable for the target audience. The objective of the game needs to be relevant to accomplish the learning goal, either in a direct or indirect fashion. Interaction with the player is required, e.g. through direct player feedback or a scoring system and it needs to be an important element to achieve the goal of the game” (Doyen et al., 2018). In a word, four structural elements of a digital serious game are storyline, goal or objective, interaction,

Table 9.1 Criteria for labelling serious games quality, adopted from the DSSH

Star rating	Quality-labelling evaluation criteria (all requirements are cumulative and must be met)
1. Star	1. The entry is a serious game, which contains all necessary game elements 2. There is a safe data storage mechanism, i.e., compliant to the Personal Data Protection Act 2012 in Singapore or the Malaysian Personal Data Protection Act 2010 3. The working mechanism and theoretical background of the game must be at least plausible in relation to the rationale and functionality of healthcare and wellness
2. Star	4. Face validity has been checked and confirmed by experts 5. The underlying mechanism has to be supported by evidence
3. Star	6. Medical, educational/psychological, and game development experts have to be involved in the development process 7. Relevant player tests need to be performed, and the results must be processed into the game to reveal the conditions and degree of player's learning outcome attainment
4. Star	8. Independent experts have validated the game in a study. At least the construct validity (proving that a high score in the game correlates to high scores on proficiency tests in real life) needs to be proven
5. Star	9. Predictive validity has been confirmed in a rigorous scientific validation study, published in a peer-reviewed journal, hereby proving that the game achieves the set learning goals outside of the game. A 5-star serious game shows good potentials to be digital medicine can be prescribed by medical doctors for treatment (e.g., Glatter, 2016)

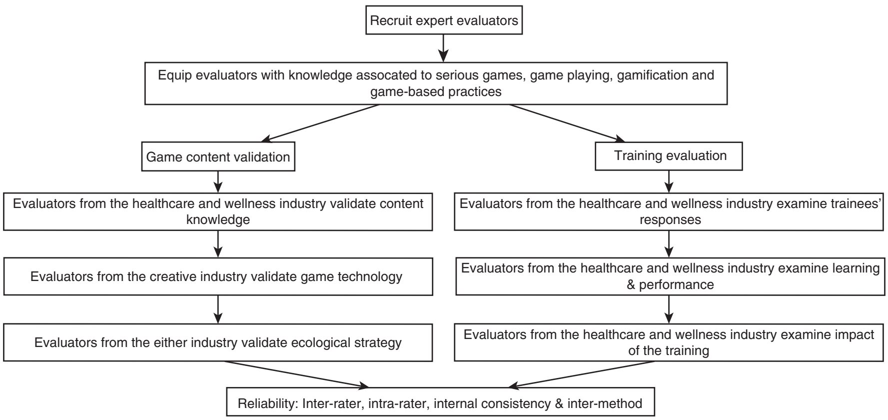


Fig. 9.2 A provisional multimodal serious games evaluation framework

and feedback, hence the notion of "necessary game elements." The provisional evaluation framework defines these elements as gameplay, feedback, and storyline.

9.2.2.1 Gameplay

Gameplay is the core of a game which consists of four structural components in the game, namely, game goal, rules, challenges, and interaction (Adams, 2013). In serious games, gameplay determines how content knowledge is aligned to the intended outcomes in the form of game goal and rules; how those intended outcomes were informed to targeted players in the form of a challenge; and how players should achieve those outcomes through interactions (Tan, Noor, & Wang, 2016). In an evaluation framework, experts in the creative industry should validate a serious game according to the novelty and fun-ability, while healthcare experts should assure appropriate alignment between the gameplay and intended outcomes.

9.2.2.2 Feedback

Feedback would determine the learnability of games (Prensky, 2007), i.e., how efficient players can learn the physics of a game world; what can and cannot be done in the game world; and how to win a mission, a quest, or the whole game (Adams, 2013). In this sense, diagnostic feedback informs players how well they have understood and mastered the gameplay at the beginning of a play session; formative feedback instructs players how to make progress when playing the game, while summative feedback debriefs players with their overall performance when the game is over. In an evaluation framework, game experts would classify the types of feedback used in the game. Then, they cross-check with healthcare-related strategies which can be deployed by medical professionals to afford long-term or sustainable engagement with targeted players.

9.2.2.3 Storyline

Game story consists of player events which allow players to act and react, narrative events which show players the contexts of the story, and in-game events which construct the game environment that players cannot change (Adams, 2013). Not all games involve storytelling; a narrative is an optional component for serious games (Jenkins, 2006). However, narrative plays an important role in certain game genres because it can connect programmed or preset events to player events through at least one storyline. In this sense, narrative supports players to immerse in the role they played or the story they walked through, hence narrative immersion (Adams, 2004). In the framework, the narrative should be evaluated by experts in the creative industry.

9.2.3 Types and Scopes of Knowledge Involved in Multimodal Evaluation

After accumulating the first star, all subsequent star ratings rely on experts who possess the necessary knowledge to conduct the evaluation. The credentials of a serious games evaluation framework are heavily depending on professional knowledge

Table 9.2 Eight types of knowledge required for serious games evaluation

Types of knowledge	Descriptors
1. Terminology	General knowledge of game playing, gamification, and game-based practices in healthcare and wellness
2. Specific facts	
3. Conventions	Professional knowledge of ways and means of analyzing game playing, gamification, and game-based practices in healthcare and wellness
4. Trends and sequences	
5. Classification and categories	
6. Criteria	
7. Methodology	
8. Principles and generalization	Professional knowledge of the universals and abstractions in synthesizing and assessing elements, components, structures, and outcomes of game playing, gamification, and game-based practices in healthcare and wellness

possessed by these experts. To qualify as an expert, one should be well-versed in eight types of knowledge on game playing, gamification, and game-based practices in healthcare and wellness. The categorization of knowledge types was adapted from a systems theory perspective (Hays, 2006), as shown in Table 9.2. Also, the serious games evaluation framework should cover three scopes of validation, i.e., content knowledge, game technology, and ecological strategy, as shown in Fig. 9.3.

9.2.3.1 Content Knowledge

The readability of the text content will be measured using existing formulae, including Flesch-Kincaid, Gunning Fog, and Coleman-Liau Index (Janan & Wray, 2014). As for non-text contents, experts from the healthcare industry can rate the content knowledge, using a set of instruments which is similar to questionnaires used by the International Age Rating Coalition (IARC) to classify digitally delivered games and apps (International Age Rating Coalition, 2016).

9.2.3.2 Game Technology

The usability of the serious game will be measured, which covers five aspects of the technology, i.e., likeability, efficiency, helpfulness, control, and learnability (Kirakowski & Corbett, 1993).

9.2.3.3 Ecological Strategy

The concept of the ecological strategy was adopted from ecology (MacArthur & Wilson, 1967). The ecological strategy claims that there are constantly other ecological factors that affect even an isolated landform in the form of an inaccessible

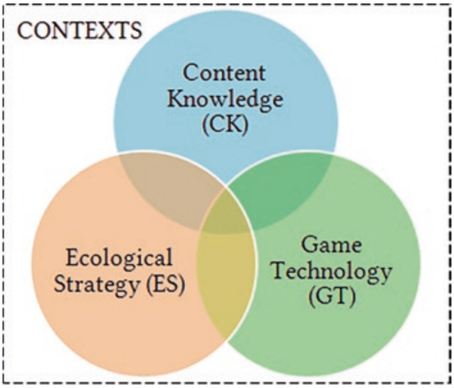


Fig. 9.3 Three scopes of an evaluation framework for serious games

island. Similarly for any game development project, the context of the game must also be taken into consideration. Broader aspects of the size of the users for such games, depth, and history of the subject which the game is built on should be taken into consideration. That would explain the acceptance of off-the-shelf games versus custom-made serious games.

9.2.4 Levels of Training Evaluation

According to Horton (2001), training can be evaluated at four levels of evaluation, i.e., the response evaluation, the learning evaluation, the performance evaluation, and the results evaluation. These levels can be applied in the multimodal training evaluation framework for serious games as described in the following subsections.

9.2.4.1 Response Evaluation

The first level of serious games evaluation is the response evaluation, in which targeted players were invited or recruited to give opinions based on their game playing experience. The evaluation can be conducted using one or a combination of several methods, including attitudinal measurement, follow-up questionnaires, interview, focus group, and access and navigation tracking. Questions asked in the level 1 evaluation should be related to the likeability and fun-engagement factor, such as the following (Iuppa & Borst, 2010):

– Did you like the game?
– What was your favorite game level?
– Was the game challenging?
– Was the game fun?

Response evaluation can be deployed to examine the Dunning-Kruger effect (Dunning et al., 2003) since it does not quantify players' advancement attributable to serious games (Iuppa & Borst, 2010). Instead, it can gather data to reveal perceived advancement after playing games.

9.2.4.2 Learning Evaluation

The second level of evaluation is the learning evaluation, where targeted players were tested or observed to identify learning gains. The evaluation can be conducted before, during, and after playing games using criterion-referenced tests, norm-referenced tests, and ipsative tests (Tan, 2013). Criterion-referenced tests can examine whether a player performed well or poorly on a given task; norm-referenced tests compare one player to his or her peers, while ipsative tests compare one player to his or her previous learning gains (Hughes, 2017).

Learning events may be embedded in the game environment or in the post-playing debriefing session (Wang & Yatim, 2019). Thus, players' positive and negative learning behaviors can be recorded as evidence of achieved learning outcomes, especially when they are engaging in hands-on activities, simulated work activities, and role-playing activities.

In the case where evaluators missed the timing to examine pre- and posttest differences or to observe in-game activities, learning evaluation can still be carried on by surveying superiors or immediate supervisors of individual players who have the authority to rate players' learning gains. However, it is worth stressing that the results of the survey might not reflect the actual learning gains.

9.2.4.3 Performance Evaluation

The third level of game evaluation is the performance evaluation which aims to examine the on-the-job performance of the real skills. The focus of performance evaluation will reveal the effectiveness of the game; thus it is related to the durability and sustainability of a specific game-based practice or solution. The evaluation can be led by an instructor as in role-playing assessments, or it can be conducted using simulation within or outside the game (Iuppa & Borst, 2010). Psychometric tests can be administered to measure the effectiveness and the reasoning behind performance improvement (Iuppa & Borst, 2010). Data collected from the measurement should be analyzed statistically to test hypotheses.

9.2.4.4 Results Evaluation

The fourth level of evaluation is the results evaluation which aims to examine the return on investment (ROI) (Horton, 2001). This evaluation is relevant to serious games that have been launched, installed, or sold. Depending on the initial design goal of specific serious game, variables related to ROI can be sales, revenue, profit margin, market share, stock price, or customer satisfaction ratings (Iuppa & Borst, 2010). For promotional or marketing games, four variables can be measured to determine the ROI, namely, brand establishment, brand recall, consumer outreach, and sales demand (Iuppa & Borst, 2010).

9.2.5 Reliability of the Evaluation

9.2.5.1 Inter-Rater Reliability

In the serious games evaluation framework, inter-rater reliability is important when a validation or evaluation task is assigned to more than one person. This type of reliability is assured by testing the extent of agreement between two or more evaluators in their assessment using the same instruments or methods under the same assessment conditions (Nutter, Gleason, & Christians, 1993).

9.2.5.2 Intra-Rater Reliability

In serious games evaluation, intra-rater reliability would be useful when a particular evaluator has been assigned several validation or evaluation tasks. The reliability concerns with the extent to which scores or ratings are consistent from one measurement to the next (Nutter et al., 1993). Measurements are collected from one rater who uses the same instruments or methods under the same conditions but at different timeframes.

9.2.5.3 Internal Consistency Reliability

When building test items for evaluation instruments, the consistency of results across items within a test is essential to ensure high internal consistency reliability. Cronbach's alpha (Cronbach, 1951) and Kuder-Richardson Formula 20 (Feldt, 1969) are two common and suitable measures for determining the level of internal consistency reliability.

9.2.5.4 Inter-Method Reliability

When different methods or instruments are used to evaluate serious games, inter-method reliability can become an issue. To achieve high reliability, the test scores or ratings must be consistent across different instruments or methods used in evaluation (Fries, Spitz, Kraines, & Holman, 1980). Standardizing the instruments or methods for serious games evaluation will make the inter-method reliability irrelevant.

9.3 Evaluating Serious Games for Health

9.3.1 Blood Transfusion Game

Every year, mass competency exercises are conducted to refresh nurses' competencies in certain key procedures and skills. Such exercises are very labor intensive as assessment is conducted in person and on a one-to-one basis. The assessment involves the assessor providing different clinical scenarios, in order to ascertain individual nurse's familiarity with practice guidelines, as well as their critical thinking and decision-making skills.

A team of nurses from the Singapore General Hospital (SGH) developed a blood transfusion game to demonstrate how a serious game could be used effectively and efficiently to evaluate nurses' competencies in a safe virtual environment (Fig. 9.4).

The game development team used the evaluation framework adopted from the DSSH to guide a team of nurses who were subject matter experts in the planning and development stages of the game. The team of subject matter experts (SME) started with the narratives of the game by developing different scenarios for the game. As the narrative is being developed, the SME defined identification attributes for specific users. The narratives started to be more specific when the SME established intended learning outcomes of the game. With a defined game narrative and a collection of learning outcomes, the game was divided into six stages. Each game stage encompasses a maximum of three learning outcomes. The duration of each game stage is capped at 10–15 min, allowing the game to be played in short sessions. With clear learning outcomes, the SME were clear on what sort of data they were expecting to see when the game was released to target users for trial.

The game developer with a clear understanding of the narrative and the learning outcomes started developing the gameplay. The users were identified, but their age group posed a problem for the developers as it stretches across a large age group. The gameplay must be captivating for a tech-savvy generation and those who were not, particularly senior staff members. Finally, the team decided to build a simulation game that was based on something the nursing staff are familiar with.

The final lap was spent on developing feedback in the game. The game should be able to provide the feedback needed by the SME to demonstrate the competency of each player. The team wanted to gather data on how often each player repeats the game in order to make progress from one game stage to the next. The team needed to know where were bottlenecks or 'difficult points' for the test group. The results were used to establish intervention which was necessary to improve work processes of blood transfusion in the hospital.

The Serious Games Association elected a team of three evaluators, comprising of a nurse, a game developer, and an e-learning entrepreneur, to become the first review panel for the blood transfusion game. The panel members were selected based on the abovementioned eight types of knowledge, which they need to be familiar with in their own domain of expertise.

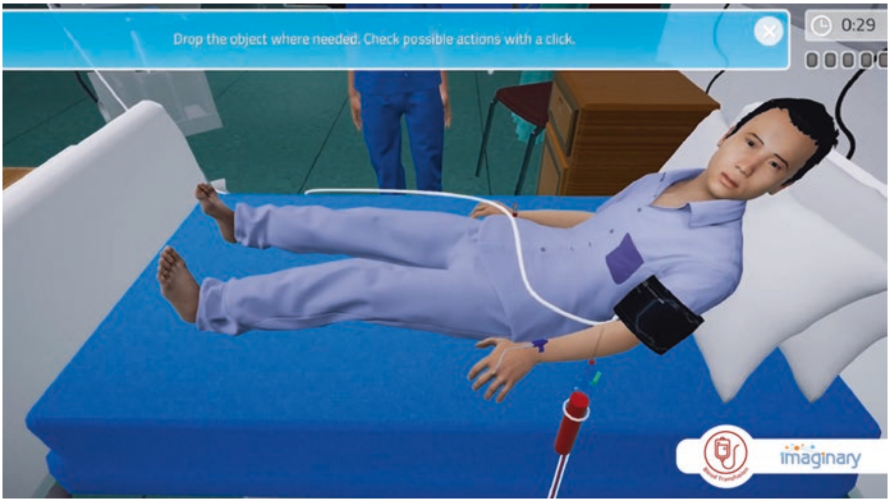


Fig. 9.4 A scenario in the blood transfusion game

1. Terminology used in the subject matter
2. Specific facts/procedures in the Singapore context
3. Conventions used in Singapore
4. Trends and sequences both locally and internationally
5. Classification and categories
6. Criteria of assessment
7. Methodology used in Singapore
8. Principles and generalization

The panel was provided with information addressing the five major areas.

9.3.1.1 Game Description

The team described the game clearly by articulating the purpose of the game, the intended user groups, and the settings where the game will be used. The team also explained the nature of game data that they are expecting to collect to show evidence of the nurses' competencies in conducting a blood transfusion for a patient in a specific scenario. The team was clear on the ownership of intellectual property associated to the game.

9.3.1.2 Rationale

Through different iteration of the narratives, the team was able to define specific learning outcomes for each game stage. With a common understanding of the learning outcome and the game outcome, the team agreed upon the rationale of each game stage. The rationale of each game stage directed a clear and precise outcome for the entire gameplay.

9.3.1.3 Functionality

The functionality of the game was established once both the nurses' team and the development team elaborated and agreed on the purpose of the game, how that could be achieved, and how the results could be measured.

9.3.1.4 Validity

The validity of the game was established when the game was released to over 2000 users to test. The game was released in two stages. In the first stage, the game was released to a single ward of 100 nurses. The criteria to expand the trial to the greater population were when the game playability was deemed stable after 2 weeks of trial. Feedback in terms of playability, engagement, content accuracy, and user-friendliness was then collated over a period of 2 months.

9.3.1.5 Data Protection

The team addressed issues on data protection in terms of the sensitivity of the data and how they will handle all issues regarding data safety and protection. This was done to ensure that data was collected correctly and stored safely.

The panel reviewed the game based on the content accuracy using the context of Singapore's guidelines and protocol for blood transfusion. This was limiting in certain ways, as the specific forms used in the process may not have the same criteria for different countries. There were already known differences in the process for blood transfusions between different institutions within Singapore. The game technology was easier to ascertain if they conform to industrial practices in game development. The game ecology was a bit of an unknown territory. As the serious games ecology in Singapore (or Asia) is not as established as in Europe or the USA, the game developers had to take a lot of the references and practices from the e-learning industry. The game was eventually given a 3-star rating.

9.3.2 Opportunities and Challenges

The panel evaluated the blood transfusion game based on the game description, rationale, functionality, validity, and data protection by using the 50 questions developed under the DSSH evaluation framework. The evidence provided by the blood transfusion game team was used to qualify the team in achieving the requirements under each star rating. In order to qualify for a 1-star rating, the following criteria need to be met:

1. The entry is a serious game, which contains all necessary game-elements.
2. There is a safe data storage mechanism.
3. The working mechanism and theoretical background of the game must be at least plausible.

Based on the storyline, gameplay, and feedback developed for the blood transfusion game, the evaluation panel was very clear that it was a serious game, which contains four structural game elements and other features of serious game, such as difficulty levelling, points system, and individual performance score.

The game data was stored in a secured cloud-based database. The game mechanics was derived from the day-to-day standard operation manual established for the nurses in real life. The game developer was very clear on how the working mechanism for the game could be developed, and the theoretical background was well understood.

To achieve the 2-star rating, the requirements were:

1. Face validity has been checked and confirmed by experts.
2. The underlying mechanism has to be supported by evidence.

As the game was being developed, the SME team constantly tests the gameplay with other senior nurses to ensure the procedure in the game follows the same procedure in real life. Besides getting feedback from other senior nurses in the hospital, the nurses also referred to the hospital's standard operating procedure manuals to confirm that the underlying game mechanism was in accordance to the latest established protocol since the Singapore Ministry of Health updates such protocols from time to time.

To achieve the 3-star rating, the requirements were:

1. Medical, educational/psychological, and game development experts have to be involved in the development process.
2. Relevant user tests need to be performed, and the results must be processed into the game.

The SME team comprised of three clinical nurses whom are qualified subject matter experts. The team is involved in the entire game development process, starting with the narration of the storyboard to the testing of the prototype game. Besides involving other clinical team members from time to time to test run different stages of the game for fact-checking purposes, there were also non-clinical individuals such as technicians, administrators, and medical education researchers providing valuable feedback on the approach and usability of the game. The feedback received from the individuals testing the game was constantly communicated back to the game developer. The main intention of constantly playtesting and adjusting was to tweak numerous minor changes rather than making a major change at the end when the game was almost completed.

Once the beta version of the game was ready for testing, the SME team selected a ward of 100 nurses to serve as the first round playtest users. The first playtest was planned to be completed over a period of 1 month. The decision to proceed to the second round of playtest was depended on the seriousness of reported bugs and changes needed. The first playtest was completed with no feedback to make any changes. The second playtest was run with 2000 playtesters over a period of 2 months. At the time of writing this report, the second round of playtest was still ongoing.

To achieve the 4-star rating, the requirement was:

1. Independent experts have validated the game in an empirical study. At least the construct validity (proving that a high score in the game correlates to high scores on proficiency tests in real life) needs to be proven.

In this game evaluation, it was rather easy for the evaluation panel to establish the rating because the team did not fulfill the requirement for a 4-star rating as they needed to conduct a game validation study. The recommendation from the evaluation panel was to put the game through a game validation study which will be conducted by an independent expert.

In this case study, it is apparent that the team managed to conduct the first level of serious games evaluation, which is the response evaluation. In the user trial test, 2000 targeted players were invited to give opinions based on their game playing experience. The team has yet to conduct the second level of evaluation, which is the learning evaluation, where targeted players were tested or observed to identify learning gains. Based on the evaluation and rating exercise done with the blood transfusion game, several opportunities and challenges were identified.

9.3.2.1 Compliance to Relevant Legislation

A key contributor to the credibility of a quality-labelling mechanism is the nature and extent of participation requirements. Other contributing factors are product-specific game technology, content accuracy, and pedagogy appropriateness. While the focus of the quality-labelling criteria relates to the quality and performance of the serious game being offered, it is also important to address the regulatory compliance. The process involves maintaining international compliance while accommodating regional differences. This approach acknowledges and avoids challenging the varying regulatory requirements that may exist in different country jurisdictions.

9.3.2.2 Consideration of Consistency of Serious Games Accreditation Centers

Besides legislative compliance, it is also important to address the consistency of evaluation by different serious games accreditation centers. As discussed earlier, there are already four variations of reliability, so when different raters are employed to evaluate the same serious game, inconsistent outcome can become an issue. To achieve high reliability, the test scores or ratings must be consistent across different centers based in different countries with varying cultures and practice. The credibility of both the quality label and the evaluation program could suffer if products bearing the quality label do not actually demonstrate comparable quality and reasonable performance although they are all adopting the same evaluation framework. Hence there is a need to establish more than one serious games accreditation center to triangulate game evaluation results.

9.3.2.3 Grounded on Sound Scientific Principles

Maintenance of stringent evaluation requirements based on good scientific approaches and methodology can assure consumers that they can trust the quality label, while all licensing applicants will be treated fairly. Further, there is a strong prevailing view that product criteria should be based on measurable indicators. The rationale is that there is a generally perceived need to assure consumers, producers, and developers that all aspects of a product's development, maintenance, deployment, and termination options have been taken into account.

9.3.2.4 Independence

A credible quality-labelling mechanism should be operated by an independent organization. The independence also extends to how product categories and criteria are determined. This is done through formal and direct representation of different stakeholders and interested groups on independent boards, panels, or advisory groups. The boards, panels, or advisory groups generally include members from the industry, consumers, academic and scientific community, and government sectors. The challenge is to strive for a balanced representation to prevent actual or perceived excessive influence by specific sector or individual stakeholders.

9.3.2.5 Open and Accountable Process

A credible program must be based on an open and accountable process that can be observed, monitored, and questioned at any time. Fair, consistent, and uniformly applied procedures must be established at each step in the process. A good-quality management system is a strong asset and highly desirable. Also, public criteria review is an essential feature of an open process. Some programs publicly announce new draft criteria through media and government information networks. Others hold public hearings or directly contact stakeholders and interested groups when requesting comments. The comments received through various means are then taken into account when preparing the final criteria.

9.3.2.6 Flexibility

In order to be credible and effective, the labelling mechanism must operate in a business-like and cost-effective manner, consistent with market forces and requirements. They must be able to respond to technological and market changes, in a timely manner. This requires, for example, periodic review and, when necessary, an update of criteria, taking into account technological and marketplace developments. Periodic review which is usually carried out every other year can ensure that standards and criteria levels keep pace with new developments. This would allow many identical programs to be upgraded at any time.

9.4 Discussion

In summary, an evaluation framework for serious games has been established to assure a correct alignment between the intended outcomes and the core mechanics of serious games. Gameplay, feedback, and storyline are three structural components, which can be aligned to three scopes of a game when being used in serious contexts, i.e., content knowledge, game technology, and ecological strategy. These scopes of serious games are validated in the evaluation. Meanwhile, game-based practices can be evaluated at four levels, to examine the responses, the learning gains, and the performance of targeted players. The results of game-based practices can be measured to determine the ROI. The results of validation and evaluation should be tested for inter-evaluator reliability, intra-evaluator reliability, and internal consistency reliability. After passing reliability tests, the outcomes of evaluation were expected to afford long-term engagement and sustain for achieving predetermined healthcare objectives.

Acknowledgements We thank Ms. Ang Shin Yuh, the Senior Nurse Manager of Nursing Research and Quality, and Ms. Andrea Choh and Ms. Teo Kai Yun who are nurse clinicians, Nursing Quality Management in SGH, who developed the blood transfusion game.

References

Abt, C. C. (1970). *Serious games*. Lanham, MD: University Press of America.

Adams, E. (2004). *The designer's notebook: Postmodernism and the 3 types of immersion*. Retrieved December 25, 2018, from http://www.gamasutra.com/view/feature/130531/the_designers_notebook_.php

Adams, E. (2013). *Fundamentals of game design* (3rd ed.). San Francisco, CA: New Riders.

Aebersold, M. (2018). Simulation-based learning: No longer a novelty in undergraduate education. *The Online Journal of Issues in Nursing, 23*, 2. https://doi.org/10.3912/OJIN.Vol23No02PPT39

Barnum, C. M. (2011). *Usability testing essentials*. Burlington, MA: Morgan Kaufmann.

Biggs, J., & Tang, C. (2007). *Teaching for quality learning at university: What the student does* (3rd ed.). Maidenhead, UK: Society for Research into Higher Education and Open University Press.

Brady, M., Devitt, A., & Jameson, E. (2016). Towards a framework for the evaluation of digital game-based learning at the third level. In *Proceedings of the 6th Irish Conference of Game-based Learning* (pp. 110–115). Trinity College Dublin.

Cant, R. P., & Cooper, S. J. (2017). The value of simulation-based learning in pre-licensure nurse education: A state-of-theart review and meta-analysis. *Nurse Education in Practice, 27*, 45–62.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Csikzentmihalyi, M. (2008). *Flow: The psychology of optimal experience*. New York, NY: Harper Perennial.

De Freitas, S. (2006). *Learning in immersive worlds. A review of game-based learning*. Bristol, UK: JISC.

Doyen, B., Mert, A., Meijer, H. A. W., Nijboer, T. C. W., & Dankbaar, M. E. W. (2018). A quality label for serious games. *Medical Training Magazine, 1*, 10–13.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*(3), 33–37. https://doi.org/10.1111/1467-8721.01235

Dutch Game Garden. (2016). *Growing Games. The evaluation of health-oriented serious games and apps: A differentiated approach*. Retrieved from https://www.dutchgamegarden.nl/wp-content/uploads/2018/08/DifferentiatedApproachHealthGameValidation_v11-2.pdf

Falstein, N. (2007). Serious games production. In *The eLearning Guild's Online Forums: Strategies and Techniques for Building Immersive Learning Simulations (ILS)*.

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder Richardson coefficient twenty is the same for two tests. *Psychometrika, 34*(3), 363–373.

Fries, J. F., Spitz, P. R. N., Kraines, R. G., & Holman, H. R. (1980). Measurement of patient outcome in arthritis. *Arthritis and Rheumatism, 23*, 137–145. https://doi.org/10.1002/art.1780230202

Glatter, R. (2016). Barton Health first to offer new digital medicine developed by Proteus Digital Health. *Forbes*. Retrieved May 21, 2019, from https://www.forbes.com/sites/robert-glatter/2016/01/13/barton-health-first-to-offer-new-digital-medicine-developed-by-proteus-digital-health/

Gloria, A. D., Bellotti, F., Berta, R., & Lavagnino, E. (2014). Serious games for education and training. *International Journal of Serious Games, 1*(1). https://doi.org/10.17083/ijsg.v1i1.11

Hays, R. T. (2006). *The science of learning: A systems theory perspective*. Boca Raton, FL: BrownWalker Press.

Horton, H. (2001). *Evaluation e-learning*. Alexandria, Egypt: American Society for Training and Development (ASTD).

Hughes, G. (2017). *Ipsative assessment and learning gain: International case studies*. Basingstoke, UK: Palgrave Macmillan.

International Age Rating Coalition. (2016). *How IARC works?* Retrieved December 28, 2016, from https://www.globalratings.com/how-iarc-works.aspx

Iuppa, N., & Borst, T. (2010). *End-to-end game development: Creating independent serious games and simulations from start to finish*. Burlington, MA: Focal Press.

Janan, D., & Wray, D. (2014). Reassessing the accuracy and use of readability formulae. *Malaysian Journal of Learning and Instruction, 11*, 127–145.

Jenkins, H. (2006). Game design as narrative architecture. In K. Salen & E. Zimmerman (Eds.), *The game design reader: A rules of play anthology* (pp. 670–689). Cambridge, MA: MIT Press.

Jones, F., Passos-Neto, C. E., & Braghiroli, O. F. M. (2015). Simulation in medical education: Brief history and methodology. *Principles and Practice of Clinical Research, 1*(2), 56–63.

Kirakowski, J., & Corbett, M. (1993). SUMI: The software usability measurement inventory. *British Journal of Educational Technology, 24*(3), 210–212. https://doi.org/10.1111/j.1467-8535.1993.tb00076.x

Lin, Y., Cheng, A., Hecker, K., Grant, V., & Currie, G. R. (2018). Implementing economic evaluation in simulation-based medical education: Challenges and opportunities. *Medical Education, 52*(2), 150–160.

MacArthur, R. H., & Wilson, E. O. (1967). *The theory of island biogeography*. Princeton, NJ: Princeton University Press.

McDougall, E. M. (2015). Simulation in education for health care professionals. *British Columbia Medical Journal, 57*(10), 444–448.

Nutter, F. W., Gleason, M. L., & Christians, N. C. (1993). Assessing the accuracy, intra-rater repeatability, and inter-rater reliability of disease assessment systems. *Phytopathology, 83*, 806–812.

Popescu, M. M., Romero, M., & Usart, M. (2013). Serious games for serious learning-using SG for business, management and defence education. *International Journal of Computer Science Research and Applications, 3*(1), 5–15.

Prensky, M. (2007). *Digital game-based learning*. New York, NY: Paragon House.

Sawyer, B., & Smith, P. (2008). *Serious games taxonomy*. Retrieved December 25, 2018, from https://thedigitalentertainmentalliance.files.wordpress.com/2011/08/serious-games-taxonomy.pdf

Serious Games Society. (2008). *Serious games initiative: About serious games initiative*. Retrieved December 25, 2018, from http://www.seriousgames.org/about.html

Sonawane, K. (2017). *Serious games market – Global opportunity analysis and industry forecast, 2016–2023*. Singapore, Singapore: Allied Market Research.

Tan, W. H. (2013). Developing aesthetics assessment skills in Bachelor of Design Programmes: An introspective account in UPSI. *International Journal of Assessment and Evaluation in Education, 3*, 92–104. Retrieved from http://ojs.upsi.edu.my/index.php/IJAEE/article/view/1949

Tan, W. H. (2015). *Gamifikasi dalam pendidikan: Pembelajaran berasaskan permainan*. Tanjung Malim, Malaysia: Penerbit UPSI.

Tan, W. H., Noor, N. F. M., & Wang, Y. (2016). Inventive problem-solving puzzle game design and development. *International Journal of Computational Vision and Robotics, 6*(3), 244–252. https://doi.org/10.1504/IJCVR.2016.077355

U.S. Army. (2017). *America's Army: America's Army Providing Grounds: The official game of the US Army*. Retrieved January 25, 2017, from https://www.americasarmy.com/

Wang, Y., & Yatim, M. H. M. (2019). A framework of childhood obesity prevention. In W. H. Tan (Ed.), *Design, motivation, and frameworks of game-based learning* (pp. 154–183). Hershey, PA: IGI Global.

Wattanasoontorn, V., Boada, I., Garcia, R., & Sbert, M. (2013). Serious games for health. *Entertainment Computing, 4*, 231–247. https://doi.org/10.1016/j.entcom.2013.09.002

Chapter 10
Scaffolding and Assessing Teachers' Examination of Games for Teaching and Learning

Mamta Shah

10.1 Introduction

The praxis of facilitating and assessing student learning with interactive, immersive, and interdisciplinary technologies is impacting the roles teachers play in contemporary classrooms (Shaffer, Nash, & Ruis, 2015). This necessitates a reconceptualization of the knowledge and skills teachers cultivate in order to creatively establish a pedagogical partnership with learning technologies and to make relevant decisions about instruction and assessment (Mehta, Henriksen, & Rosenberg, 2019). Digital games are one type of complex learning technologies that have not only pervaded K-12 classrooms but also catalyzed dynamic shifts in teachers' roles in scaffolding and assessing student learning (Groff, 2018; Kangas, Koskinen, & Krokfors, 2016). However, the field is nascent in terms of the role, presence, and identity of teachers in augmenting the impact of games on student learning processes and outcomes (Chee, Mehrotra, & Ong, 2015; Sanchez-Mena, Marti-Parreno, Sanchez-Mena, & Aldas-Manzano, 2017; Shah & Foster, 2018). Specifically, there is a dearth of research on theoretical and methodological frameworks that can operationalize, guide, and examine the development of teachers' knowledge and motivation to adopt technological pedagogical innovations such as games. Related to this issue is the lack of empirically tested scalable models for teacher professional development in the context of game-based learning (Caldwell, Osterweil, Urbano, Tan, & Eberhardt, 2017; Foster, Shah, & Duvall, 2015; Molin, 2017). This chapter argues for the centrality of teachers in augmenting teaching, learning, and assessing with games in K-12. In doing so, the chapter presents the Game Network Analysis (GaNA) methodological framework (Foster, 2012) and illustrates its application in a study with pre-service teachers in a university context. GaNA offers one way of scaffolding and assessing teachers' examination of games for learning as an

M. Shah (✉)
Drexel University, Philadelphia, PA, USA
e-mail: mamta@drexel.edu

D. Ifenthaler, Y. J. Kim (eds.), *Game-Based Assessment Revisited*, Advances in Game-Based Learning, https://doi.org/10.1007/978-3-030-15569-8_10

important first step toward the calls for advancements in the field as outlined in the current volume.

10.2 Games and Teachers in K-12

Shaffer, Squire, Halverson, and Gee (2005) argued that digital games are conducive environments for learning because they can be used for promoting "situated understandings, effective social practices, powerful identities, shared values, and ways of thinking of important communities of practice" (Shaffer et al., 2005, p. 7). Over the last decade, educational researchers have continued theorizing and examining gaming and game-making activities for promoting affective-cognitive-motivational-social dimensions of learners' development (Foster, 2011; Gaydos & Devane, 2019; Siyahhan, Ingram-Goble, Barab, & Solomou, 2017). For instance, Kafai and Burke (2015) contended that learning through making games, that is, constructionist gaming, can (a) support students to develop a variety of digital and technical literacy skills, (b) provide opportunities for students to explore aspects of their identity and collaborate with peers, and (c) help gaming cultures become more inclusive. Research has also burgeoned on investigating the impact of gamification and game-based learning (GBL) as instructional approaches for promoting students' knowledge and motivation in a variety of academic domains (Boyle et al., 2016; Clark, Tanner-Smith, & Killingsworth, 2016; de Freitas, 2018; Wouters & van Oostendorp, 2013). For instance, Raphael (2018) demonstrated the efficacy of employing gamification strategies within a blended learning environment (a) to engage students in simulated professional experiences, (b) to scaffold students' proficiency with targeted skills, and (c) to support students in developing conceptual knowledge, while supporting students' motivational orientations in interactive environments. The focus of this chapter is on a nascent area of research in the field, that is, educating teachers in using games to augment students' learning processes and outcomes, such as those outlined above (Franklin & Annetta, 2011; Molin, 2017).

Recent studies have illustrated that teachers' intervention and facilitation in a game-based classroom can lead to meaningful learning experiences for students (Barzilai & Blau, 2014; Spires & Lester, 2016). For instance, a case study by Watson, Mong, and Harris (2011) delineated teachers' adeptness in identifying and leveraging teachable moments to support students' understanding of concepts in a history unit. However, researchers agree that the nature of teaching with games is complex and dynamic, evolving from the time a teacher identifies a game to the time he/she incorporates the game within a curriculum and finally introduces the game in the classroom (Eastwood & Sadler, 2013; Hanghøj & Hautopp, 2016; Silseth, 2012). As teachers engage in implementing games in their curriculum in a recursive manner, they develop adaptive expertise to engage in differentiated instructional approaches that complements the affordances of the games and mirrors students' personal engagement with the curriculum (Bell & Gresalfi, 2017; Marklund & Taylor, 2015; Shah & Foster, 2014a). However, such skillful use of games is not intuitive to all teachers. In addition, not all teachers have supportive ecosystems to develop competence in using games (Stieler-Hunt & Jones, 2017). Large-scale surveys have brought to the spotlight that pre-service and in-service teachers' goal to incorporate GBL in their practice is impacted by a lack of professional development (PD) opportunities that can systematically guide them in using games for teaching, learning, and assessment (Fishman, Riconscente, Snider, Tsai, & Plass, 2014; Ruggiero, 2013). Teachers using games "wish it were easier to find games that aligned with their curriculum" and "[are] not sure how to integrate games into their teaching" (Takeuchi & Vaala, 2014). These trends are problematic and must be addressed in order for game-based learning to be accessible to more teachers in K-12 to enhance instructional and assessment approaches. Scholars have recommended that explicating the practices involved before, during, and after game-based interventions through systematic models that can offer teachers the skills to leverage the affordances of games might be a beneficial approach (Kangas et al., 2016; Molin, 2017).

10.3 Teacher Education in Game-Based Learning

Until almost a decade ago, teachers were an underrepresented group in game studies (Hwang & Wu, 2012), and educational research focusing on teacher education in GBL was scarce (Becker, 2007; Franklin & Annetta, 2011). The earliest shifts in the field began with surveying pre-service and in-service teachers for (a) their perceptions, attitudes, and beliefs about the instructional use of games and (b) the perceived opportunities and barriers in introducing games in classrooms (Kenny & Gunter, 2011; Kenny & McDaniel, 2011), specific cultural contexts (Allsop, Yeniman Yildirim, & Screpanti, 2013), and academic domains (Demirbilek & Tamer, 2010). Parallelly, case studies of in-service teachers were emerging to situate teachers' roles and intervention as crucial in (a) creatively repurposing games for curricular use (Eastwood & Sadler, 2013; Hanghøj & Brund, 2011), (b) overcoming the design- and context-imposed limitations of games (Jaipal & Figg, 2009; Ketelhut & Schifter, 2011), and (c) making student learning in games explicit and meaningful (Barzilai & Blau, 2014; Silseth, 2012). Whereas different models of game-based classroom interventions were being tested and documented (e.g., variety of game genres, nature of learning goals), there was a growing recognition that one of the key deterrents to the widespread and sustainable adoption of GBL was the lack of teacher professional development opportunities (Groff, McCall, Darvasi, & Gilbert, 2016; Ruggiero, 2013; Takeuchi & Vaala, 2014).

Scholars have found that professional development positively impacts teachers' perceptions and intention for using games in their classroom. One such conclusion was reported by An (2018) who facilitated 25 students (mostly K-12 teachers) enrolled in a semester-long online professional course to develop the skills for (a) interacting with other educators and parents about the learning potentials of games, (b) selecting and analyzing games that might be relevant, (c) incorporating games in their instruction, and (d) facilitating game design activities with their students. In another study, Hsu, Liang, and Su (2015) assigned 25 in-service teachers to a pedagogy-oriented course and 24 to a technology-oriented course. The researchers then assessed the impact of the instructional sequences afforded by these versions on teachers' technological pedagogical content knowledge for games (TPACK-G) or their knowledge to use games that integrated methods and content and acceptance of game-based learning. This quasi-experimental study found that teachers in the technology-oriented group, that is, teachers who were facilitated to acquire knowledge of a game first (objective of the game, functions of the game), were better equipped to align the game with decisions about content and pedagogy in designing their lesson plans (Hsu et al., 2015).

However, few researchers have proposed theoretical and practical structures to aid teachers in adopting GBL; their potential for offering a comprehensive package for scaffolding and assessing teachers' knowledge of GBL over time remains to be witnessed. For instance, Wu's (2015) proposed typology may be useful as many teachers experience difficulties in identifying games that align with their curricular needs. However, while the typology can aid teachers in identifying games and, by extension, determine a game's content and pedagogical characteristics, teachers may not necessarily know how to leverage the potentials of the games and design experiences that alleviate the constraints of the game to meet a set of learning goals. Similarly, Jong and Shang (2015) offered Virtual Interactive Student-Oriented Learning Environment (VISOLE), a structure for teachers to support student learning with games. However, the application of VISOLE may be restricted because of its reliance on online games exclusively with a capability of back-end activity access for teachers. Not all teachers may have access to games with such exclusivity and have the supportive infrastructure (e.g., technology, money) in their schools to be able to use online games only. Furthermore, the frameworks or the typology does not prompt teachers to consider the nuances of facilitating GBL to engage learning in academic domains (Foster, 2008).

The lack of guiding game-based learning frameworks has led to large-scale projects with sufficient resources struggling to support participating and interested teachers in a meaningful manner (Caldwell et al., 2017). This issue is compounded further by a paucity of studies that focus on developing and examining teachers' knowledge of GBL, particularly during teachers' pre-service years (Barbour et al., 2009; Kennedy-Clark et al., 2015; Sardone & Devlin-Scherer, 2010). The Game Network Analysis (GaNA) framework offers one approach (a) to make GBL accessible to pre-service and in-service teachers, (b) for teacher educators to design professional development interventions, and (c) for researchers to capture change in teacher's knowledge and motivation to incorporate GBL in their practice through formative and summative assessments focusing on three focal areas: game analysis, game integration, and ecological conditions.

10.4 Theoretical Framework

Game Network Analysis (GaNA) was developed as a methodological process for designing and facilitating game-based learning (GBL) experiences (Foster, 2012), thus, making it a good fit for this volume. GaNA includes a network of pedagogical and analytical frameworks that allows users (teachers, in this study) to focus on the pedagogy and content of games as well as the process for employing GBL in a given ecology, in formal and informal learning settings (Shah & Foster, 2015). GaNA is comprised of an analytical lens for game analysis and selection by helping teachers approach a game as a curriculum with context-attuned constraints and affordances for technology, pedagogy, and content (Foster, 2012). GaNA includes *Play Curricular activity Reflection Discussion* (PCaRD) model that aids teachers to systematically leverage the affordances of a game and overcome the limitations of a game for specific learning goals by designing congruent and anchored instruction, reflection, and discussion activities that extend student engagement in a curricular unit beyond game play. Teachers adopt a variety of roles to maintain a synergy between emergent teachable moments mediated during PCaRD, challenges inherent in a typical school structure, and students' knowledge and motivation to learn in a specific academic domain (Foster & Shah, 2015b). GaNA facilitates teachers in identifying opportunities for inquiry, communication, construction, and expression (ICCE) to foster transformative learning experiences anchored in the game and design opportunities for ICCE during curricular, reflection, and discussion activities (Foster & Shah, 2015a). The decisions teachers make during game analysis and game integration are guided by ecological conditions impacting the successful use of GBL experiences. These conditions include social dynamics, organizational and technological infrastructure, and pedagogical culture of the context in which GBL is to be introduced and sustained (see Fig. 10.1). As such, GaNA assumes teachers' knowledge and skills for using games should comprise of three focal and interconnected areas: game analysis, game integration, and ecological conditions. These knowledge areas are essential for teachers to leverage the affordances of games in designing curricular interventions, facilitating and assessing student learning.

10.4.1 Rationale and Process of Game Analysis

The purpose of analyzing a game is to gain deep familiarity with the affordances and constraints of the environment such that this knowledge informs teachers' decisions to implement the game for supporting student learning in a chosen context. The first step in this pursuit involves helping teachers go past the perception that games, like any technology, are unbiased and stand-alone artifacts (Mishra & Koehler, 2006). Instead, teachers need to approach games as designed experiences (Squire & Barab, 2012) whose cognitive, pedagogical, and experiential affordances

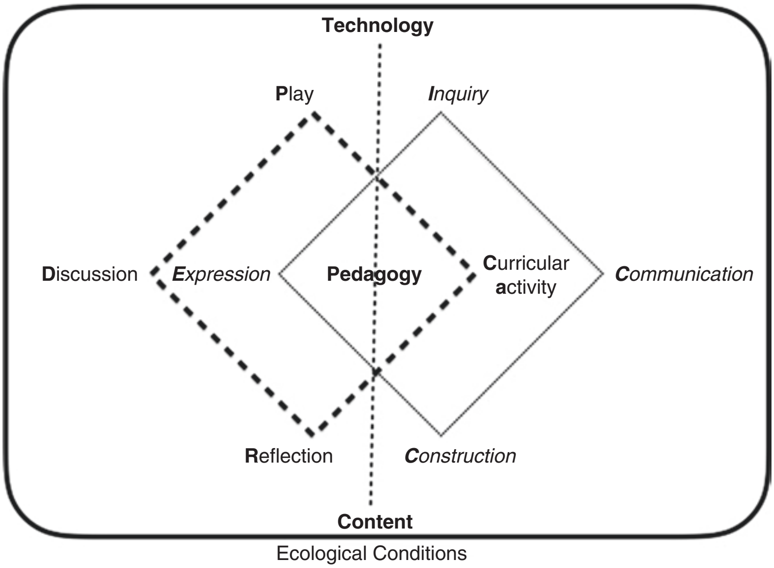


Fig. 10.1 The Game Network Analysis

and constraints could be leveraged in partnership with teachers' expertise (Moline, 2009).

As a result, the initial process of game analysis encompasses practicing direct and vicarious methods that can yield relevant knowledge about the game. These methods include teachers playing the game, researching it (e.g., looking for information about the game on the publisher's website), and observing another individual play the game (e.g., watching a YouTube video) (Aarseth, 2003; Foster, 2012). Doing so allows teachers to establish a level of comfort with the game and the process of analyzing the game. This activity also yields preliminary insights about the game in relation to technical requirements (e.g., platform for running the game and ease of installation), pedagogy in general (e.g., objective of the game, intended target group, customization options, icons and multimodal literacy needed), embedded content, and pedagogy specific to the content. This initial game knowledge may be sufficient for teachers to decide whether the game warrants further examination such that it could possibly be integrated within a curriculum and the larger context (Zhao, Pugh, Sheldon, & Byers, 2002).

As teachers begin to delve deeper into a game, it necessitates documentation of what infrastructure/resources will be needed to use the game and identification of what students will learn that is relevant to the curricular goals. Additionally, teachers need to examine the nature of experiences the game is likely to engage students in, focusing on whether the game can adapt to different learning orientations (Foster, 2011), how it scaffolds the knowledge students construct, and the kinds of opportunities it presents for students to personally connect with the learning experience within the game and the academic domain (Foster, 2008). Thus, teachers' firsthand assessment and detailed documentation provides them with insights about the educational merits and limitations of a game. This activity also allows teachers to foresee the kind of curricular activities that will be required outside of the game to help students make connections between their play experience and the desired learning objectives and facilitate students in articulating their newly formed knowledge. *In summary, the process of game analysis assists teachers in game exploration, selection, and evaluation.*

10.4.2 Rationale and Process of Game Integration

While the process of game analysis allows teachers to identify the characteristic features of a game, the objective of game integration is for teachers to leverage the potential of a game and augment its impact on student learning through teachers' expert intervention – a process akin to distributed cognition (Plass, Homer, & Kinzer, 2015). The key determinants to successful game integration is for teachers (a) to learn how to use a game as their pedagogical partner that complements and extends teachers' technical pedagogical and content knowledge and (b) to use the game as an anchor for facilitating a social, affective, motivational, and cognitive learning experience for students (Dewey, 1956; Shah & Foster, 2014b). The first step toward this is to understand that the process of game integration is iterative and dynamic. It should allow for scaffolding student experiences and informing immediate and future developments in their learning trajectory through the game. Typically, this involves immersion in naturalistic game playing and engagement in curricular activities that build upon students' game playing experience, followed by reflections and discussions to articulate the connections made between the game and learning objectives (Silseth, 2012). Such a routine allows teachers and students to go past the novelty of learning with games and establish a structure to focus on the learning objectives, while continuing to learn through play-based activities.

Teachers who effectively facilitate a game-based classroom perform multiple roles (Kangas et al., 2016). Teachers create and nurture a naturalistic learning environment where sharing skills and knowledge, coaching one another, and problem solving with peers are encouraged. They act as participant observers in order to gain additional insights about the game through students' play experiences and to facilitate and assess student learning and engagement. Through curricular activities, teachers choose between providing instruction and designing opportunities for students to apply and demonstrate their emerging understanding of the curriculum. Reflections and discussions further allow teachers to prompt students to articulate and discuss the connections and gaps in their learning related to the desired objectives. Teachers also use this time for providing feedback and summary. *Thus, through game integration, teachers gradually make the process of learning in games intentional, explicit, and meaningful for students. Practicing such a routine also enables teachers to identify teachable moments and improvise during a GBL session and in the subsequent sessions.*

10.4.3 Rationale and Process of Ecological Conditions Affecting Game Use

Although game analysis precedes game integration as a procedure, conceptually they occur simultaneously. Teachers must think about game integration as they analyze a game conceptually and pragmatically. Similarly, the process of game integration deepens teachers' game knowledge (Shah & Foster, 2014b). Another layer of expertise that teachers must add in the process of introducing and facilitating GBL is to consider the context as they make decisions in relation to game analysis and integration. The purpose of being aware and skilled in working around the conditions within the learning context to the best possible extent is to ensure that organizational infrastructure, social dynamics, and established pedagogical practices are in harmony for teachers to nurture students learning through games even when unexpected changes may be experienced (Nardi & O'Day, 1999; Zhao et al., 2002).

10.5 Method

The study reported in this chapter is situated in a larger exploratory convergent mixed-method (Creswell & Clark, 2007) doctoral research project that was undertaken with the objective of cultivating pre-service teachers' knowledge and skills for integrating digital games in K-12 classrooms. In this study, an 11-week methods course designed using the aforementioned principles of GaNA scaffolded and assessed teachers' ability to analyze games, integrate games in lesson plans, and consider specific ecological conditions that impact the introduction and sustainability of GBL in learning settings. Data (self-report and applied) was obtained by way of multiple sources to ascertain the extent to which participants cultivated the aforementioned methodological knowledge and skills of GBL. Thus, the convergent design was beneficial in ascertaining what knowledge participants claimed to develop and were able to exhibit about game analysis and integration and the conditions impacting GBL in education, when they were tested, surveyed, interviewed, and observed in the study. Further, participants' knowledge about GBL was demonstrated in course assignments, classroom interactions, and discussions. Data was collected simultaneously and triangulated to answer the research question: "To what extent did an 11-week methods course scaffold and assess pre-service teachers' knowledge and skills for GBL?"

10.5.1 Participants

The researcher (a doctoral candidate in learning technologies at the time of the study) designed and taught "Integration of Digital Games in K-12 Classrooms," as a three-credit special topic elective course at a private university in a city in the mid-Atlantic area of the USA in Spring 2013 (April–June). Participants were selected using convenient sampling from a pool of approximately 200 on-campus undergraduate and graduate pre-service teachers. Given the exploratory nature of the study and the methodological focus on the course, enrollment was open to pre-service teachers from all concentrations and college years. Participants were not required to have any minimum game-playing experience in order to enroll for the course and content to participate in the study. Thus, only those pre-service teachers who were interested in the GBL elective course were registered for it.

A total of 18 pre-service teachers enrolled in the course and consented to participate in the study; however, 14 participants completed the course. Participants included five undergraduate students (from freshman and senior years) and nine graduate students (ranging from program beginners to pre-service teachers ready to graduate). The participants varied in terms of their concentration areas, subject areas, and focus age groups in their teacher education programs. These included teaching learning and curriculum, science of instruction, elementary education (PreK-4), secondary education (chemistry, physics, and language arts), and special education. The average age of the participants was 24 years, ranging from ages 18 to 45 years with four males and ten females. At the start of the study, participants were not regular game players. Majority of them reported playing casual mobile games such as Angry Birds and Words with Friends for less than an hour a week in order to pass time when bored. At the time of study, none of the participants had received prior training in GBL. Only four participants had completed a school practicum experience as part of their teacher education program.

10.5.2 Settings

The intervention involved two educational settings: *physical and virtual*. The class met on campus for a 3-hour class each week. The classroom had furniture that was fixed in a lecture-style classroom arrangement (five rows of table chairs facing the instructor). While the classroom arrangement was not the preferred kind (e.g., flexible, seminar style), the small size of the class allowed the researcher to create an environment for participants to have optimal interaction with each other and the researcher and to create a collaborative space for the participants to share their experiences and expertise with their peers.

The following online games were used for the study: Citizen Science, Hot Shot for Business, and Spent. The following PC-based games were used: Food Force and RollerCoaster Tycoon 3 (RCT3). These games were chosen for several reasons. First, all the games except RCT3 are available online at no cost. RCT3 is a popular game that can be purchased at a cost of only $5. Second, these games focus on a diverse range of content areas and skills and target age groups that could use the game. A Blackboard Learn® course shell was also used by the researcher to share course resources and submit graded assignments. Participants also used Blackboard

Learn® to perform classroom (e.g., document findings from game analysis, reflection in PCaRD) and weekly (e.g., discussion on readings) activities.

10.5.3 Procedure for Scaffolding and Assessing Teachers' Examination of Games for Teaching and Learning

"Integration of games in K-12" was designed to be a participatory course. As such, individual growth and contribution were valued as much as learning with peers. Through classroom activities and assignments, the course objectives were for participants (a) to become knowledgeable about identifying the affordances and constraints of digital games along three dimensions (technology, pedagogy, and content) and (b) to become skilled in the process of incorporating games in the design of classroom activities to achieve curricular goals. The guiding essential questions of the course were as follows:

1. Why use games for learning in K-12 education?
2. What are the potentials and barriers to implementing GBL in K-12 education and how can they be addressed?
3. What skills and knowledge do teachers need in order to facilitate synergistic practices between student engagement, interdisciplinary learning objectives, and pedagogy through digital games?

Additional questions/prompts were provided to scaffold participating pre-service teachers' development in the course that tapped into teachers evolving knowledge and motivation to adopt GBL. For instance, (a) what associations are you making with your insights from this course and your teaching philosophy? (b) What skills, knowledge, and experiences of yours will help you implement GBL as a teacher?

Overall, the structure of this 11-week course was designed to scaffold participants' understanding of the methods involved in GBL though a judicious mix of theory and practice. Participants were afforded with opportunities to experience, examine, and implement GBL through course activities and assignments. Participants engaged in authentic experiences (e.g., analyzing games), a collaborative learning environment (e.g., analyzing games as a group), and opportunities for reflection and articulation (e.g., presenting findings from game analysis and exploring the curricular use of a game that was analyzed). Tables 10.1 and 10.2 explain the schedule for scaffolding and assessing change in teachers' knowledge of GBL.

A 2-h and 50-min face-to-face class was held each week for the entire duration of the course. Typically, for each week, the first 80 min of the class was dedicated to activities and topics related to game analysis (part A). This was followed by a break of 10 min. Thereafter, the remaining 80 min focused on topics and activities related to game integration (part B). Weekly online discussion boards also were set up for participants to discuss readings based on weekly themes. Together, the intervention focused on instruction, knowledge construction, and skill development

Table 10.1 Weekly schedule for scaffolding teachers' knowledge of game-based learning

Activity	Duration in minutes
Discussing assigned readings based on weekly themes[a]	15–20
Analyzing games individually or in groups using GaNA guide	40
Discussing findings from the game analysis activity	15–20
Experiencing the PCaRD model through different vantage points (student, teacher)	50–55
Unpacking PCaRD and discussing issues faced by teachers during game integration[b]	15–20

[a]Example of themes: (a) teachers' technology knowledge and technology integration (Zhao, 2002; Zhao et al., 2002), (b) the relationship between digital media and learning (Gee, 2005; Squire & Jenkins, 2003)

[b]Examples of techniques used: participants analyzing exemplary cases of in-service teachers presented in empirical studies to understand the nature and relevance of teacher intervention in game-based classrooms (e.g., Silseth, 2012); virtual synchronous interaction with an in-service teacher who had previously incorporated GBL using GaNA (Shah & Foster, 2014a)

Table 10.2 Data collection schedule for assessing change in teachers' knowledge of game-based learning

	Intervention										
Data sources	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11
Background survey	Pre										
Teachers' Knowledge of GBL survey	Pre										Post
Game Integration Scenario test	Pre					Mid					Post
Focus group interview											Post
Observations											
GaNA guide											

pertaining to GBL through a focus on game analysis, game integration, and ecological conditions (see Table 10.1).

10.5.4 Measures for Scaffolding and Assessing Teachers' Examination of Games for Teaching and Learning

The *background survey included* a total of 21 close-ended, multiple choice, and open-ended items. Questions focused on ascertaining data about participants' (a) demographic and psychographic characteristics, (b) experience with digital media as part of their teacher education program, (c) experience with game playing, (d) prior training in GBL, and (e) expectations from the course. The survey was inspired by (a) Hayes and Orhnberger's (2013) work on investigating pre-service teachers' gaming and digital media practices and (b) the work of Zhao and colleagues (2002) on generating a portrait of teachers' beliefs about the pedagogical use of technology (games, in this study). The background survey informed the need for any modifications in the early weeks of the intervention and interpreting the nature of knowledge participants acquired in relation to GBL.

The *focus group interview* included 4 open-ended prompts and was led by an expert in qualitative research. Eight participants who volunteered to participate in the focus group were asked (a) to share what they learned from the course, (b) to discuss the associations they made about the potential of GBL in K-12 education and their future practice, and (c) to express any gaps or concerns that might hinder them in using games in the future. The relevance of focus group has been discussed in the context of exploratory studies, particularly to gain rich insights about the impact of an intervention (Liamputtong, 2011), and for participants to be heard in their own voices (Given, 2008). Responses from the focus group were intended to interpret the extent to which (a) participants had internalized GBL as a part of their future instructional goals as a teacher and (b) GaNA provided an all-encompassing method for educating teachers to use games.

The *Game Integration Scenario (GIS) test* included ten open-ended items. The test situated the participants in a hypothetical but realistic scenario which prompted teachers to articulate decisions about game analysis, game integration, and ecological conditions. Examples of questions included the following: (a) Why did you select this game? (b) How will the game be implemented? (c) What, if any, opportunities and challenges do you anticipate in facilitating the GBL lesson? The GIS test was designed to represent authentic questions technology-using teachers are likely to consider when planning a lesson with technology (in this case, a digital game) (Niess, 2008; Zhao, 2003). It was useful in assessing change in participants' emerging ability to apply their knowledge of GBL over the duration of the study (mechanical, meaningful to regenerative, Zhao, 2003), especially since participants were not going to be observed in their student teaching experience or a real classroom setting where they could implement GBL lessons. Details about the constructs (TPACK, PCARD, CITE) are measured in the test.

Teachers' Knowledge of Game-Based Learning (TKGBL) survey was developed to assess participants' knowledge about GBL through a 5-point Likert scale survey form (1 = strongly disagree; 2 = disagree; 3 = neither agree or disagree; 4 = agree; 5 = strongly agree). It consisted of subscales for assessing participants' knowledge of game analysis, game integration, and conditions impacting game use in education. The game analysis subscale (52 items) was created by modifying the Survey of Pre-service Teachers' Knowledge of Teaching and Technology (Schmidt et al., 2009) to gather data about participants' knowledge of selecting games for teaching and supporting student learning. The game integration subscale (39 items) collected data about participants' knowledge of employing games for teaching and supporting student learning. The ecological conditions subscale (6 items) assessed participants' awareness and skills needed to address technological, pedagogical, and social conditions surrounding technology integration in schools (Zhao et al., 2002). Cronbach's alpha obtained from a split-half reliability analysis indicated that the TKGBL survey, which included the three subscales of game analysis (TPACK), game integration (PCaRD), and ecological conditions (CITE), had good to excellent reliability (see Shah & Foster, 2015). Details about the constructs measured in the survey through each subscale, examples of items, and scoring can be found in another report (Shah & Foster, 2015).

An *observation protocol* was followed which involved note-taking about events as they occurred, while teaching a class (the condensed account) based on the weekly themes, adding details to earlier notes immediately after each class (the expanded account), reflecting on the progress of the course before planning for the forthcoming class (the daily log), and connecting the researcher's insights with the progress participants were making with respect to their knowledge of GBL through the course experiences (the ongoing analysis of interpretations; Haslam, 1987). As such, the observations, video, and field notes further aided in assessing and interpreting the knowledge participants acquired about GBL.

The *GaNA guide* was created to provide participants with a template that facilitated participants through the process of game analysis and game integration. Overall, GaNA guide gave all participants common language to execute, document, discuss, and design game analysis and game integration, both in class and outside (as part of course assignments). GaNA guide was instrumental in prompting participants (a) to determine the pedagogical and content stance in specific games and its relation to the participants teaching stances and learning goals, (b) to anchor the game to the participants' expertise in lesson planning and relevant education standards, and (c) to ascertain the game/GBL lesson's reliance on contextual conditions (Alexander, Eaton, & Egan, 2010; Halverson, 2005) (see Table 10.3).

10.5.5 Data Analysis

Overall, participants' knowledge of game-based learning was assessed using inductive and deductive analyses. Data obtained from background survey and focus group interview was analyzed using thematic analysis (Guest, MacQueen, & Namey, 2012) to identify shifts and insights in participants' thoughts about (a) the use of digital games in K-12 education, (b) their expectations from the course and what they learnt from the course, (c) the knowledge and skills teachers need to implement GBL in schools, and (d) the use of GBL in their future practice. A matched-paired t-test was used to assess the change in participants' self-reported knowledge on the TKGBL survey (pre-post) and on the GIS test (pre-mid-post). The significance level for all tests was set at $p < 0.05$.

This use of multiple data sources spread over the period of 11 weeks afforded the researcher to employ between-method triangulation to determine the extent to which the findings converged (Denzin, 1970). For instance, to understand the

Table 10.3 Example of items on the Game Network Analysis (GaNA) Guide

Knowledge of	Examples of items
Game analysis	
Technology (general)	Describe both the ease and difficulty of using this game (e.g., cost, installation, saving game)
Pedagogy (general)	What is the instructional style of this game? Or, how would you describe its teaching approach? (Inquiry)
	How does the game allow demonstration of understanding and accomplishment of objectives? (Construction)
	Does the game provide between-player communication? If so, please describe (communication)
	How can players customize their game play experience? (Avatars, selecting game play options, customization before game begins and during the game) (Expression)
Content	Which national and state core curriculum standards (such as NETS or content area standards) were addressed in this game?
Pedagogy (specific to the area of concentration)	How will students learn about the topic/academic content in the game? (Inquiry)
	What level of knowledge construction for the topic/academic content is possible in the game (content-only; content and application, but separate; content and application integrated – learning and doing are integrated well)? (Construction)
	How does the game guide and inform (feedback, instruction, adaptive to each player, tutorial, relevance, formative/summative) students in learning about the topic/academic content? (Communication)
	What opportunities are available for players to feel connected personally through freedom of expression to learn, perform, and demonstrate their learning of the topic/academic content? (Expression)
Game integration	
Play	List questions that you may use to observe students as they are playing (e.g., are they gaining opportunities for ICCE) List questions that you may use to observe the classroom (physical settings, seating, technical infrastructure)
Curricular activity	Develop a problem-based case or activity that addresses the objective of this lesson. Provide any additional material that is developed and used for the curricular activity How is the curricular activity connected with game play?
Reflection	Develop a reflection prompt or case that addresses the objective of this lesson. Provide any instruction(s) that will be given in relation to the reflection prompt or question
Discussion	Provide a driving question(s) to initiate the discussion What lessons are learned from the application of PCaRD in this session for future lessons?

change in participants' knowledge of game analysis, the TKGBL survey yielded self-report findings of how knowledgeable the participants believed they were to analyze games, whereas, the GIS test yielded objective findings for how well the participants were able to demonstrate their ability to analyze games and use the information to design game-based learning lesson plan. Furthermore, participants' activities logged in discussion boards and GaNA guide afforded additional insights about how their knowledge evolved over the duration of the course. Triangulation was essential for ensuing confidence in the findings from the study (Denzin, 1970). This also helped to reduce the researcher bias.

10.6 Results

To answer the research question, "To what extent did an 11-week methods course scaffold and assess pre-service teachers' knowledge and skills for GBL?," quantitative findings are presented for the group at large, and interpretive findings are presented for one case from the start of the intervention, during the intervention, and at the end of the intervention.

10.6.1 Group Findings

Table 10.4 summarizes the descriptive statistics for change in participants' knowledge on the TKGBL survey and the Game Integration Scenario (GIS) test. *T*-test results from the TKGBL survey indicated that participants had statistically significant knowledge gains for game analysis (TPACK), game integration (PCaRD), and ecological conditions impacting technology integration (CITE) (see Table 10.5).

Table 10.4 Descriptive statistics for Teachers' Knowledge of Game-Based Learning survey and Game Integration Scenario test

Groups	Means	SD	Std. err	*N*
Pre-survey of teachers' knowledge of game analysis (TPACK) (subscale of knowledge survey)	20.73	3.24	0.86	14
Post-survey of teachers' knowledge of game analysis (TPACK) (subscale of knowledge survey)	27.9	3.50	0.93	14
Pre-survey of teachers' knowledge of game integration (PCaRD) (subscale of knowledge survey)	31.55	4.91	1.31	14
Post-survey of teachers' knowledge of game integration (PCaRD) (subscale of knowledge survey)	42.31	3.96	1.06	14
Pre-survey of teachers' knowledge of conditions for integrating technology in education (CITE) (subscale of knowledge survey)	5.59	1.13	0.30	14
Post-survey of teachers' knowledge of conditions for integrating technology in education (CITE) (subscale of knowledge survey)	8.51	1.17	0.31	14
Pre-Game Integration Scenario test	13.07	2.63	0.70	14
Mid-Game Integration Scenario test	18.85	3.63	0.97	14
Post-Game Integration Scenario test	32.46	7.70	2.05	14

Table 10.5 Paired *t*-tests analysis of the teachers' knowledge of survey

Source	df	*t*	*p*	*d*
Pre-post game analysis knowledge	13	−9.78	0.001	2.12*
Pre-post game integration knowledge	13	−8.74	0.000	2.41**
Pre-post ecological conditions awareness and skills	13	−9.42	0.000	2.53***

Note: $P < 0.01$, $*R^2 = 0.52$, $**R^2 = 0.59$, $***R^2 = 0.61$

There was a significant difference in pre-service teachers' knowledge of game analysis from pretest ($M = 20.73$; SD = 3.24) to posttest ($M = 27.9$; SD = 3.50) from participating in the methods course in game-based learning. Similarly, a significant difference in pre-service teachers' knowledge of game integration from pretest ($M = 31.55$; SD = 4.91) to posttest ($M = 42.31$; SD = 3.96) was found. Lastly, there was a significant difference in pre-service teachers' knowledge of conditions impacting technology integration in education from pretest ($M = 5.59$; SD = 1.13) to posttest ($M = 8.51$; SD = 1.17). The effect sizes indicated that the course had a large effect on participants' knowledge of GBL.

Participants made statistically significant gains on the Game Integration Scenario test from pretest ($M = 13.07$, SD = 2.63) to mid-test ($M = 18.85$, SD = 3.63) to posttest ($M = 32.46$, SD = 7.70). Pretest scores were used as a covariate. Once again, the effect size indicated that the course had a large effect on participants' ability to apply the knowledge and skills involved in GBL as shown in Table 10.6.

10.6.2 Max: At the Start of the Intervention

Max was a 26-year-old graduate-level student who had completed a school practicum at the time of participating in this study. In his background survey and initial classroom discussions, Max reported playing casual games, mostly by himself, on mobile devices and some computer strategy games for 1–3 h each week with the following being his top three favorite games: Mario Kart, the Zelda games, RollerCoaster Tycoon. He engaged in the following game-related practices: visit game websites, read reviews and/or discussion boards, and help or guide others when playing. Max had considered using games for teaching but had never implemented them since he "never felt properly informed or trained in accomplishing this in a meaningful way." Max believed that if games were integrated well in schools, they could be beneficial in making content more relatable to the students. Max wanted to use games (a) to reinforce concepts in his content area (English/language arts), (b) to introduce themes or ideas for an upcoming unit, and (c) to assess student learning. At the start of the course, Max believed that in order for teachers to implement GBL in classrooms teachers needed to know how the games work, how to make the games educational, how to relate the games to the content area, and how

Table 10.6 Paired *t*-tests analysis of the Game Integration Scenario test

Source	df	*t*	*p*	*D*
Pre-mid Game Integration Scenario	13	−5.556	0.000	1.82*
Mid-post Game Integration Scenario	13	−6.389	0.000	2.26**
Pre-post Game Integration Scenario	13	−8.437	0.000	3.37***

Note: $P < 0.01$, $*R = 0.44$, $**R^2 = 0.54$, $***R^3 = 0.72$

to make the activity as productive and meaningful as possible. These ideas also captured what Max expected to learn from the course.

10.6.3 Max: During the Intervention

During the course, Max readily collaborated with different peers in in-class activities. Max was willing to explore a variety of games and listen keenly about the games his peers explored. Individually and with peers, Max was forthcoming in sharing his findings from analyzing games and conceptualizing ways to integrate specific games. Furthermore, Max drew upon his experience of playing games informally, and the content and pedagogical expertise he had acquired as part of his graduate program (see Table 10.7).

Since the course could not incorporate opportunities for classroom observation and implementation, participants learnt about contextual conditions impacting the implementation of games in the classroom through readings that explicitly highlighted the issues, challenges, and barriers teachers experience (Baek, 2008; Tuzun, 2007). In discussing these readings, Max shared his insights in classroom and online discussions. Max believed that a school's infrastructure was a "realistic and tangible barrier to game integration." He believed he learnt a few options to tackle the issue from this course. For instance:

> [the instructor] had mentioned that there are numerous grants for technology available, and that applying to, or encouraging an administrator to apply for these, is a viable option to get funding for more technology in the school… It will take a good deal of extra work on the [teacher's] part to accomplish these, and will require perseverance. But if game integration is the goal, these are ways of accomplishing it when an infrastructure barrier is in the way.

As part of his mid-term assignment, Max individually analyzed the learning affordances of the interdisciplinary game *Minecraft* along the TPC and ICCE dimensions. He documented an overview of the game play. This was followed by an assessment of *Minecraft* for teaching English/language arts concepts. Toward the end of the course, Max developed an action plan for employing Minecraft using PCaRD for teaching an English/language arts class in a ninth-grade classroom (see Shah, Foster, Scottoline, & Duvall, 2014).

10.6.4 Max: At the Conclusion of the Intervention

The course experiences expanded his knowledge of the process of game-based learning. He felt more informed about the importance of analyzing games. Max believed he was more prepared than before to analyze games. He shared the following in an online discussion forum of the class at the end of the course:

> From my time in the course, it's become extremely clear that educators need to focus on playing the games and deeply analyzing them in order to successfully integrate games in the classroom. If a teacher does not do a proper investigation into the game—whether or not it is good for learning, interesting to play, or applicable to the curriculum—it is likely the game integration will not go very well.

Max was one of the eight participants who volunteered to partake in the focus group discussion. When asked about what he learnt about GBL from the course, Max responded that successful game integration requires careful planning and analysis beforehand. He elaborated upon this point:

> Well…there was a whole game analysis guide that we had and we practiced it a lot. Using that and going very deep into the game and looking at it from almost every conceivable angle was really.. that seems like the way to me game integration could be done successfully, paired with like lesson design like the PCaRD lesson plan that we were studying. You know having that structure and consistency is what really helped.

In response to a question about what skills students learn when playing games, Max talked about learning in games as occurring both in the game and activities beyond it that teachers use to helps students make a connection between the content and the game. He believed that this approach was better than just telling students what they learn because:

> **Max:** I think it builds more of a realistic connection between sort of the world so to speak and the content. If you are having an experience whatever that may be in the game and then later you recognize that "Oh, what I am learning here directly relates to that!" I think that sort of connectivity is important.
>
> **Facilitator:** It is quite a discovery!
>
> **Max:** Yeah!

Max concluded that the intervention convinced him to use games in his future practice. Although some of his peers were concerned about the lack of resources in the city's schools, Max was positive that even in schools where resources may seem scarce, teachers can "sort of seek it as best as you can."

10.7 Discussion and Conclusions

This chapter exemplified the efficacy of Game Network Analysis (GaNA) framework in scaffolding and assessing 14 pre-service teachers' knowledge of game-based learning (GBL) in an 11-week methods course. In doing so, the chapter makes few contributions to the emergent field of teacher education in game-based learning

Table 10.7 Max's participation in game analysis and integration activities

Week	Matt	Example of findings
2	Played *Fatworld* with Kelly	Fatworld attempts to highlight the issues in America concerning nutritional policy and the socioeconomic, political, and health factors that affect the everyday lives of individuals. The game is designed to set the player free in a small "world," where they must make a meal plan, buy groceries, and create income in effort to maintain their lifestyle and survive. This game seems to be designed for students who are a bit older, perhaps middle school ages. This could be a beneficial game for health class or social science classes, where the focus can either be on diet or the politics of nutrition. There is no specific content being taught, but the themes presented in the game could be used to lead a discussion or a reflective assignment. Players learn through experience in the game and the consequences of the decisions that they make on the health and lives of their character. For example, if your character is or becomes overweight, due to their diet, it can become more difficult to move around the town, as they will become fatigued and not be able to walk for a sustained period of time The overall goal of the game is to live to be 100 years old. The main consequence presented by the game is the ability for your character to die. There are also economic aspects of the game, as your character can run a business. We were not able to get far enough into the game to explore this portion, but it's possible the success or failure of your business is another thing for the players to consider, as this will impact their finances and abilities to maintain their diet plans It's an interesting game and something that likely could be used in the classroom to present certain topics and give students a tangible way to deal with these kinds of issues
3	Played *Spent* individually. The whole class analyzed a common game followed by a discussion	It seems to me that Spent is a game designed for young adults and even older adults, with the goal of highlighting the economic issues facing many Americans. On a pedagogical level, the game's use of budgeting and dealing with real life scenarios is successful. While playing, the player can really feel the weight of many of the decisions and situations they are faced with. Often times, these situations are incredibly emotional, and the game uses strong, extreme language in the presentation. There were instances where the two options presented were extreme and did not quite reflect reality; but overall, it is well done. Many of the scenarios caused me to pause and evaluate possible consequences of my decision. It's an incredibly interesting game and one that I think is genuinely important for more people to experience. It gives players the opportunity to walk in the shoes of someone in a low-income or financially disadvantaged situation and to begin to form empathetic feelings toward those who find themselves in these situations. I would absolutely use this game in the classroom. It could be used in a variety of subjects: English if there is a story regarding low-income characters, history/social science to discuss income inequality, and even math, to discuss budgeting. It's also a good tool for economic study and the risk/reward system

(continued)

Table 10.7 (continued)

Week	Matt	Example of findings
4	Played Hot Shot for Business individually. Paired with Peter to document findings using GaNA guide	The game is fairly easy to use as long as you have access to a computer and the Internet. There is no cost to play the game, and no additional technology is required (*Technology*) To meet the objective, players must utilize marketing, consumer data, and risk assessment to make the right decisions that will enable their business to generate sufficient revenue. They can listen to the advice of the characters in the game, as well as analyze other data like news reports and research to see what customers want to buy. Students must be able to follow instructions and data in order to succeed (pedagogy, inquiry) The game addresses NETS 3, research and information; 4, critical thinking, problem solving, and communicating; and 6, technology operations and concepts (*content*) The game guides with tutorials and explanations at every new step of the game. In addition, there are performance reviews every game week that allow students to learn about the content and reflect on their understanding (*pedagogy, specific to content, communication*)
7	Designed a lesson plan for Citizen Science using GaNA guide with Hannah and Kelley. The game was individually analyzed by all in week 6	To use the game for English/language arts – logical vs. emotional appeals Students will be able to identify logical and emotional appeals through game play and class activities and use Citizen Science for collecting evidence to create valid arguments with logical appeals (*lesson goals*) What makes an argument valid? What are the elements of a logical appeal? Emotional appeal? Why is it important to make logical appeals in persuasive writing? In what circumstances is an emotional appeal valid? (*Essential questions*) 5–10-min direct instruction on appeals; examples given from the game and other sources. Give situations and have students give appeals for the situation. Play devil's advocate during discussion, and give statements that students will identify as emotional or logical (*curricular activity*) During the game, were there more emotional arguments or logical arguments? Why did the characters in the game require logical arguments to persuade them? (*Discussion*)

(Molin, 2017) that speak to the focus of this volume on revisiting game-based assessments (Groff, 2018).

First, GaNA offers one theoretical and methodological approach to frame teachers' knowledge, that is, game analysis, game integration, and ecological conditions (Foster, 2012; Shah & Foster, 2015). The analytical and pedagogical frameworks that constitute GaNA allow teachers, teacher educators, and researchers to systematically scaffold and assess teachers' ability (a) to select, explore, and evaluate the learning affordances of games, (b) to synergistically use game knowledge and their curricular expertise to design experiences targeted toward specific goals, and (c) to be conscious of the context in which the game is to be introduced. A methodological framework such as GaNA is useful because teachers who are interested in adopting

GBL have expressed concerns about the lack of systematic guidance in identifying games, obtaining a contextual knowledge about them, and integrating games in their curriculum (Ruggiero, 2013; Takeuchi & Vaala, 2014). This gap is a crucial one to address because teachers' knowledge of games is inextricably tied to their (a) decisions related to designing and iteratively refining game-based curricula (Eastwood & Sadler, 2013; Shah & Foster, 2014b), (b) ability to identify teachable moments and facilitate meaningful discourse (Silseth, 2012; Watson et al., 2011), and (c) efforts at assessing change in students' knowledge and motivation (e.g., Barab et al., 2009; Foster, 2011). Furthermore, few studies have supported pre-service teachers' exploration of embedded content in games and the pedagogical implications of game genres (Barbour et al., 2009; Becker, 2007; Gopin, 2018). Even fewer studies have facilitated participants to incorporate their game knowledge in designing learning activities, leading a game-based session, and evaluating student learning in a game-based classroom (Sardone & Devlin-Scherer, 2010). GaNA, then, offers a valuable approach to comprehensively frame teachers' knowledge for using games and to systematically facilitate the development of those skills. This conceptual and pedagogical operationalization is important so that researchers and teacher educators can support teachers in leading contemporary classrooms wherein student learning and outcomes are facilitated in synergy with the context-attuned affordances of the chosen games (and other interactive and interdisciplinary play-based environments) and teachers' praxis (Mehta et al., 2019; Shaffer et al., 2015; Shah, Petrovich, Foster, Schaar, & Chen, 2019).

Second, the study illustrated how formative (e.g., Game Network Analysis Guide) and summative (e.g., Teachers' Knowledge of Game-Based Learning survey) assessments were created using the GaNA framework (a) to support participating pre-service teachers to examine games as a form of curriculum and (b) to afford the researcher to qualitatively and quantitatively capture the change in teachers' game literacy and the extent to which it was integrated with teachers' design of game-based lesson plans. The case study of Max, a male pre-service English/language arts teacher, illustrated how the 11-week course afforded him opportunities (a) to examine the situated affordances and constraints of games through direct and vicarious ways, individually and in collaboration with peers, (b) to document findings about multiple games, and (c) to reflect on the possibilities of repurposing games for curricular use. Scalable models that simultaneously prompt consideration for guiding and examining changes in teachers' examination of games, both for the facilitator and the teacher, are an essential first step toward developing teachers' competence in identifying, designing, and enhancing teaching, learning, and assessment possibilities with playful environments such as games (Hsu et al., 2015; Jong & Shang, 2015; Shah & Foster, 2018; Wu, 2015).

Third, as the nascent field advances, it will be important to not consider teachers' knowledge and skills in adopting technological and pedagogical innovations as all encompassing. Teachers intention to use games and their ability to see themselves in the role of a professional who is able to facilitate learning with these complex technologies will be an equally important goal that will require scaffolding and assessing (Chee et al., 2015; Sanchez-Mena et al., 2017; Shah & Foster, 2018).

Pursuing such a direction in future examinations will facilitate teachers to not only examine what knowledge and skills do they want to facilitate and assess with games but also why those may be important.

In conclusion, new media tools such as games and maker tools have galvanized the energy around play as a medium of learning in novel ways. Games are dynamic environments for engaging students in meaningful learning opportunities (Foster, 2008) and assessing what they learn (Groff, 2018). Learners are afforded with individual, participatory, and connected learning opportunities in and out of school to experiment with complex ideas and possible roles and to explore new interests and deepen existing ones. Yet, less attention is given to the praxis of teaching and teachers' preparation in using these environments. Promoting teachers' capacity to examine games and integrate them well, especially early in their careers, will catalyze the nature of how games and other playful environments are leveraged for their teaching, learning, and assessment endeavors.

Acknowledgements I want to extend my gratitude to Drs. Aroutis Foster, Elizabeth Haslam, Diane Jass Ketelhut, and Sarah Reynolds Ulrich. Their guidance and support for my doctoral study are invaluable.

References

Aarseth, E. (2003). *Playing research: Methodological approaches to game analysis.* Paper presented at the Digital Arts and Culture conference, Melbourne, Australia.

Alexander, G., Eaton, I., & Egan, K. (2010). Cracking the code of electronic games: Some lessons for educators. *Teachers College Record, 112*(7), 1830–1850.

Allsop, Y., Yeniman Yildirim, E., & Screpanti, M. (2013). Teachers' beliefs about game based learning: A comparative study of pedagogy, curriculum and practice in Italy, Turkey and the UK. In *The Proceedings of The 7th European Conference on Games Based Learning* (pp. 1–10), 3–4 October, Porto, Portugal.

An, Y. (2018). The effects of an online professional development course on teachers' perceptions, attitudes, self-efficacy, and behavioral intentions regarding digital game-based learning. *Educational Technology Research and Development, 66*, 1505–1527.

Baek, Y. K. (2008). What hinders teachers in using computer and video games in the classroom? Exploring factors inhibiting the uptake of computer and video games. *CyberPsychology and Behavior, 11*(6), 665–671.

Barab, S. A., Scott, B., Siyahhan, S., Goldstone, R., Ingram-Goble, A., Zuiker, S. J., & Warren, S. (2009). Transformational Play as a Curricular Scaffold: Using Videogames to Support Science Education. *Journal of Science Education and Technology, 18*(4), 305–320.

Barbour, M. K., Rieber, L. P., Thomas, G. B., & Rauscher, D. (2009). Homemade PowerPoint games: A constructionist alternative to WebQuests. *Tech Trends, 53*(5), 54–59.

Barzilai, S., & Blau, I. (2014). Scaffolding game-based learning: Impact on learning achievements, perceived learning, and game experiences. *Computers & Education, 70*, 65–79.

Becker, K. (2007). Digital game-based learning once removed: Teaching teachers. *British Journal of Educational Technology, 38*, 478–488.

Bell, A., & Gresalfi, M. S. (2017). Teaching with videogames: How experience impacts classroom integration. *Technology, Knowledge and Learning, 22*(3), 1–14.

Boyle, E. A., Hainey, T., Connolly, T. M., Gray, G., Earp, J., Ott, M., … Pereira, J. (2016). An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. *Computers & Education, 94*, 178–192.

Caldwell, K. E. H., Osterweil, S., Urbano, C., Tan, P., & Eberhardt, R. (2017). "I Just Don't Know Where to Begin": Designing to facilitate the educational use of commercial, off-the-shelf video games. In M. Ma & A. Oikonomou (Eds.), *Serious games and edutainment applications* (Vol. II, pp. 625–848). Cham, Switzerland: Springer.

Chee, Y. S., Mehrotra, S., & Ong, J. C. (2015). Professional development for scaling pedagogical innovation in the context of game-based learning: Teacher identity as cornerstone in "shifting" practice. *Asia-Pacific Journal of Teacher Education, 43*(5), 423–437.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research, 86*(1), 79–122.

Creswell, J. W., & Clark, V. L. P. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

de Freitas, S. (2018). Are games effective learning tools? A review of educational games. *Journal of Educational Technology & Society, 21*(2), 74–84.

Demirbilek, M., & Tamer, S. L. (2010). Math teachers' perspectives on using educational computer games in math education. *Procedia - Social and Behavioral Sciences, 9*, 709–716.

Denzin, N. K. (1970). *The research act in sociology*. Chicago, IL: Aldine.

Dewey, J. (1956). *The child and the curriculum; and, the school and society*. Chicago, IL: University of Chicago Press.

Eastwood, J. L., & Sadler, T. D. (2013). Teachers' implementation of a game-based biotechnology curriculum. *Computers & Education, 66*, 11–24.

Fishman, B., Riconscente, M., Snider, R., Tsai, T., & Plass, J. (2014). *Empowering educators: Supporting student progress in the classroom with digital games*. Ann Arbor, MI: University of Michigan.

Foster, A. (2008). Games and motivation to learn science: Personal identity, applicability, relevance and meaningfulness. *Journal of Interactive Learning Research, 19*, 597–614.

Foster, A., & Shah, M. (2015a). The ICCE framework: Framing learning experiences afforded by games. *Journal of Educational Computing Research, 51*(4), 369–395.

Foster, A., & Shah, M. (2015b). The play curricular activity reflection and discussion model for game-based learning. *Journal of Research on Technology in Education, 47*(2), 71–88.

Foster, A., Shah, M., & Duvall, M. (2015). Game network analysis: For teaching with games. In M. L. Niess & H. Gillow-Wiles (Eds.), *Handbook of research on teacher education in the digital age* (pp. 380–411). Hershey, PA: Information Science Reference.

Foster, A. N. (2011). The process of learning in a simulation strategy game: Disciplinary knowledge construction. *Journal of Educational Computing Research, 45*, 1–27.

Foster, A. N. (2012). Assessing learning games for school content: Framework and methodology. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives*. New York, NY: Springer.

Franklin, T., & Annetta, L. (2011). Special issue: Digital games and simulations in teacher preparation. *Journal of Technology and Teacher Education, 19*(3), 239–242.

Gaydos, M. J., & Devane, B. M. (2019). Designing for identity in game-based learning. *Mind, Culture, and Activity, 26*(1), 61–74.

Gee, J. P. (2005). Good Video Games and Good Learning. *Phi Kappa Phi Forum 85*(2), 33–37.

Given, L. M. (2008). *The Sage encyclopedia of qualitative research methods*. Los Angeles, CA: Sage.

Gopin, E. (2018). Finding and evaluating great educational games. In I. Management Association (Ed.), *Gamification in education: Breakthroughs in research and practice* (pp. 323–337). Hershey, PA: IGI Global.

Groff, J., McCall, J., Darvasi, P., & Gilbert, Z. (2016). Using games in the classroom. In K. Schrier (Ed.), *Learning, education and games* (Bringing games into educational contexts) (Vol. 2, pp. 19–41). Pittsburgh, PA: ETC.

Groff, J. S. (2018). The potentials of game-based environments for integrated, immersive learning data. *European Journal of Education, 53*(2), 188–201.

Guest, G., MacQueen, K. M., & Namey, E. E. (2012). *Applied thematic analysis*. Los Angeles, CA: Sage.

Halverson, R. (2005). What can K-12 school leaders learn from video games and gaming? *Innovate: Journal of Online Education, 1*(6). Retrieved from http://www.innovateonline.info/index.php?view=article&id=81

Hanghøj, T., & Brund, C. E. (2011). Teachers and serious games: Teachers roles and positionings in relation to educational games. In *Serious games in education: A global perspective* (1st ed., pp. 125–136). Copenhagen, Denmark: Aarhus Universitetsforlag.

Hanghøj, T., & Hautopp, H. (2016). Teachers' pedagogical approaches to teaching with Minecraft. In *10th European Conference on Games Based Learning: ECGBL 2016*. Academic Conferences and Publishing Limited.

Haslam, E. L. (1987). *The design, implementation, and evaluation of a college inquiry writing course* (Unpublished dissertation). University of Pennsylvania.

Hayes, E., & Orhnberger, M. (2013). The gamer generation teaches school—The gaming practices and attitudes towards technology of pre-service teachers. *Journal of Technology and Teacher Education, 21*(2), 153–177.

Hsu, C.-Y., Liang, J.-C., & Su, Y.-C. (2015). The role of the TPACK in game-based teaching: Does instructional sequence matter? *Asia-Pacific Education Researcher, 24*(3), 463–470.

Hwang, G. J., & Wu, P. H. (2012). Advancement and trends in digital game-based learning research: A review of publications in selected journals from 2001 to 2010. *British Journal of Educational Technology, 43*(1), E6–E10.

Jaipal, K., & Figg, C. (2009). Using video games in science instruction: Pedagogical, social, and concept-related aspects. *Canadian Journal of Science, Mathematics and Technology Education, 9*(2), 117–134.

Jong, M. S.-Y., & Shang, J. (2015). Impeding phenomena emerging from students' constructivist online game-based learning process: Implications for the importance of teacher facilitation. *Educational Technology & Society, 18*, 262–283.

Kafai, Y. B., & Burke, Q. (2015, January 1). Constructionist gaming: Understanding the benefits of making games for learning. *Educational Psychologist, 50*(4), 313–334.

Kangas, M., Koskinen, A., & Krokfors, L. (2016). A qualitative literature review of educational games in the classroom: The teacher's pedagogical activities. *Teachers and Teaching, 23*, 451.

Kennedy-Clark, S., Galstaun, V., & Anderson, K. (2015). Death in Rome: Using an Online Game for Inquiry-Based Learning in a Pre-Service Teacher Training Course. In I. *Management Association (Ed.), STEM Education: Concepts, Methodologies, Tools, and Applications* (pp. 1118-1132). Hershey, PA: IGI Global. https://doi.org/10.4018/978-1-4666-7363-2.ch060.

Kenny, R., & Gunter, G. (2011). Factors affecting adoption of video games in the classroom. *Journal of Interactive Learning Research, 22*(2), 259–276.

Kenny, R. F., & McDaniel, R. (2011). The Role Teachers' Expectations and Value Assessments of Video Games Play in Their Adopting and Integrating Them into Their Classrooms. *British Journal of Educational Technology, 42*(2), 197–213.

Ketelhut, D. J., & Schifter, C. C. (2011). Teachers and game-based learning: Improving understanding of how to increase efficacy of adoption. *Computers & Education, 56*(2), 539–546.

Liamputtong, P. (2011). *Focus group methodology: Principles and practices*. Los Angeles, CA: Sage.

Marklund, B. B., & Taylor, A.-S. A. (2015). Teachers' many roles in game-based learning projects. In R. Munkvold & L. Kolås (Eds.), *Proceedings of the 9th European Conference on Games Based Learning* (pp. 359–367). Reading, UK: Academic Conferences and Publishing International Limited.

Mehta, R., Henriksen, D., & Rosenberg, J. M. (2019, January 1). It's not about the tools. *Educational Leadership, 76*(5), 64–69.

Mishra, P., & Koehler, M. (2006). Technological pedagogical content knowledge: A new framework for teacher knowledge. *Teachers College Record, 108*(6), 1017–1054.
Molin, G. (2017). The role of the teacher in game-based learning: A review and outlook. In M. Ma & A. Oikonomou (Eds.), *Serious games and edutainment applications* (Vol. II, pp. 649–674). Cham, Switzerland: Springer.
Moline, T. (2009). Descriptors of quality teachers and quality digital games. In *Handbook of research on effective electronic gaming in education* (pp. 652–669). Hershey, PA: IGI Global.
Nardi, B. A., & O'Day, V. (1999). *Information ecologies: Using technology with heart*. Cambridge, MA: MIT.
Niess, M. L. (2008). Guiding preservice teachers in developing TPCK. In *American Association of Colleges for Teacher Education. Handbook of technological pedagogical content knowledge (TPCK) for educators* (pp. 223–250). New York, NY: Routledge for the American Association of Colleges for Teacher Education.
Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist, 50*(4), 258–283.
Raphael, R. (2018). Towards a model of playful learning: Gamification strategies in the i2Flex classroom. In I. Management Association (Ed.), *Gamification in education: Breakthroughs in research and practice* (pp. 397–414). Hershey, PA: IGI Global.
Ruggiero, D. (2013). *Video games in the classroom: The teacher point of view.* Paper presented at the Games for Learning Workshop of the Foundations of Digital Games Conference, Chania, Greece.
Sanchez-Mena, A., Marti-Parreno, J., Sanchez-Mena, A., & Aldas-Manzano, J. (2017). The role of perceived relevance and attention in teachers' attitude and intention to use educational video games. *International Journal of Emerging Technologies in Learning, 12*(3), 154–168.
Sardone, N. B., & Devlin-Scherer, R. (2010). Teacher candidate responses to digital games: 21st-century skills development. *Journal of Research on Technology in Education, 42*(4), 409–425.
Schmidt, D. A., Baran, E., Thompson, A. D., Mishra, P., Koehler, M. J., & Shin, T. S. (2009). Technological pedagogical content knowledge (TPACK): The development and validation of an assessment instrument for preservice teachers. *Journal of Research on Technology in Education, 42*(2), 123–149.
Shaffer, D. W., Nash, P., & Ruis, A. R. (2015). Technology and the new professionalization of teaching. *Teachers College Record, 117*(12), 1–30.
Shaffer, D. W., Squire, K. R., Halverson, R., & Gee, J. P. (2005). Video games and the future of learning. *Phi Delta Kappan, 87*(2), 104-111.
Shah, M., & Foster, A. (2014a). The inquiry, communication, construction and expression (ICCE) framework for understanding learning experiences in games. Special issue on applied research in immersive environments for learning. *International Journal of Virtual and Personal Learning Environments, 5*(2), 1–14.
Shah, M., & Foster, A. (2014b). Undertaking an ecological approach to advance game-based learning: A case study. Special issue on game-based learning for 21st century transferable skills: Challenges and opportunities. *Journal of Educational Technology & Society, 17*(1), 29–41.
Shah, M., & Foster, A. (2015). Developing and assessing teachers' knowledge of game-based learning. *Journal of Technology and Teacher Education, 23*(2), 241–267.
Shah, M., & Foster, A. (2018). Promoting teachers' identity exploration: The way forward in teacher education for game-based learning. In E. Langran & J. Borup (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference* (pp. 486–494). Washington, DC: Association for the Advancement of Computing in Education (AACE).
Shah, M., Foster, A., Scottoline, M., & Duvall, M. (2014). Pre-service teacher education in game-based learning: Analyzing and integrating Minecraft. In M. Searson & M. Ochoa (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference 2014* (pp. 2646–2654). Chesapeake, VA: AACE. Retrieved from http://www.editlib.org/pv/131186

Shah, M., Petrovich, M., Foster, A., Schaar, R., & Chen, D. (2019). Change in role identity of an environmental science educator who desires to facilitate learning by making. In K. Graziano (Ed.), *Proceedings of Society for Information Technology & Teacher Education International Conference* (pp. 1635–1645). Las Vegas, NV: Association for the Advancement of Computing in Education (AACE). Retrieved from https://www.learntechlib.org/primary/p/207863/

Silseth, K. (2012). The multivoicedness of game play: Exploring the unfolding of a student's learning trajectory in a gaming context at school. *International Journal of Computer-Supported Collaborative Learning, 7*(1), 63–84.

Siyahhan, S., Ingram-Goble, A., Barab, S., & Solomou, M. (2017). Educational games to support caring and compassion among youth: A design narrative. *International Journal of Gaming and Computer-Mediated Simulations, 9*(1), 59–74.

Spires, H. A., & Lester, J. C. (2016). Game-based learning: Creating a multidisciplinary community of inquiry. *On the Horizon, 24*, 88–93.

Squire, K., & Barab, S. (2012). *Games, learning, and society: Learning and meaning in the digital age* (pp. 279–305). Cambridge, UK: Cambridge University Press.

Squire, K., & Jenkins, H. (2003). Harnessing the power of games in education. *Insight, 3*, 5–33.

Stieler-Hunt, C. J., & Jones, C. M. (2017). Feeling alienated–teachers using immersive digital games in classrooms. *Technology, Pedagogy and Education, 26*(4), 457–470.

Takeuchi, L. M., & Vaala, S. (2014). *Level up learning: A national survey on teaching with digital games*. New York, NY: The Joan Ganz Cooney Center at Sesame Workshop.

Tuzun, H. (2007). Blending video games with learning: Issues and challenges with classroom implementations in the Turkish context. *British Journal of Educational Technology, 38*, 465–477.

Watson, W. R., Mong, C. J., & Harris, C. A. (2011). A case study of the in-class use of a video game for teaching high school history. *Computers & Education, 56*(2), 466–474.

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.

Wu, M. L. (2015). *Teachers' experience, attitudes, self-efficacy and perceived barriers to the use of digital game-based learning: A survey study through the lens of a typology of educational digital games* (PhD dissertation). Michigan State University.

Zhao, Y. (2003). What teachers need to know about technology? Framing the question. In Y. Zhao (Ed.), *What should teachers know about technology* (pp. 1–14). Greenwich, CT: Information Age Publishing.

Zhao, Y., Pugh, K., Sheldon, S., & Byers, J. L. (January 01, 2002). Conditions for Classroom Technology Innovations. *Teachers College Record, 104*, 482–515.

Part III
Best Practice Implementations

Chapter 11
Assessing Game-Based Mathematics Learning in Action

Fengfeng Ke, Biswas Parajuli, and Danial Smith

11.1 Introduction

Digital learning environments, such as games and simulations, emphasize learning in action. In these learning environments, knowledge is constituted in actions or “intellectual practices” in which a learner work with open and question-generating knowledge objects, study problems through active engagement, and acquire understanding via reflective inquiries (Dewey, 1910; Eriksson & Lindberg, 2016). Because knowledge is present in what learners do, how they do it, what tools they use, and how they communicate in and about their doing, it is important to assess knowledge production in context and learning in action rather than testing abstracted and isolated skills and understanding. Tracking and diagnosing the learners’ learning in progress also enables the provision of learner-adaptive learning supports, thus fostering both autonomous and guided intellectual practices in the digital learning environment.

Prior research suggested that process-oriented data mining and learning analytics methods, such as Bayesian networks, social networks or the structural analysis, the visual or graphical analysis of event paths, and the sequential analysis of time series, can capture the complex and open-ended learning trajectories in a digital learning environment (Bakhshinategh, Zaiane, ElAtia, & Ipperciel, 2018; Ke, Shute, Clark, & Erlebacher, 2019; Manjarres, Sandoval, & Suárez, 2018; Papamitsiou & Economides, 2014). However, empirical research that provides an in situ examination and a rich description of the learning-in-action assessment via educational data mining, especially that examining mathematical learning in action (e.g., Ayers & Junker, 2006; Pardos, Heffernan, Ruiz, & Beck, 2008), is still limited.

F. Ke (✉) · B. Parajuli · D. Smith
Educational Psychology and Learning Systems, Florida State University, Tallahassee, FL, USA
e-mail: fke@fsu.edu; bparajuli@fsu.edu; dsmith@fsu.edu

D. Ifenthaler, Y. J. Kim (eds.), *Game-Based Assessment Revisited*, Advances in Game-Based Learning, https://doi.org/10.1007/978-3-030-15569-8_11

Via a design-based research approach, we explored the feasibility and validity of using the approach of evidence-centered design (Mislevy, Almond, & Lukas, 2003) along with Bayesian networks to assess mathematical learning in a game-based learning environment. We iteratively tested, compared, and refined the core elements of the assessment models and alternative processes of exploiting game-based performance data, via longitudinal data sets collected during the course of 42 gaming sessions (50 min per session) across 3 academic semesters. This current investigation focused on extracting design and implementation heuristics related to the game-based, learning-in-action assessment, by addressing the following questions:

- What are the salient patterns and strategies of operationalizing the learning-in-action assessment in a game-based math learning environment?
- What is the feasibility and validity of the learning-in-action assessment via the Bayesian network?

11.2 Theoretical Framework

11.2.1 Evidence-Centered Design

Evidence-centered assessment design (ECD, Mislevy et al., 2003; Mislevy, Haertel, Riconscente, Rutstein, & Ziker, 2017) is an established conceptual design framework for developing and implementing performance-based assessment. ECD consists of conceptual and computational models that work together. The three principal processes or models are the competency model, the evidence model, and the task model. The *competency model* (or domain competency modeling) outlines in a structured fashion a collection of knowledge, skills, and other attributes to be assessed. The *task model* identifies the features of selected tasks for learners that provide evidence about their target competencies. The *evidence model* serves as the bridge between the competency model and the task model. It transmits evidence elicited by tasks specified by the task model to the competency model by connecting the evidence model variables and competency model variables statistically. A design claim of ECD is that the modularized assessment specification allows it to be reassembled in different configuration for different assessment purposes. Different from the conventional forward-design assessment process, ECD integrates and encourages exploratory method of educational data mining (Mislevy, Behrens, Dicerbo, & Levy, 2012).

In this study, we adopted the evidence-centered design approach to develop game-based assessment. Specifically, we started by defining the claims to be made about participants' math competencies (i.e., competency modeling), establishing what actions or elements of gameplay constitute valid evidence of the claim (i.e., evidence modeling), and determining the nature and form of game tasks that elicit that evidence (task modeling). Although assessment design flows from competency to task modeling, diagnosis flows in the opposite direction. That is, the learners' performance (e.g., recorded by game logs) during a game level/task provides the evidence or data (e.g., logged scores of observable variables) that are passed on to the competency model, which in turn updates the claims (e.g., probabilities) about relevant competencies.

11.2.2 Data-Driven, Game-Based Assessment

Major published reports on the status of US mathematics education indicate critical shortcomings in achievement of US students at the middle and high school level (Carnoy & Rothstein, 2013). Research suggests that digital games present a realistic framework and a story context for experimentation and situated understanding and have positive cognitive and motivational effects on the development of multi-stranded mathematical proficiency: understanding, problem solving, and positive disposition (Clark, Tanner-Smith, & Killingsworth, 2016; Ke, 2016). Yet a major challenge of using games for math learning is to use in-game diagnostic assessment to capture game-based math competency development.

Recent studies (e.g., Dede, 2012; Kang, Liu, & Qu, 2017; Levy, 2014; Shaffer et al., 2009; Shute & Ventura, 2013; Taub, Azevedo, Bradbury, Millar, & Lester, 2018) have exemplified the applicability of game-based learning assessment via educational data mining. These studies all adopted a data-intensive, evidence-based approach by collecting, measuring, analyzing, and reporting dynamic data about learner performance and contexts (e.g., online log data) to understand learning and the environments in which it occurs. Multiple methods of data analyses (e.g., quantitative psychometric modeling, network or structural analysis, and path analysis) and visualization (e.g., algorithms, models, network graphs, or spatial and chronical maps) were used. These study results suggested that educational data mining can and should be used to exploit game-based performance data to inform on students' on-task or off-task behaviors, competency development related to the targeted subject matter, and hence the effectiveness and design of digital learning environments.

11.3 Method

Adopting the design-based research approach (Sandoval & Bell, 2004) and the multiple-case study method, we explored the heuristics of translating the aforementioned theoretical insights or conceptual claims into the design and practice of game-based math learning assessment. By iteratively designing, infield testing, and refining variant processes of assessment modeling along with alternative approaches of gameplay data mining, we aimed to delineate and examine "analytic generalizations" governing the design of evidence-centered, learning-in-action assessment (Yin, 2013). The research involves multiple phases of theory-driven design efforts that are examined and refined via infield testing. In the initial phase, we conceptualized core assessment models and the architecture of the game-based learning-in-action assessment. We then designed and infield-tested a Bayesian network in Netica to drive the assessment of game-based math learning. We also explored and compared the feasibility and validity of using a package of alternative data mining methods, including random forest, k-nearest neighbors classification (KNN), support vector machine (SVM), and Gaussian Bayes, to analyze the gameplay data. In phase 3, we refined the assessment models, the Bayesian network, and the game logging structure based on the collected gameplay data and the performance of the assessment and statistical models. We also explored and compared alternative approaches of training and validating the developed Bayes network.

11.3.1 Game-Based Learning Environment

We developed a 3D game-based learning environment (called E-Rebuild) that situates mathematical problems in the context of architectural design quests. In this multi-episode game, the overarching goal is to rebuild a disaster-damaged space, while fulfilling preset design criteria and needs. Each game level (or task) of E-Rebuild embodies a multimodal, math context problem. Aligned with the Common Core State Standards for mathematics Grade 6–8, the targeted mathematical competencies include (a) understanding and using ratio reasoning and proportional relationships to solve mathematical problems, (b) solving math problems involving area and volume, and (c) solving math problems using numerical and algebraic expressions.

Endorsing the perspectives of reflective inquiry and epistemic practices (Dewey, 1910; Eriksson & Lindberg, 2016), we designed core gameplay actions of E-Rebuild, including site surveying, item collection and trading, structure building, and allocation, to make players interact with and encode the architecture-themed math problem. It was conjectured that these gameplay actions would delineate players' interactions with the game-based math problem and capture their actions (or evidence) of problem interpretation, math knowledge application, and problem solving. For example, via the structure-building action, players engaged in *composing and decomposing* geometric shapes (such as stacking cuboids for a stadium stair bench), *covering* or calculating the area of a given space (such as painting a basketball court), *surrounding* or calculating the perimeter of a structure (such as fencing a farm), or *filling* or calculating the volume of a vessel or cavity (such as a fish pond). In other words, this building action aimed to activate and externalize students' comprehension and application of the verbal, graphical, and numerical representations of a geometrical math problem and the related math concepts.

11.3.2 Data Collection and Analysis

Gameplay data sets were collected via game logs from 120 middle school (6-8th grade) students during the course of 42 gaming sessions (50 min per session) across three academic semesters (or three iterative case studies). Students recruited in each semester were a different student population. The design and infield testing of the game-based learning assessment with the participants in each academic semester composed an iterative case study. Participants in each case study played E-Rebuild over 12 to 16 gaming sessions and received a mathematical knowledge test before and instantly after the gaming sessions. The test evaluated the application of math knowledge targeted by E-Rebuild and was created and validated by a panel of math educators and measurement experts. The test contains two equivalent forms (for the pre- and posttest) with each form containing seven multistep math word problems (Ke & Clark, 2019). The average Cronbach's alpha of this math knowledge test is 0.68. The results of the external knowledge test would act as a validation of the game-based, learning-in-action assessment strategy.

Participants' gameplay data were collected via the game log (that was developed and iteratively refined during the research) as well as screen video capturing. We conducted a systematic behavior analysis with the video-captured gameplay performance of participants (Ke, 2019). The behavior analysis results, especially on the observable patterns of competency-related task performance, helped to corroborate and inform the development of the game log as part of the measurement model capturing participants' game task performance.

11.4 Results

11.4.1 Patterns and Strategies of Assessing Learning in Action from Gameplay Data

11.4.1.1 Domain Modeling and Learning Game Mechanics Conceptualization

A unique challenge of game-based learning and assessment lies in the modeling or analysis of the targeted competency, its elements, dependencies, and situations for competency development. When analyzing and modeling the targeted math competencies, we experienced two major challenges: (a) the dependencies or relationships among the competency elements within a targeted mathematics practice, whether they are conjunctive or complementary or combined, are not definite; and (b) the granularity level or the depth of a graphical model that maps the targeted math competencies is not definite. The former challenge is related to the inherent complexity of the subject matter. We constructed a graphical model for each target competency (see Fig. 11.1) by consulting expert mathematical educators and referring to the

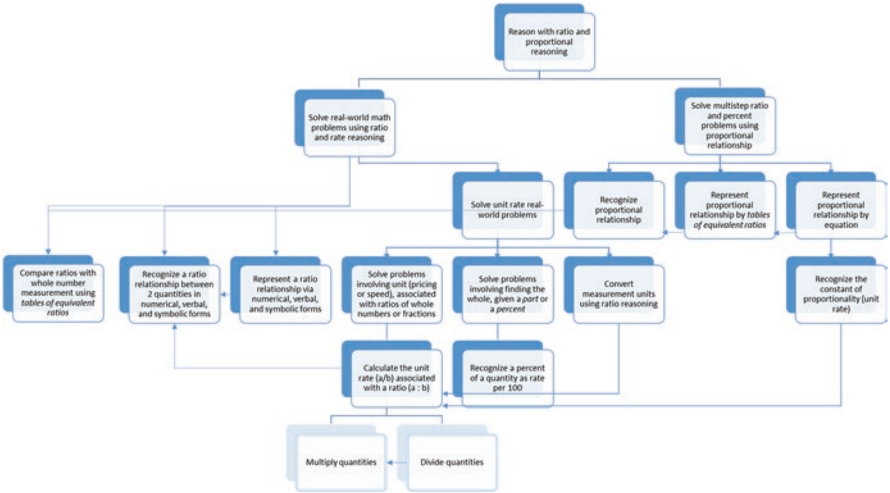


Fig. 11.1 A draft graphical model for domain competency modeling

Common Core State Standards for Mathematical Practices (CCSSI, http://www.corestandards.org/Math/Practice/). The relationships between parallel children nodes (e.g., whether they compose conjunctive or complementary prerequisites) and the parent node, however, are not specific or definite. Content experts and the mathematics literature lack a conclusive perspective on the development process and the substantive structure of the composite proficiency. To address this issue, we iteratively refined the graphical competency model based on (a) our participants' performance with the competency-related tasks during gameplay and (b) the trusted resource that contributed to defining the connections between Common Core State Standards for mathematics (e.g., the Math Common Core State Standards Mapper by UCLA).

In our initial design effort of domain modeling, we had strived for an in-depth, high-granularity, and multilayer modeling of each latent competency. These graphic models were then converted to Bayesian network (BN) models. Later during the assessment implementation stage, we found it challenging to train and calibrate a complicated BN using small data sets. Moreover, in spite of the fact that the process of domain modeling was highly integrative with the design of game mechanics, not all children nodes in the competency model could be effectively represented or integrated into the interactive game-based learning environment as game actions. This made it difficult to extract direct observable variables for each and every children node from the gameplay data. Both the game design literature (e.g., Ke, 2016; Klopfer, Osterweil, & Salen, 2009) and our infield testing findings indicated that the learning game mechanics should depict the most salient (rather than all) components of the targeted competency. In consequence, we simplified the initial domain models to highlight and select only salient children nodes that were aligned with and directly captured by the designed learning game mechanics. The core game actions developed, at the same time, were mapped with the targeted nodes (see Table 11.1). Mapping the core game mechanics with the domain competency model enabled the provision of both game-based learning and learning-in-action assessment. The refined, simplified domain models guided the refinement of the BN models for the learning-in-action assessment implementation.

11.4.1.2 Evidence and Task Modeling: Constructing and Mapping Assessment Models with Game-Based Learning Tasks

Instead of developing measurement models in a linear sequence, in our project, the assessment specifications were reassembled and constructed interactively to better capture game-based learning. Following the integrative development of the domain model and core game mechanics, we proceeded with the evidence and task modeling simultaneously. The task model for E-Rebuild was constructed as a template of game challenges that specified the core parameters of a family of game-based learning tasks. These core game task parameters comprised the key game action and rule set (or the game mechanic), an architectural design scene (e.g., building a shelter

Table 11.1 Part of the tabular model mapping game mechanics with competency nodes

6.G.A.1	Collect	Collect items by identifying task-relevant math properties
6.G.A.1	Paint	Paint target objects based on the design specifications
6.G.A.1	ProtectFloor	Cover the floor to protect it from the elements
6.G.A.2	EmptyInventory	Place/sell all items from inventory
6.G.A.2	FillVolume	Fill a volume with blocks
6.G.A.4	Fold2D	Create folds in correct locations with matching angles and direction
6.G.A.4	Fold3D	Fold object in 3D space to match object displayed
6.RP.A.3	Collect	Collect all items marked collectable
6.RP.A.3	LivingArea	Provide enough floor space for all people
6.RP.A.3	FillArea2D	Use items of the required type and size to partially fill a space
7.G.B.4	EmptyInventory	Place/sell all items from inventory
7.G.B.4	LivingArea	Provide enough floor space for all people
7.G.B.4	Paint	Paint all unpainted objects
7.G.B.5	Angle	Place item with the same angle as the target
7.G.B.6	FillArea2D	Use items of the required type and size to partially fill a space
7.RP.A.3	Collect	Collect all items marked collectable
8.G.A.1	PlaceItems	Place items in the corresponding locations affected by a transform
8.G.A.1	WithinTolerance	Purchase and use the prescribed amount of blocks
8.G.A.2	PlaceItems	Place items in the corresponding locations affected by a transform
8.G.A.3	Angle	Place item with the same angle as the target
8.G.A.3	Distance	Place object in a marked position
8.G.A.3	MinimumAmount	Place at least the required number of objects
8.G.C.9	FillVolume	Fill a volume with blocks

with shopping containers or an adobe house using blocks), the interactive objects and their properties (e.g., the geometric shape, location, area, perimeter, and volume), the resources (e.g., time and material credit), the constraints (e.g., the variables known and unknown and the design objectives to be fulfilled), and the interaction interfaces (e.g., number entry, object maneuvering control, or navigation choice making). Evolving around the game mechanics that were mapped with the targeted competency, each task model was associated with the competencies specified in the domain model. The drafted task models and their parameters were iteratively tested with the project participants and refined based on the observed interactions between the players and exemplary tasks. Our refinement of the task models and parameters, especially with the game actions, rules, and user interaction interfaces, was driven by our observation on whether and how much a task model "necessitated" and extracted competency-related gameplay (Ke et al., 2019). In consequence, we obtained eight task models (or eight task templates) that underlie the instantiation and development of 43 game levels packaged under five game episodes in the current version of E-Rebuild (https://mileresearch.coe.fsu.edu/erebuild/).

As a principal process of evidence modeling, we constructed a graphical measurement model that extends the previous, developed domain model by mapping the children nodes with each task model. This graphical model, however, was later found difficult to be refined or used due to the complexity of the task models and the multi-way relationships between each task model and every children node in the domain model. As an alternative, we then constructed a Q-matrix (Almond, 2010) as a tabular, evidence rule set that relates the children nodes of the domain model to observable evidence in the game tasks. This Q-matrix (Fig. 11.2) specified how various observed variables of major game task model would extract the practice (and evidence) of different competency facet(s) and they would collectively afford the learning and assessment of the targeted competencies. We found the Q-matrix being operative as an assembly model that enables game designers and assessment experts to (a) estimate whether, when, and what tasks generated will accumulate enough evidence and (b) gauge the difficulty, discrimination quality, and hence the balance and sequence of the generated/instantiated tasks across game task models and episodes.

11.4.1.3 Data Capturing, Performance Pattern Recognition, and Observable (Evidence) Extraction During Game Log Design

We found the game log design to be the most iteratively refined process of evidence modeling. Game logs are XML files that has elements with tags like *root*, *Name*, *Level*, and *Time* which are enclosed within "<" and ">." E-Rebuild examined in this chapter has 43 game levels, and each level has its own set of observables. Table 11.2 lists a few game levels along with their observables. The last entry of Table 11.2 is the union of the observables from all the game levels. Each game level logs only a subset of the total observables.

Episode	Level	Comprehend a ratio relationship via numerical, verbal, and symbolic forms	Solve problems involving finding the whole, given a part or a percent	Calculate the unit rate (a/b) associated with a ratio (a : b)	(De)compose quadrilaterals and polygons into (right) triangles and rectangles	Compute the area and perimeter of triangle and rectangle	Find surface areas of 3D figures using nets of rectangles and (right) triangles	Compute the volume of right rectangular prisms (V = l w h, V = b h)	Compute area and circumference of a circle using formulas
Island	Collect Container 1	0	0	0	0	1	0	0	0
Island	Collect Family 1	1	0	1	0	0	0	0	0
Island	Build Training	0	0	0	0	0	0	0	0
Island	Fill Training	1	0	0	0	1	0	0	0
Island	Build 1	1	0	1	1	1	0	0	0
Island	Fill 1	1	0	1	0	1	0	0	0
Island	Collect Container 2	0	0	0	0	1	0	0	0
Island	Collect Family 2	1	0	1	0	0	0	0	0
Island	Build 2	1	0	1	1	1	0	0	0
Island	Fill 2	1	0	1	0	1	0	0	0
Island	Collect Container 3	0	0	0	0	1	0	0	0
Island	Collect Family 3	1	0	1	0	0	0	0	0
Island	Build 3	1	0	1	1	1	0	0	0
Island	Fill 3	1	0	1	0	1	0	0	0
Desert	Copy Training 1	0	1	1	1	1	0	0	0
Desert	Copy Training 2	0	1	1	1	1	0	0	0
Desert	Placement Training 1	0	0	0	0	0	0	0	0
Desert	Build 1	1	1	1	1	0	1	0	0
Desert	Angle Build 1	1	1	1	1	0	1	0	0
Desert	Location Build 1	1	1	1	1	0	1	0	0
Desert	Fill 1	1	0	1	0	1	0	0	0
School	Place 1	1	1	1	0	1	0	0	0
School	Place 2	0	0	0	0	0	0	0	1
School	Fill1	1	0	1	0	1	0	0	0
School	Paint 1	0	1	1	0	0	0	0	1
School	Stadium 1	0	1	1	1	0	0	0	0
School	Stadium 2	0	1	1	1	0	0	0	0
School	Paint 2	0	1	1	1	0	0	0	0
Farm	Angle 1	0	1	1	0	0	0	0	0
Farm	Perimeter 1	0	1	1	0	1	0	0	0
Farm	Area 1	0	1	1	0	1	0	0	0
Farm	Volume 1	0	1	1	0	0	0	1	0
		16	14	25	11	17	3	1	2

Fig. 11.2 Part of a tabular, assembly model (Q-matrix)

Table 11.2 Sets of observables for exemplary game levels (Ke et al., 2019)

Exemplary level	Observables
21ContainerCollect	Time, NumWrong, MaterialCredits
26FamilyPlacement	Time, NumWrong, AssignmentComplete, MaterialCredits
SchoolAssignment01	Time, AssignmentComplete, NumAssignments, NumFailedAssignments, Num-FamilyCollected, LevelComplete
IslandBuild02	Time, NumBlocks, NumTrades, Total-Lost, MaterialCredits, Distance, Size, Angle, BuildingComplete, LevelComplete
All levels	Angle, AssignmentComplete, BuildingComplete, Distance, LevelComplete, MaterialCredits, NumAssignments, NumBlocks, NumFailedAssignments, NumFamilyCollected, NumTrades, NumWrong, Size, Time, TotalLost

The development and selection of observable variables reflected the initial hypotheses of the domain and task modeling, as well as the result of the behavioral analysis with the actual gameplay performance of the project participants. We did a systematic coding with participants' gameplay behaviors, by identifying the salient behavioral patterns and their occurrence frequencies, the type and sequence of variant behaviors related to the task performance and math knowledge application, and the contexts of the task engagement (see more details in Ke, 2019). The behavioral coding, as a form of the performance pattern recognition and dimension reduction (Mislevy et al., 2012), assisted the extraction and selection of salient observables from the gameplay data for learning evidence capturing and accumulation. It also addressed the issue of overloading the game log with nonessential observables, thus facilitating the later data/evidence processing.

Data Processing: Categorize Raw Observables

We consolidated the participants' gameplay data in multiple XML game logs along with the corresponding external knowledge test scores into a single CSV format data sheet. Each entry or row in the output CSV file corresponded to a particular user and his/her game log for a particular game level played at a certain time. All raw values in the consolidated data file were converted to categorical values, namely, low, medium, and high. Having only three categories greatly reduced the complexity of a prediction model. We defined a distinct categorization rule for each observable based on a pair of thresholds. The category thresholds can be determined by an expert or computed from the training data itself. We took an integrative approach that mixed these two strategies. Let's assume an observable variable x has a set of observed values S_x. We first iteratively removed the first two values in S_x which are farthest from the mean of S_x in the hopes of retaining an outlier-free set of values. We then computed the mean μ and the standard deviation σ of the remaining values in S_x. Experts of the E-Rebuild game and learning assessment then validated the computed μ and σ. Finally, the two thresholds are computed as:

$$t_x^1 = \mu - \sigma$$

$$t_x^2 = \mu + \sigma$$

After we computed the two thresholds t_x^1, t_x^2 for the variable x such that $t_x^1 < t_x^2$, a categorization rule $f_x(v)$ mapped an input value v to one of the three categories *{Low, Medium, High}* as follows:

$$f_x(v) = \begin{cases} \text{Low, if } v < t_x^1 \\ \text{Medium, if } t_x^1 \le v < t_x^2 \\ \text{High, if } t_x^2 \le v \end{cases}$$

11.4.1.4 Training and Calibrating the Statistical Evidence Model

We used Netica to create a Bayesian network (Fig. 11.3) and conducted experiments on the preprocessed categorical data set. The goal was to learn a model based on the training set about the relationships between observables across all gameplay and the

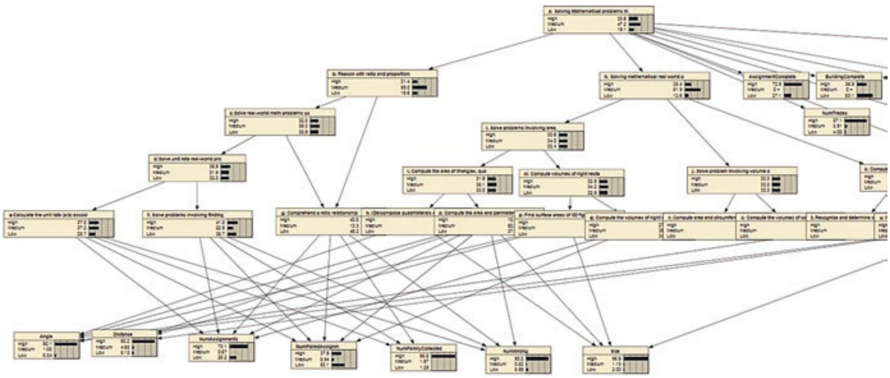


Fig. 11.3 Part of a Bayesian network model

users' competencies. Here are two reasons why we chose Bayesian network for the in-game learning assessment:

- It is flexible in allowing domain experts to encode the domain knowledge by defining nodes (i.e., competency and observable variables) and edges (i.e., the relationships among the variables) in a directed acyclic graph.
- It handles missing data. This feature is a requirement in our case because each game level provides only a subset of the observables.

Prepare Model Input and Pretest

We fed the training data set into the Bayesian network via the Netica case file format. To establish initial values for the conditional probability tables in our Bayesian network models, we considered each game level played as an individual case. The initial values for the students' mathematic ability were taken from their external math test results. Initially the values for the student's mathematic ability were taken directly from the pretest. The network was trained using these values along with the game log data. The network was then tested against the posttest data. Our initial results were subpar, with an approximately 70% error. To combat this, we tried using a weighted average and treating the number of logs as a measure of time. We had two values coming from the pre-gaming and the post-gaming tests. Possible solutions included picking the maximum or simply taking the average of the two. Instead, we tried to capture the progress of a dynamic learner as they played the game levels in a chronological sequence. We assumed that students improved their math skills after every game level and transitioned from the competency they showed in the pre-game tests to their final competency shown in the post-game test. So, if a student with m_{pre}, m_{post} scores in the pre- and post-gaming tests played n

game levels $\{l_1, l_2, \ldots l_n\}$ in sequence, the eventual math score m_i for level l_i was computed as:

$$m_i = \left(1 - \frac{i}{n}\right) \times m_{\text{pre}} + \frac{i}{n} \times m_{\text{post}}$$

Training and Calibration

We used a part of students' data to learn the math competency prediction model and used the rest to test its validity. For this, we first split the data set into the training and test sets based on the students. We varied the number of students in the training set to see how the number of training examples affected the prediction accuracy. We selected three numbers at random for training and kept the rest for testing. In this way, we ran three different experiments on this student population. We further validated the model trained on the full population data with the full data for an independent population.

Given a training data set, we used expectation maximization to train the Bayesian network model in Netica. After the training was complete, we fed the test data to the model. For each input test case, the trained model emitted probabilities for each node. We picked the category that has the highest marginal probability as the predicted output.

11.4.2 Validity of BN-Based Game-Based Learning Assessment and Alternative Data Mining Methods

We conducted an association analysis between the in-game assessment results (i.e., Bayesian network predictions for individual learners) and the external post-gaming math test results to validate the Bayesian network-based assessment mechanism. The correlation analysis was conducted to examine the consistency between the predicted result of the current Bayesian network model (i.e., low, medium, or high in the targeted competency) and the categorized posttest performance of the middle school participants. The analysis result indicated a significant association between the BN prediction results and the external test results, $r = 0.40$, $p = 0.02$.

We have also examined the potential of using a package of other educational data mining methods to analyze the gameplay data, including random forest, *k*-nearest neighbors classification (KNN), support vector machine (SVM), and Gaussian

Table 11.3 Prediction correctness scores of alternative data mining methods (gathered over 10 random runs)

Random forest	KNN	SVM	Gaussian Bayes
0.34	0.28	0.32	0.24

Bayes. Their prediction correctness scores are presented in Table 11.3. It appears that the random forest algorithm can be another promising statistical method for the game-based, learning-in-action assessment.

11.5 Conclusions and Discussion

Our design-based research outputs confirm the feasibility and validity of using and integrating the methods of educational data mining with the evidence-centered design framework in assessing game-based learning in action. The results suggested that the processes of domain or competency analysis, evidence modeling, task modeling, and data capturing/processing for assessment implementation are highly integrative and interactive rather than linear or sequential. As observed, the design and implementation conjectures of the learning-in-action assessment, similar to those for the design of a digital learning environment, should be iteratively tested and refined during infield testing. Through infield testing a package of data capturing, processing, modeling, and validation strategies, we hope to present functional and data-driven conjectures governing the design and implementation of the game-based, domain-specific learning assessment. Specifically, the game-based, learning-in-action assessment evolves around four major operational practices: (a) domain competency modeling along with core game mechanics conceptualization; (b) developing task models and the Q-matrix; (c) developing the game log that encompasses performance data capturing, pattern recognition, and observables extraction; and (d) training, substantiating, and comparing statistical models for data processing and assessment implementation. This emerged operational framework is not meant to be prescriptive, but works more as an illustrative case of the evidence-centered assessment design along with educational data mining in the game-based, domain-specific learning setting.

Multiple assessment design and implementation issues have emerged during this current study and warrant further investigation in the future research. *First*, an exploratory approach in identifying the evidence (or observable) of learning in action, as well as the training and calibration of the statistical model of assessment, is in need of a longitudinal data set collected from a large, heterogeneous learner population. Such a need makes the development and validation of the game-based learning, especially domain-specific game-based learning, cost intensive and time-consuming. Future research should focus on exploring the data mining methods and/or statistical models that can cope with a small data set or a homogenous sample issue. *Second*, future researchers and practitioners should examine and control the moderating effect of a student's gameplay skills (developed across game levels) in the student's model when predicting the student's domain competency development. A conjecture is that assigning a specific and variant weight of evidence to a student's initial-level performance and to his later one can balance or skew the prediction of his/her proficiency state. In this study, we have considered the task performance of each level as a separate data record, due to a small sample size. The risk of such a strategy in skewing the prediction result should be further investigated and compared with a strategy that treats the congregated, multilevel task performance of a single student as a data record. *Third*, E-Rebuild is more a game-based learning environment than a game-based assessment tool. An assumption of E-Rebuild is that students will develop proficiency across game levels or tasks. This assumption is in conflict with the assumption of a computerized assessment tool that one's proficiency state remains consistent across assessment items. In consequence, it remains a question whether initiating the conditional probability tables for the competency nodes in the Bayesian network using only the pretest data would skew the prediction on a student's proficiency state during game-based learning. Our strategy of capturing the progress of a dynamic learner as they play the game levels in a chronological sequence—using a weighted average of the pre- and posttest data while treating the number of logs as a measure of time—should be replicated and examined in different digital learning settings. *Last but not least*, specifying a game-based learning task model, like that in E-Rebuild, involves defining not only the targeted competency facets but also their observable variables and inter-variable relationships in an ill-structured problem. Therefore, specifying the difficulty index of each instantiated task involves the estimation and combination of multiple facets, such as the task's psychometric property, the task structure, and the gaming skill required. Moreover, the task's parameters may interact with individual students' characteristics and proficiency states, making the difficulty index of each task a dynamic and individualized parameter. Addressing this dynamic difficulty index issue is important for the sequencing and adaptive presentation of game-based tasks for individual learners and should be a focus of the game-based learning and assessment research.

Acknowledgements The work reported in this chapter was supported by the National Science Foundation, grant no. *1318784.* Any opinions, findings, and conclusions or recommendations expressed in these materials are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Almond, R. G. (2010). I can name that Bayesian network in two matrixes! *International Journal of Approximate Reasoning, 51*(2), 167–178.

Ayers, E., & Junker, B. W. (2006, July). Do skills combine additively to predict task difficulty in eighth grade mathematics. In *Educational data mining: Papers from the AAAI workshop*. Menlo Park, CA: AAAI.

Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies, 23*(1), 537–553.

Carnoy, M., & Rothstein, R. (2013). *What do international tests really show us about U.S. student performance?* Washington, DC: Economic Policy Institute.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research, 86*(1), 79–122.

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.ets.org/Media/Research/pdf/session4-dede-paper-tea2012.pdf

Dewey, J. (1910). Science as subject-matter and as method. *Science, 31*, 121–127.

Eriksson, I., & Lindberg, V. (2016). Enriching 'learning activity' with 'epistemic practices'–enhancing students' epistemic agency and authority. *Nordic Journal of Studies in Educational Policy, 2016*(1), 32432.

Kang, J., Liu, M., & Qu, W. (2017). Using gameplay data to examine learning behavior patterns in a serious game. *Computers in Human Behavior, 72*, 757–770.

Ke, F. (2016). Designing and integrating purposeful learning in game play: A systematic review. *Educational Technology Research and Development, 64*(2), 219–244.

Ke, F. (2019). Mathematical problem solving and learning in an architecture-themed epistemic game. *Educational Technology Research and Development*. https://doi.org/10.1007/s11423-018-09643-2

Ke, F., & Clark, K. M. (2019). Game-based multimodal representations and mathematical problem solving. *International Journal of Science and Mathematics Education*. https://doi.org/10.1007/s10763-018-9938-3

Ke, F., Shute, V., Clark, K., & Erlebacher, G. (2019). *Interdisciplinary design of the game-based learning platform: A phenomenological examination of the integrative design of game, learning, and assessment*. New York, NY: Springer.

Klopfer, E., Osterweil, S., & Salen, K. (2009). *Moving learning games forward*. Cambridge, MA: The Education Arcade. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.687.5017&rep=rep1&type=pdf

Levy, R. (2014). *Dynamic Bayesian network modeling of game based diagnostic assessments* (CRESST Report 837). National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved from https://files.eric.ed.gov/fulltext/ED555714.pdf

Manjarres, A. V., Sandoval, L. G. M., & Suárez, M. S. (2018). Data mining techniques applied in educational environments: Literature review. *Digital Education Review, 33*, 235–266.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 2003*(1), i–29.

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., Haertel, G., Riconscente, M., Rutstein, D. W., & Ziker, C. (2017). Evidence-centered assessment design. In *Assessing model-based reasoning using evidence-centered design* (pp. 19–24). Cham, Switzerland: Springer.

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society, 17*(4), 49–64.

Pardos, Z., Heffernan, N., Ruiz, C., & Beck, J. (2008, June). The composition effect: Conjunctive or compensatory? An analysis of multi-skill math questions in ITS. In R. de Baker, T. Barnes, & J. E. Beck (Eds.), *Educational data mining 2008 proceedings* (pp. 147–156).

Sandoval, W. A., & Bell, P. (2004). Design-based research methods for studying learning in context: Introduction. *Educational Psychologist, 39*(4), 199–201.

Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., … Mislevy, R. (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., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. Cambridge, MA: MIT.

Taub, M., Azevedo, R., Bradbury, A. E., Millar, G. C., & Lester, J. (2018). Using sequence mining to reveal the efficiency in scientific reasoning during STEM learning with a game-based learning environment. *Learning and Instruction, 54*, 93–103.

Yin, R. K. (2013). Validity and generalization in future case study evaluations. *Evaluation, 19*(3), 321–332.

Chapter 12
Bridging Two Worlds: Principled Game-Based Assessment in Industry for Playful Learning at Scale

V. Elizabeth Owen and Diana Hughes

12.1 Introduction

In recent years, a large body of research in game-based assessment has emerged (e.g., Baker, Chung, & Delacruz, 2012; Ifenthaler et al., 2012; Mislevy et al., 2014) focused on games as immersive, complex environments in which active engagement and player interaction fuel learning progression (e.g., Gee, 2003; Shute, 2011). Games are designed experiences that can provide immersive contexts for supporting self-regulated learning and higher-order thinking skills (Rieber, 1996; Squire, 2006; Steinkuehler & Duncan, 2008). Comprised of well-ordered problems providing just-in-time information to support player progress, well-designed games provide formative feedback within cycles of appropriately challenging play (Gee, 2005). Indeed, good games effectively harness formative assessment to foster ongoing feedback cycles and customized player difficulty levels (Shute & Kim, 2014). In order to maintain this immersive context for learning, games often consist of ongoing assessment balanced with engaging mechanics and narrative (cf. Squire, 2011).

Thus, games as engaging, interaction-driven systems can be natural vehicles for assessment in an authentic context (Gee, 2012). Methods like evidence-centered design (Mislevy, Almond, & Lukas, 2003) can help to structure this assessment in the context of educational games, with emphasis on aligning target knowledge, skills, or abilities (KSAs) with desired evidence and in-game assessment tasks designed to elicit such data. ECD can be applied in the context of simulations and games (Mislevy, 2011) to support design and measurement of complex competencies in immersive contexts (e.g., Clarke-Midura, Code, Dede, Mayrath, & Zap, 2012; DiCerbo et al., 2015). These well-structured game environments can also provide rich interaction data streams afforded by digital platforms, integrating evidence-based design structures into a comprehensive, context-rich event stream—thus

V. E. Owen (✉) · D. Hughes
Age of Learning, Los Angeles, CA, USA
e-mail: diana.hughes@aofl.com

D. Ifenthaler, Y. J. Kim (eds.), *Game-Based Assessment Revisited*, Advances in Game-Based Learning, https://doi.org/10.1007/978-3-030-15569-8_12

enabling a broad range of analyses, including investigation of emergent learner patterns afforded through disciplines like educational data mining (EDM; Baker & Yacef, 2009). Though seemingly very different approaches, recent scholarship establishes ways in which ECD and more exploratory analysis methods (e.g., consistent with EDM) can be compatible and even complimentary, particularly with emergent student patterns informing iterative refinement of evidence models and core assessment mechanics (DiCerbo et al., 2015; Mislevy, Behrens, DiCerbo, & Levy, 2012). The established integration of these approaches can be applied in an educational games context to enable "data mining" and "feed insights back into improved design" (Mislevy et al., 2014, p. 59). Indeed, empirical research in ECD-based educational games has captured emergent, EDM-based insights with implications for iterative design (e.g., Shute et al., 2015; Slater, Bowers, Kai, & Shute, 2017; Stephenson, Baker, & Corrigan, 2014). These established interdisciplinary principles (with foundations of evidence-based core design enriched by insights from a broad, EDM-enabled data stream) applied to assessive game design will be referred to as integrated game-based assessment (iGBA) for reference throughout the chapter. This kind of integrated approach enables a data-driven system, with formative assessment for just-in-time feedback, student-responsive learning pathways, and deep insights into emergent learning patterns for data-driven design and intelligent personalization.

Application of these principles to large-scale production of learning games can be vital to expanding the benefits of iGBA in practice for impact on playful, engaged learning at scale. Practices that support this application to game industry are therefore important, necessitating the development of tools and processes that are viable and efficient and work within industry-based production paradigms. Challenges include implementing principled learning design in very short development timelines, incorporating importance of production values and polish to support user adoption and financial sustainability, creating research-based tools accessible to nonacademic game developers, and effectively integrating these tools into industry-standard development practice.

In addressing these challenges, this chapter offers an example of a working iGBA practice in an industry context. Using this integrated approach, game-based assessment principles are embedded in the design and production processes at *Age of Learning*—specifically, in a personalized game-based learning system called *Mastering Math* (*MM*), designed to help young learners build a strong understanding of fundamental math concepts. In this game production environment, the implemented iGBA practices are lean enough for fast iteration and practical use by game designers and developers, with a culture of rich, structured data built in from the first considerations of game design and development. This detailed data stream, in turn, allows for iterative, data-driven design for improved learning and engagement. A range of methods can thus be employed for analysis, including evaluation of designed assessment, as well as broader explorations with learning analytics and educational data mining (LA/EDM; Baker & Siemens, 2014) to enhance intelligent system response in real time. These rich, iGBA-based data also have implications outside the system, as event-stream insights can be surfaced to teachers to allow classroom-based intervention and student support. *MM* integrates well into formal learning environments, as recent classroom-based research suggests, providing promising learning results with *MM* and informing future development of the system.

The following pages detail *Mastering Math* as an industry-produced system based in iGBA principles, designed as a data-driven, immersive learning experience positioned for educational impact on students at scale. Theoretical foundations are discussed, as well as implementation of iGBA through viable practices in an industry setting to produce *MM*. Enabled by an iGBA approach, *MM* as a data-driven system is covered, with examples of data-enabled insights for iterative core design and player-responsive pathways. Application to classroom environments is then discussed, as iGBA-based systems like *MM* can also surface salient student progress information to teachers for in-person intervention. In a school-based context, we review recent empirical research with *MM*, as well as implications of the system for game-based playful learning impact at scale.

12.2 Games, Learning, and Assessment: A Review

Well-designed games provide roles, goals, and agency (Squire, 2011) in a series of well-designed problems, providing just-in-time information and formative feedback as players progress (Gee, 2005). As this kind of dynamic, player-responsive environments, games have the power to offer a meaningful context for learning and assessment—enabling an interactive learning environment where "structure and motivation are optimized without subverting personal discovery" (Rieber, 1996, p. 44). Thus fueled by interaction, good games can be seen as natural formative assessment vehicles in which player input drives system response in an authentic, immersive learning context (e.g., Plass, Homer, Kinzer, & Perlin, 2012; Shute, 2011). Indeed, this chapter focuses on games as formative assessment environments, in which students are assessed by the system, receive just-in-time feedback, and may learn or change behaviors as a result, which is then picked up by the game for the next round of formative feedback (e.g., Ke, Shute, Clark, & Erlebacher, 2019). This kind of assessment provides "formative information whenever possible (i.e., give useful feedback during the learning process instead of a single judgment at the end); and…has as its primary goal improvement of learning" and is "assessment *for* learning, in contrast to 'summative assessment' (or assessment *of* learning)" (Shute & Kim, 2014, p. 311).

To this end, in order to sustain these engaging, player-responsive environments, games need to react to player performance on core game mechanics from moment to moment. In an educational context, this translates into fundamental game interaction generating learning performance data as players move through the system. Generating quality in-game learning evidence can be supported through early consideration of key design factors, such as *what* learning goals need to be assessed, *how* they will be assessed through game interaction design, and what *evidence* these designed interactions will provide. These elements have been synthesized in evidence-centered design (ECD), a canonical approach to evidence-based design in educational contexts (Mislevy et al., 2003). ECD is an assessment framework which "enables the estimation of students' competency levels and further provides evidence supporting claims" about the target competency (Shute, 2011, p. 508). Elements in the ECD process cover research on what to assess (domain analysis and modeling), the design of the core assessment structures (Conceptual Assessment Framework, CAF), and implementation (assessment implementation and delivery) (Fig. 12.1). The CAF in particular has been focused on as a central part of assessment design, which aligns elements of competency, evidence, and designed tasks (Shute, 2011).

ECD can be applied in the context of simulations and games (e.g., Mislevy, 2011) to support design and measurement of complex competencies in immersive contexts (e.g., Clarke-Midura et al., 2012; DiCerbo et al., 2015; Shute, Wang, Greiff, Zhao, & Moore, 2016). Recent efforts include deeper research into ECD specifically for games, with consideration for viability of implementation in game development. For example, researchers working with GlassLab developed a paradigm for ECgD (evidence-centered game design), an extension of ECD optimized specifically for embedding assessment in games (Mislevy et al., 2014). These efforts were applied to game-based assessment practices in GlassLab's game development lab for projects like S*imCityEDU*. In related work, MIT researchers and game designers developed a simplified ECD-based approach to designing learning games (Groff,

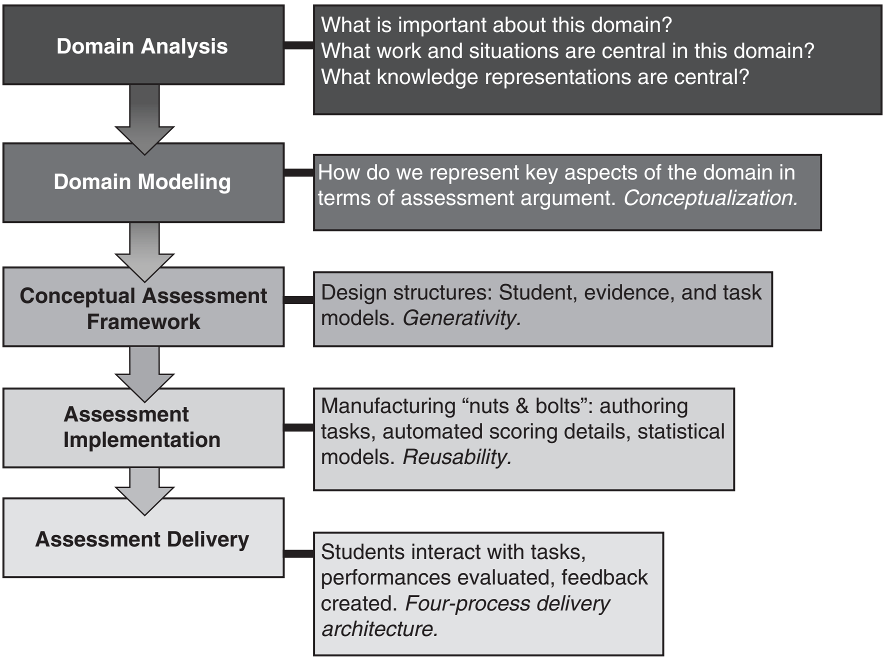


Fig. 12.1 A full-scale ECD model (Mislevy, 2011)

Clarke-Midura, Owen, Rosenheck, & Beall, 2015). This evidence-based approach was called "balanced design" and was written for an audience of game developers—with the purpose of presenting a viable learning design approach to be integrated with current best practices in game design. Figure 12.2 shows a central evidence-based approach (aligning competency, evidence, and task design) in template form for game designers and populated with sample design.

Building on the foundation of a strong central design, these games now have the digital affordance of generating large data streams (cf. DiCerbo, 2014; Hao, Smith, Mislevy, von Davier, & Bauer, 2016), which can be leveraged to better understand learner patterns for iterative core design and refined formative feedback (Shute & Kim, 2014). Detailed event-stream data can enable analysis of designed assessment as well as support insight into emergent learning patterns (e.g., DiCerbo et al., 2015; Plass et al., 2013). Recent research asserts this data can be leveraged to enhance ECD-based design using EDM and exploratory methods (DiCerbo et al., 2015), particularly for feature generation and feature selection informing iterative design of evidence and task models (Mislevy et al., 2012). These emergent feature patterns obtained through "data mining" can "feed insights back into improved design" in ECD-based systems (Mislevy et al., 2014, p. 59).

This is particularly helpful in the medium of games, which by nature can contain rich narratives and immersive interactions that work in tandem with core assessment mechanics. Indeed, games as a medium encourage the discovery of an underlying rule system through boundary testing and subversive play (Aarseth, 2007; Salen & Zimmerman, 2004). Failure-based formative feedback as a mechanism often drive discovery in well-designed games (Juul, 2013), in which unexpected player pathways and productive failure can support learning performance (e.g., Bielaczyc &

Content Model	Task Model		Evidence Model	
Learning Objective	**Quest**	**Task/Action**	**Data Collected**	**Interpreting Evidence**
Recognize patterns in data sets	ST1.1	Turn in data summary to support/refute government claim	Data summary (see Table 1.7 for possible data summary submissions)	*Correct:* Player knows to use a large enough sample size and the correct measure. *If Incorrect:* *Species other than blackburn:* Player likely does not know what a blackburn is *Trait other than body length:* Player did not understand what needed to be measured *If result does not agree with their answer in Tumbler 1.2:* See Quest 1 7a *If sample size is too small:* Player does not understand the importance of a large enough sample
Use models and simulations to make inferences and conclusions	EV3.3	Students use a simulator to see how environmental pressures can affect trials.	Students turn in EvoGlobe and respond to questions. Data collected includes Globe setting and responses.	*Correct:* See EvoGlobe settings EV3.3a table. *Potential reasons for Incorrect EvoGlobe:* 1. They don't understand principle of natural selection. 2. They don't understand how to interpret EvoGlobe. 3. They don't understand how to use the EvoGlobe.

Fig. 12.2 A sample evidence-based design template for game designers, core to the "balanced design" approach (Groff et al., 2015, p. 10)

Kapur, 2010; Owen, Anton, & Baker, 2016). Rich student interaction data can thus support a broad range of analyses critical to understanding learning in the context of educational games. These large digital event streams enable the application of methods tailored to high-volume educational data, such as learning analytics and educational data mining (LA/EDM; Baker & Siemens, 2014). These approaches empower the use of large educational data streams to mine organic learner patterns related to elements like student performance, affect, and behavior (Baker & Yacef, 2009). Recent game-based research has applied these kinds of methods, building on ECD game design foundations to leverage event-stream data for insight into emergent player patterns linked to learning performance (Baker & Clarke-Midura, 2013; (Martinez-Garza & Clark, 2017), noncognitive skills (e.g., Shute et al., 2015, and desirable student behaviors (e.g., Kerr & Chung, 2012; Sweet & Rupp, 2012). In a wide range of other game-based research (e.g., Owen & Baker, 2019), LA/EDM analyses have been used to uncover emergent learner patterns related to elements like strategy (e.g., Asbell-Clarke, Rowe, & Sylvan, 2013; DiCerbo & Kidwai, 2013), student attrition (Hicks et al., 2016; Ramirez, 2016), player profiles (Canossa, Badler, El-Nasr, Tignor, & Colvin, 2015; Slater et al., 2017), and learner affect (Kai et al., 2015; Rodrigo & Baker, 2011). Rich event-stream data in game environments enables such investigations, which can inform potent data-driven design and personalized formative feedback (e.g., Ke et al., 2019).

To enable these kinds of powerful insights about playful learning in well-designed learning games, the design of data frameworks for event-stream collection is key (e.g., Owen, Wills, & Halverson, 2012). Architecture of these data streams can serve as a synthesis point for designed assessment mechanics (e.g., from a game-based task model) and resultant evidence, while capturing a context-rich data stream of all player interactions (allowing EDM-based investigations into emergent student patterns). Ideally, a well-designed game will have game events that can be interpreted directly in terms of the types of competencies and learning that the designer wants to measure (Shute & Kim, 2014). The process of this design produces data that can also be interpreted in consideration of other interactions and features, particularly in the context of larger event-stream log files (Owen & Baker, 2018). Within recent efforts in structuring serious game data (cf. Chung, 2015; Hao et al., 2016), a strong framework for game data capture provides several key functions: comprehensive data, clear organization (e.g., alignment with core assessment/progress mechanics), and consistency in nomenclature of events throughout the game. This also allows for interpretable data models, which can directly inform data-driven design of the game itself—i.e., when data is interpretable and design-aligned, outcomes of analysis can be more easily translated to direct feedback into design. In practice, this can take the form of several core types of recorded events. Specifically, *player actions* and *system feedback* are captured as part of a high-resolution data framework (e.g., Danielak, 2014); additionally, strong data frameworks effectively mark player *progression* (e.g., the user's current level, quest, or objective) within the scope of the full game. Each core event type may then then associated with important *context* (e.g., timestamp and player ID) as well as the *result* of the player action (e.g., win/lose, score, or more fine-grained outcomes) when appropriate (cf. Serrano-Laguna et al., 2017; Stenerson, Salmon, Berland, & Squire, 2014).

An example from recent research, the ADAGE data framework (Assessment Data Aggregator for Game Environments; Halverson & Owen, 2014) offers one representation of how these core events can be structured. ADAGE calls *progression* elements "units," or repeating progress mechanics.[1] Built to be flexible, units don't have to be nested or linear and record a designed assessment *result* or key performance outcome when applicable. In labeling events directly in relationship to designed core mechanics, the unit structure helps data be interpretable and clearly aligned with design. This structure also allows for comprehensive data collection. In the context of each unit and with complete data on user ID, timestamp, session, and other standard contextual data, *player action* and *system events* are recorded as well. Each tap and drag of users are recorded (player action), and system events represent the game's feedback to the player as a result of action (or inaction)—e.g., inactivity prompts, tutorial pop-ups, increased in-game scaffolding, etc. Using this kind of structured framework across games also allows for consistency in data labels, key to interpretability of data for analysis across broad systems of games. Designed to be flexible, this kind of framework can be applied across a wide range of games and simulations and allows for implementors to customize events as needed to fit the context. Below is a simplified representation of an ADAGE-based data framework (Fig. 12.3).

12.2.1 Integrated GBA and Data-Driven Systems

In following this trajectory of playful learning research, key elements can be incorporated into an integrated approach to game-based assessment—grounded in tenets of ECD for principled embedded assessment design, which serve as touchstones in broader, comprehensive event-stream data collection (consistent with principles of EDM) to enable a wide range of learning insights. This integrated GBA approach informs and enables data-driven game-based learning systems.

First, with integrated GBA as part of development from inception, game design is founded in the consideration of learning evidence, in which designed assessment becomes seamlessly embedded into core verbs of play (cf. Corrigan et al., 2015; Grace, 2014; Zagal, Mateas, Fernández-Vara, Hochhalter, & Lichti, 2005). This evidence-based design then allows for formative feedback in response to learning performance (from interaction with designed assessments) throughout play. This just-in-time player feedback, already a compelling affordance in good games (Gee, 2005), can thus be used in service of game progress *and* learning assessment, which have been seamlessly integrated in evidence-based design foundations. Learning

[1] Units can be big or small, and each game can have multiple units. For example, in World of Warcraft, a unit of progress would be quests; in Words with Friends, it might be a turn or the unit of a game itself.

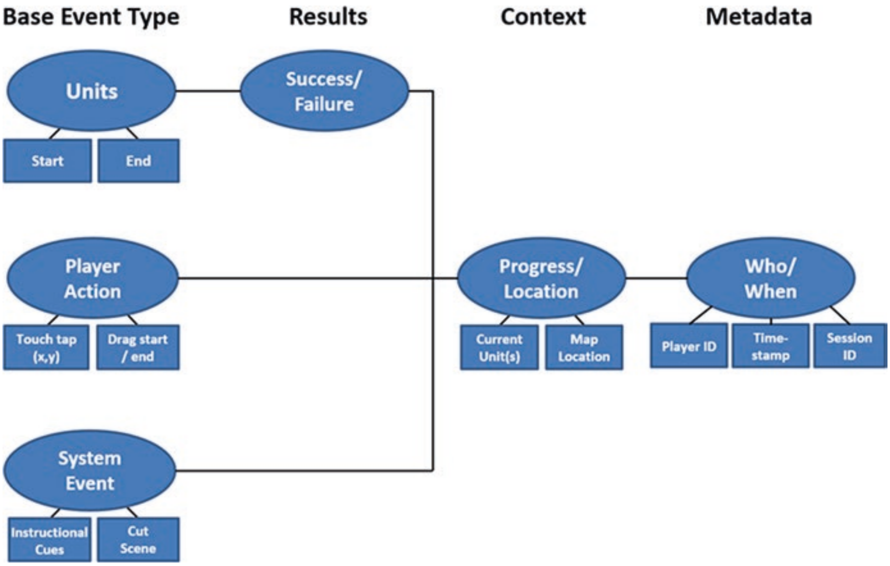


Fig. 12.3 A simplified representation of an ADAGE-based data framework

performance can then be gauged throughout play progression, providing the ability to give tailored feedback and dynamically adjusting personalized pathways.

As assessment and progress mechanics inform event-stream data structures, the resultant data is clearly labeled in alignment with design. This labeling can then be held consistent across games (e.g., Fig. 12.3), while collecting a comprehensive range of player interactions and system feedback events. This systems-view approach to data collection makes the event stream more scalable—both in terms of games' scope (providing consistently labeled data across multiple games) and player volume (with clear and consistent structure throughout large data streams). A result of iGBA in practice, this data can then be used to support ongoing analysis for iterative design, honing the game experience to support learning and engagement. In addition to supporting direct tweaks to game design, the event-stream data can fuel emergent insights into player patterns, enabling EDM-based methods to provide behavior detection and predictive modeling for more intelligent personalization and adaptive game overlays.

iGBA can thus enable data-driven game systems, in which ongoing formative assessment supports feedback and personalization. The aligned event-stream data can then effectively close the loop of data-informed iterative design—both in terms of directly honing core game mechanics and leveraging LA/EDM methods to provide adaptive overlays of event-stream prediction and behavior detection for more intelligent, player-responsive personalization. This personalization and iterative design, in turn, can support engagement and learning for strong educational impact through immersive play.

12.2.2 Integrated GBA in Practice

Bringing an iGBA approach into more widespread practice can help scale the potential learning impact unlocked by principled design of immersive educational games. One related sector poised to reach youth at scale, for example, is that of commercial game companies. High-polish, commercial-grade game experiences are highly sought after in today's world of booming mobile-based game use. In the context of learning games, this is especially relevant in reaching students in both formal and informal learning environments (e.g., voluntary use outside of school)—which means competition, engagement-wise, with top-shelf recreational games and apps. Paired with iGBA and principled game design, commercial game development with a broad user base thus has the potential for high impact through premium playful learning experiences. However, bringing theory into large-scale industry practice comes with challenges. Research-based design principles must be formalized into a set of tools and processes—all of which need to be efficient to use in a fast-paced production environment and accessible to nonacademic development team members, while still preserving tenets of iGBA. These tools and processes, moreover, need to fit into production practices as defined by industry-standard software development paradigms (e.g., Agile development[2]).

One example of this kind of production company is Age of Learning[3] (AofL), with a mission to make high-quality, engaging play experiences for a broad user base—but with a unique focus on educating young learners. In recent R&D efforts, AofL has taken a research-based approach to creating a personalized, game-based learning system. This system, *Mastering Math*, serves as one example of a data-driven system which puts principles of integrated GBA into industry-based production process—with potential to bring the impact of principled, research-based design into practice to reach young learners at scale.[4] The following pages discuss the methods and challenges of implementing iGBA into industry practice, as well as empirical research exploring the potential for impact of *MM* as a game-based, data-driven learning system.

12.3 *Mastering Math*: System Design and iGBA in Production Practice

Age of Learning's *Mastering Math* is a game-based adaptive learning system that helps elementary-age children build a strong understanding of fundamental number sense and operations, ranging from counting to ten to adding and subtracting

[2] https://www.forbes.com/sites/stevedenning/2016/08/13/what-is-agile/#3cc14e4026e3, https://www.scrumalliance.org/.

[3] https://www.ageoflearning.com/.

[4] ABCmouse, AofL's flagship product, has over 1 million users in the system to date.

three-digit numbers using the standard algorithm. The concepts covered in *MM* serve as the building blocks that allow for the development of quantitative thinking, critical to developing a strong mathematical foundation.

The app features approximately 130 games, covering number sense and operations concepts and skills for pre-kindergarten through second grade (see Fig. 12.4 for examples). Consistent with good assessment practices, every game is designed with a clear learning objective, learning task, and evidence in mind; and each learning objective is supported by an interactive instruction level, as well as several layers of scaffolding and feedback.

To support personalized instruction, *MM* recommends and customizes learning games tailored to the needs of the individual. Using engaging characters and scenarios, individualized learning pathways, and continuous assessment built into every level of every game, *MM* seeks to help children achieve proficiency through practice that incorporates repetition and variation. Based on each student's performance, the adaptive system decides what games to recommend and at which difficulty level using a predetermined network map of learning objectives and their prerequisite relationships (i.e., a node map). Adaptivity functions within individual games to provide scaffolding with each level of skill difficulty, between games to adjust to students' difficulty needs, and across the system to give players a customized pathway between skills based on performance. Assessment is embedded throughout the play experience, including game-based pretests and final assessment tasks at a granular skill level.

To develop *MM*, Age of Learning built a cross-disciplinary team of educators, learning and data scientists, and professional game developers. This team collaborated to create a game-based learning solution built upon rigorous academic curriculum, developed with a high degree of polish and engagement value, and grounded in principles of evidence-centered design, game-based assessment, and educational data mining.

Fig. 12.4 A snapshot of different games within the *Mastering Math* system

12.3.1 From Theory into Practice: iGBA Formalized into the Production Process

Mastering Math is an example of a data-driven system, enabled by fundamentals of ECD and EDM built into the development process from inception. The transfer of these integrated GBA principles into viable industry practice (particularly in the context of large-scale game development companies) is a critical and challenging part of bringing research benefits into mainstream practice for positive impact on learners at scale. This transfer requires the development of practices that (1) fit into game production methodology, (2) efficiently use team resources and apply across game contexts, and (3) are accessible to nonacademic team members—while still preserving core iGBA principles. Toward this end, the *MM* team has implemented a set of industry-based tools and practices that support evidence-based design and data collection from early stages of game development.

For baseline game development at AofL, an industry-standard methodology for software development called Scrum is used, a type of Agile development, with emphasis on short production cycles and iterative design/development phases. Generally in Scrum production, early phases constitute core conceptual design, mid-stages involve building versions of the game (e.g., Alpha and Beta[5]) with iterative user testing, and last stages produce final polished game builds. Building on this core Scrum structure, *MM* implements key iGBA design practices within each of these phases, emphasizing interdisciplinary collaboration (cf. Ke et al., 2019) as a natural, integral part of the production process. These tools and processes put iGBA principles into viable practice, embedded in a Scrum environment as development steps critical to completion of each phase's work.

12.3.2 Early and Mid-development Phase: Principled Game Design

In early stages of core conceptual design, game designers and curriculum experts (with support from the learning analytics team) came together to set the learning game foundation. First, curriculum experts work to review and define "curricular competencies" or granular learning objectives to be assessed, a core element of evidence-centered design in learning games (Mislevy et al., 2014). Next, designers and curriculum experts work together to create core evidence-based design, which is iterated throughout early and mid-development and sets the foundation for the event-stream data schema.

[5] In software development, an Alpha build is a version of the software that contains core features but not yet final functionality or polish. A Beta build is a version that is feature complete and in the early stages of final polish and tuning.

Fig. 12.5 *Mastering Math* knowledge graph overview of PreK-2 number sense and operations

In reviewing and defining learning objectives to be focused in the design of *MM*, curriculum experts quickly uncovered the challenge of no existing national PreK math curriculum (and with existing K math curriculum, learning objectives were very coarse in grain size). Investigation started with a meta-analysis of several different existing math standards for number sense and operations (including common core[6] standards, as well as non-adopting states with customized sets of standards [e.g., Texas, Florida, and Indiana]). In refining learning objectives that were granular enough for each game activity in the system and extending a curriculum map to PreK, the team also utilized canonical curricular research in learning trajectories and Building Blocks early math curriculum (Clements & Sarama, 2004; Sarama & Clements, 2004) to ultimately produce a knowledge graph of PreK-K number sense and operations for early math education. The final map defined over 130 granular learning objectives that set the foundation for the MM focus (Fig. 12.5).

Next in early (pre-production) phases, game designers and curriculum experts came together to create the core evidence-based design of each game. This was a collaborative process, in which curriculum experts brought knowledge of classroom-based practices of learning and assessment for a given learning objective to the discussion with game designers, who represented knowledge of digital design for engagement. Together, they built core design of each game, synthesizing traditional practices in teaching and measuring skills with engaging context and viable game mechanics designed to elicit evidence of a given underlying learning objective. To do this, using the principles of evidence-based game design, the team utilized a core design template aligning the game's learning objective (LO),[7] in-game task design, and resultant evidence of knowledge, skill, or ability gain. (We'll call this aligned

[6] http://www.corestandards.org/Math/Content/K/NBT/.

[7] Used synonymously with KSAs for *Mastering Math*

design template the GBA blueprint moving forward.) This principled alignment was established for each of the 130+ games in the *MM* system, each with a highly granular learning objective related to foundational number sense. Grounded in the ECD-based approach utilized in *balanced design* (Fig. 12.2; Groff et al., 2015), this allowed evidence of learning to drive game design from nascent stages, while offering an approach accessible and lean enough to be sustainable with the limited resources in industry production environments. With clear evidence of student performance aligned to granular learning objectives, this enabled subsequent design of scaffolding and formative feedback personalized to each player (an extension of the GBA blueprint). Built into the production cadence, this evidence-based process was embedded in the first phase of Scrum development (pre-production, core conceptual design). This GBA blueprint for each game (e.g., Fig. 12.6) then became the foundation for all the steps in the development process moving forward, including additional game design elements, and creating playable versions of the game for user testing and iteration. Significantly, it also served as a principled through line to the final phases of production, directly informing structure of all event-stream data collected in the system. Specifically, core assessment mechanics (e.g., the task model and resulting evidence for a given LO) defined early in the GBA blueprint became milestones in the event-stream data schema for the corresponding game, anchoring data design from early in the process.

A key goal in collaborating around this GBA blueprint is to create a rigorous yet engaging learning environment. (This is foundational in principle, since the premise of GBA is that it's situated in well-designed games that leverage the medium's capability to engage and offer meaningful context for learning.) In this process, evidence-aligned tasks became key verbs of play, around which designers then create a meaningful context for learning—vital to authentic assessment and player engagement. This is what makes learning game designers such a rare breed; "finding the fun" in design of purely commercial games is hard enough, but balancing the creation of fun mechanics and narrative with gathering core evidence of learning is even more of a challenge. Good games, indeed, consist of roles, goals, and agency (Norton, 2008; Squire, 2006) in a well-ordered series of problems with just-in-time information (Gee, 2005). In this development stage, that synthesis of narrative, meaningful context with embedded verbs of play (the game-based task model) for authentic assessment and formative feedback was created. For *MM*, this was especially important, since the app aims to support both formal and informal learning environments (and are thus in competition engagement-wise with top-shelf recreational games and apps). This is also key to the ability to scale in two ways: (1) the *MM* system is carefully themed in narrative across all games to create a seamless experience with the games' main characters—thus, plausibly building game cohesive narrative that spans over 100 granular learning objectives that is critical to system scale. (2) Sustaining interest and playability is key to retaining students in the system to support individual learning growth over time—and engagement appeal, in turn, enables popularity of a game system and enhances potential impact for many more learners. Simply put, we can help students learn the most if they're entertained enough to voluntarily spend time in the system. This means that

	Overview	Details
Learning Objective (KSA)	Cardinality. Student understands that the last number counted represents the quantity of items in a set.	• Data set: 1-5 • CPDS: Concept • Grade: PreK • Scaffolding (easy): Numerals on shapeys chests; counting out loud.
Evidence	Evidence that student has achieved the learning objective: student identifies the correct total quantity of a set of counted objects from a set of choices.	Response:Player taps on the balloon of his/her choice. Distractors/options: Three options total (two distractors).Multiple choice. Submits: No submit button. Correct/incorrect: If the selected balloon matches the target value (number of counted shapeys), the answer is correct. See SCORING section for related details.
Aligned Task(s)	Student taps on a set of given objects to count them and chooses a container that holds that specific quantity.	Context:Player is presented with a group of Shapeys, and balloons strung to baskets. Prompts: "We need to find the right basket to carry these Shapeys." "Count the Shapeys. Tap each Shapey as you count." (player taps each Shapey at least once) "Which basket is just the right size for these Shapeys?" (player taps a balloon)

Fig. 12.6 A GBA template example of a game teaching cardinality, also embeds alignment of GBA core design with event-stream data (overview of task models and evidence comprise milestones; subsequent details inform player actions and system events)

engagement, game experience, and polish are highly valued—and when synthesized with rigorous evidence-based design, potential for playful learning impact grows even further.

Honing this GBA-balanced design (with both evidence-based rigor *and* engagement) took place in the central phase of Scrum development, in which versions of the game were iteratively built by core development team (e.g., artists, engineers, and animators) and playtested. For each game, this iteration usually took place in the form of Alpha and Beta builds over several Scrum sprints. In this process, game versions were tested with a sample of players multiple times, and observational data drove refinement of design. This mid-production playtesting phase was a cornerstone of commercial game development, since players rarely interact with the game in exactly the way creators might anticipate (Salen & Zimmerman, 2004; Schell, 2008). Throughout this iterative user testing, designers and curriculum experts (both embedded in the scrum team) would work together to refine the GBA blueprint design, ultimately resulting in the level of detail shown in the right-hand column in Fig. 12.6. Elements of this detail (e.g., interactive character description, narrative, and detailed game mechanics) directly translated into player action and system event features in the event-stream data (telemetry) schema, serving to embed considerations of EDM-enabled telemetry design from inception. In the end, this process helped create a polished manifestation of engaging, evidence-based design and set the stage for final event-stream data collection.

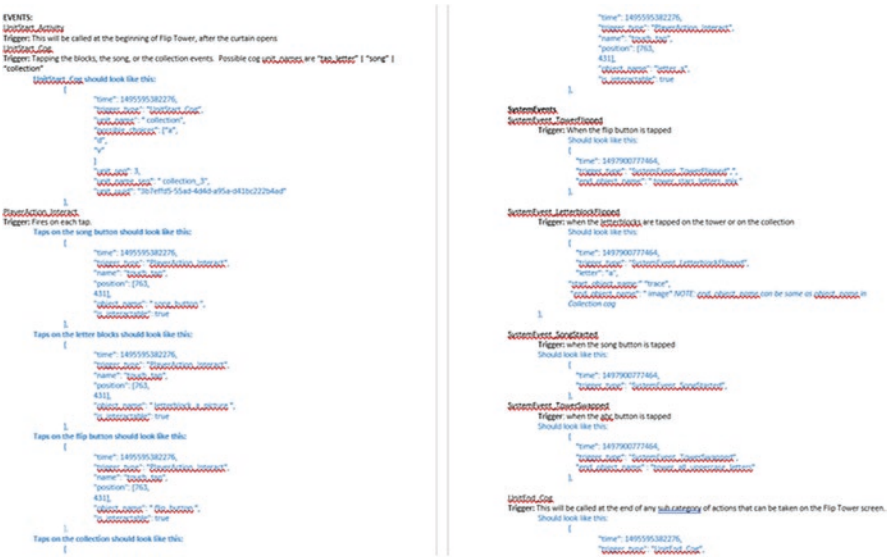
12.3.3 Data-Centered Final Phases

The final phase of game development sees the implementation of an integrated GBA data collection schema, incorporating core tasks from evidence-based design to inform structured collection of a comprehensive data stream (supportive of a wide range of analysis from LA/EDM). In this last development stage, several tools and processes have been developed to support research-grounded implementation. These include final data specification, API-based implementation, and data integrity practices in the phase before game release.

Based on recent research in flexible, comprehensive GBA-based data collection schemas like ADAGE (Halverson & Owen, 2014) and related efforts (e.g., Hao et al., 2016; Kerr & Chung, 2012), a conceptual data framework was first designed that identified task-based milestones in play, as well as all player actions and system events (collecting at the level of every click), and clearly identified corresponding results and performance outcomes. This framework was created before the *MM* system went into production, serving as cornerstone of the data-driven, game-based learning design (see Fig. 12.3).

In the final phase of live production, this data schema was used as a basis for writing data specifications for each game. Each game's data specification aligned with the GBA blueprint from the first development stage (thus integrating curriculum, game design, and analytics efforts) and incorporated information on game context and narrative—all using the language of the common data framework. As a result, the designed data streams were *clearly* aligned with core GBA design (for ease of interpretation and feedback for refining game design), *consistently* structured across games, and *comprehensive* (collecting all player interactions with the learning environment) (e.g., Owen & Baker, 2018). In this sense, *MM*'s data framework embodies iGBA, merging an evidence-based design approach (with task/evidence data milestones) with principles of LA/EDM in leveraging comprehensive, event-stream data for collection. An example of one game's data specification is given below (Fig. 12.7), modeled after JSON format to support implementation by developers in subsequent steps. Core events include milestones (or repeating core mechanics identified in the GBA template, here called "units"), as well as all player actions and system events, with persistent game context information throughout, and performance results embedded in UnitEnd events.

For execution in a fast-paced production environment, however, tools that formalized this integrated GBA data framework into implementable code were critical. The data specifications, because they were built on a collective data framework, included global events common to each game (see Fig. 12.3). Because the *MM* system is so large in scope, event types common between games (identifying information salient to learning analysis like student performance, as well as comprehensive player interaction data) were deliberately designed. Thus, this common structure was key to supporting scale, both in the system (for application to over 130 games) and in usage (affording a data stream that allowed for clear parsing and interpretability at high volume). Further supporting scale, it was also conducive to the creation of a data collection tool that could be used for implementation of data


```
EVENTS:
UnitStart_Activity
Trigger: This will be called at the beginning of Flip Tower, after the curtain opens
UnitStart_Cog
Trigger: Tapping the blocks, the song, or the collection events. Possible cog unit_names are "tap_letter" | "song" |
"collection"
    UnitStart_Cog should look like this:
        {
            "time": 1495595382276,
            "trigger_type": "UnitStart_Cog",
            "unit_name": "collection",
            "possible_choices": ["a",
            "d",
            "v"
            ]
            "unit_seq": 3,
            "unit_name_seq": "collection_3",
            "unit_uuid": "3b7effd5-55ad-4d4d-a95a-d41bc222b4ad"
        },
PlayerAction_Interact
Trigger: Fires on each tap.
    Taps on the song button should look like this:
        {
            "time": 1495595382276,
            "trigger_type": "PlayerAction_Interact",
            "name": "touch_tap",
            "position": [763,
            431],
            "object_name": "song_button",
            "is_interactable": true
        },
    Taps on the letter blocks should look like this:
        {
            "time": 1495595382276,
            "trigger_type": "PlayerAction_Interact",
            "name": "touch_tap",
            "position": [763,
            431],
            "object_name": "letterblock_a_picture",
            "is_interactable": true
        },
    Taps on the flip button should look like this:
        {
            "time": 1495595382276,
            "trigger_type": "PlayerAction_Interact",
            "name": "touch_tap",
            "position": [763,
            431],
            "object_name": "flip_button",
            "is_interactable": true
        },
    Taps on the collection should look like this:
        {
            "time": 1495595382276,
            "trigger_type": "PlayerAction_Interact",
            "name": "touch_tap",
            "position": [763,
            431],
            "object_name": "letter_a",
            "is_interactable": true
        },

SystemEvents
SystemEvent_TowerFlipped
    Trigger: When the flip button is tapped
        Should look like this:
        {
          "time": 1497900777464,
          "trigger_type": "SystemEvent_TowerFlipped",
          "end_object_name": "tower_stars_letters_mix"
        },

SystemEvent_LetterblockFlipped
    Trigger: when the letterblocks are tapped on the tower or on the collection
        Should look like this:
        {
          "time": 1497900777464,
          "trigger_type": "SystemEvent_LetterblockFlipped",
          "letter": "a",
        "start_object_name": "trace",
        "end_object_name": "image" NOTE: end_object_name can be same as object_name in
        Collection cog
    },

SystemEvent_SongStarted
    Trigger: when the song button is tapped
    Should look like this:
        {
          "time": 1497900777464,
          "trigger_type": "SystemEvent_SongStarted",
        },
SystemEvent_TowerSwapped
    Trigger: when the abc button is tapped
    Should look like this:
        {
          "time": 1497900777464,
          "trigger_type": "SystemEvent_TowerSwapped",
          "end_object_name" : "tower_all_uppercase_letters"
        },

UnitEnd_Cog
Trigger: This will be called at the end of any sub category of actions that can be taken on the Flip Tower screen.
    Should look like this:
        {
            "time": 1495595382276,
            "trigger_type": "UnitEnd_Cog",
```

Fig. 12.7 A sample data specification for a *Mastering Math* game

```
"master_account_member_id": [redacted],
"appName": "Mastery",
"gameplay_session_id": "388656241601",
"ip": "[redacted]",
"session": "dfbd0765-cb1e-4518-8abe-41791d703daf",
"payload": [
  {
    "unit_name": "TreesPretest",
    "trigger_type": "UnitStart_Activity",
    "time": 1557523723067,
    "unit_seq": 0,
    "unit_name_seq": "TreesPretest_0",
    "unit_uuid": "396e8560-14fc-493e-a6a1-c1052c4cce43"
  },
  {
    "unit_name": "tutorial_round",
    "target_value": 4,
    "trigger_type": "UnitStart_Round",
    "possible_choices": [
      4,
      2,
      1
    ],
    "unit_name_seq": "tutorial_round_0",
    "unit_seq": 0,
    "time": 1557523734930,
    "unit_uuid": "8ddae2c1-1fb0-4639-aff6-8c136305b6dd"
  },
  {
    "actionevent_name": "IP_1",
    "time": 1557523746797,
    "trigger_type": "SystemEvent_InactivityPrompt"
  },
  {
    "name": "touch_tap",
    "is_interactable": true,
    "trigger_type": "PlayerAction_Interact",
    "object_value": 4,
    "object_name": "answer_2",
    "time": 1557523749600,
    "position": [
      996,
      788
```


member_id	activity_name	unit_name	trigger_type	possible_ choices	name	object _value	position	submit _type	duration	payload-adaptive _score
blinded	UnityGame: Trees Pretest 1 - 5	TreesPretest	UnitStart_Activity							
blinded	UnityGame: Trees Pretest 1 - 5	tutorial_round	UnitStart_Round	[4, 2, 1]						
blinded	UnityGame: Trees Pretest 1 - 5	tutorial_round	SystemEvent_InactivityPrompt							
blinded	UnityGame: Trees Pretest 1 - 5	tutorial_round	PlayerAction_Interact		touch_tap	4	[996, 788]			
blinded	UnityGame: Trees Pretest 1 - 5	tutorial_round	SystemEvent_Submit					Hint		
blinded	UnityGame: Trees Pretest 1 - 5	tutorial_round	UnitEnd_Round						14767	1
blinded	UnityGame: Trees Pretest 1 - 5	first_round	UnitStart_Round	[3, 5, 1]						
blinded	UnityGame: Trees Pretest 1 - 5	first_round	PlayerAction_Interact		touch_tap	3	[558, 783]			
blinded	UnityGame: Trees Pretest 1 - 5	first_round	SystemEvent_Submit					Normal		
blinded	UnityGame: Trees Pretest 1 - 5	first_round	UnitEnd_Round						2566	1
blinded	UnityGame: Trees Pretest 1 - 5	second_round	UnitStart_Round	[1, 2, 3]						
blinded	UnityGame: Trees Pretest 1 - 5	second_round	PlayerAction_Interact		touch_tap	1	[1036, 758]			
blinded	UnityGame: Trees Pretest 1 - 5	second_round	SystemEvent_Submit					Normal		
blinded	UnityGame: Trees Pretest 1 - 5	second_round	UnitEnd_Round						2300	1
blinded	UnityGame: Trees Pretest 1 - 5	third_round	UnitStart_Round	[5, 2, 3]						
blinded	UnityGame: Trees Pretest 1 - 5	third_round	PlayerAction_Interact		touch_tap	5	[1396, 782]			
blinded	UnityGame: Trees Pretest 1 - 5	third_round	SystemEvent_Submit					Normal		
blinded	UnityGame: Trees Pretest 1 - 5	third_round	UnitEnd_Round						2566	1
blinded	UnityGame: Trees Pretest 1 - 5	fourth_round	UnitStart_Round	[2, 5, 1]						
blinded	UnityGame: Trees Pretest 1 - 5	fourth_round	PlayerAction_Interact		touch_tap	2	[561, 728]			
blinded	UnityGame: Trees Pretest 1 - 5	fourth_round	SystemEvent_Submit					Normal		
blinded	UnityGame: Trees Pretest 1 - 5	fourth_round	UnitEnd_Round						2766	1
blinded	UnityGame: Trees Pretest 1 - 5	fifth_round	UnitStart_Round	[4, 2, 1]						
blinded	UnityGame: Trees Pretest 1 - 5	fifth_round	PlayerAction_Interact		touch_tap	4	[982, 776]			
blinded	UnityGame: Trees Pretest 1 - 5	fifth_round	SystemEvent_Submit					Normal		
blinded	UnityGame: Trees Pretest 1 - 5	fifth_round	UnitEnd_Round						2133	1
blinded	UnityGame: Trees Pretest 1 - 5	TreesPretest	UnitEnd_Activity						49736	

Fig. 12.8 Sample data streaming from the MM data collection API (details blinded for privacy)

specifications. Thus, the learning analytics team created a data collection API, formalizing research-based GBA data structures into a tool that was accessible and efficient for game engineers. This API was similar to related tool development efforts (e.g., Danielak, 2014; Kevan & Ryan, 2016) but for expedience was tailored to *MM*'s common game structures and Unity development software and designed to inherently reinforce principled GBA data collection. Software engineers on the Scrum team implemented data specifications into the games in late Beta using this API during the last development phase. Sample output from the API follows below, both in "raw" JSON format and in CSV format (with each line as a separate event). These data follow the format outlined in the ADAGE-based data schema and formalized in the architecture of the Unity data collection API (Fig. 12.8).

Lastly, final development stages required completed, accurate data collection before the game could launch. Integrating data testing for quality assurance (QA) during two separate cycles (first in the late Beta build and again in the final release build) helped ensure that end-to-end data output was complete and accurate. These cycles of QA testing (involving initial test cases, identifying bugs, reporting bugs, and then verifying fixes) were supplemented with the use of an original, automated data integrity tool called DVIT (Data Validation and Integrity Tool; Keylor & Beukers, 2018). This way, the API and subsequent data integrity testing (Figs. 12.7 and 12.8) served as a final through line for evidence-based design from the early stages of development, while furnishing a rich, comprehensive data stream that fueled the system's personalized formative feedback and data-driven iterative design for learning and engagement.

12.3.4 Mastering Math *Development Process: Summary*

Overall, to put research-based principles into sustainable industry practice, core tenants of iGBA were formalized into tools and processes and embedded throughout the Scrum development process—from a formal GBA blueprint in nascent development to aligned game narrative design and Alpha/Beta production and a GBA data schema translated into a data collection API (and data integrity tools) in final phases. Table 12.1 shows a summary of this implementation, with key iGBA elements of curricular competency design; collaborative, evidence-based game design; and aligned event-stream data implementation mapped to stages of game development (from pre-production to release).

In implementing this process, the investment required was nontrivial, asking assessment designers, learning analysts, designers, and curriculum experts to collaboratively create, commit to, and utilize tools that created mutual value (e.g., the GBA blueprint). A ramp-up period of several months before any production began was needed to create and foster buy-in to key elements of the process (e.g., the GBA blueprint, conceptual data schemas, and the corresponding API) and required commitment of resources by executives who held the vision important. In addition, the first few sprints once game development started were a bit bumpy, taking some iteration to get the cadence of iGBA tool/process use right. However, once the timing, buy-in, and coordination between team members around iGBA tool and process

Table 12.1 Overview of the *Mastering Math* development process (implementation of an iGBA approach)

Stage	Build	Action	Process/tools
Early and mid-development	Pre-production	Curricular research (curricular competency)	• Curriculum experts review and define LOs
		Collaborative game design	• Collaboration: game designer and curriculum expert • Tool: GBA template is used to define initial core design (milestones for data specification are established)
	Alpha and Beta	Design iteration	• User testing • Tool: GBA template iteration (corresponding details of data specification are defined)
Late development	Late Beta	Data implementation	• Data specification design per game is finalized (based on GBA template and Beta build) • Tool: MM learning analytics API is used to efficiently implement data specification
	Prerelease	Data integrity	• QA cycles (check, submit bugs, verify fixes) • Tool: DVIT-automated QA tool support

use was stabilized, it resulted in a scalable development approach that carried *MM* through to completion.

Building this development process enabled the creation of a broad-scope, rigorous, and engaging game-based learning system positioned to reach a large number of students in both formal and informal learning environments. Specifically, integrated GBA in practice allows *Mastering Math* to function as a data-driven system, in which tasks provide evidence-based formative feedback and personalization and insights from large, click-stream log files can inform iterative design and intelligent personalization.

12.3.5 Closing the Data-Driven Design Loop: Post-production Analysis

After the initial Scrum phases and launch of the game, analysis of the *MM* event-stream data can inform iterative design to hone the playful learning experience for students. The well-structured event-stream data—enabled by implemented iGBA—supports multiple approaches to analysis, which in turn can provide insight into player patterns to inform iterative design and intelligent personalization. These analyses then serve to close the iterative design loop in *MM*'s data-driven system to support ongoing learning and engagement.

MM's event-stream data, enabled by iGBA-based development, enables many kinds of analysis for learning insight and design iteration. For example, as discussed in a recent review of LA/EDM methods for game-based learning analytics (Owen & Baker, 2019), data visualization is a primary category of analysis that can directly support communication of insights to designers. *MM* leverages this kind of analysis for iterative design, with regularly updated data visualizations of player data being sent to designers. For instance, currently in deployment are simple indices that represent in-system learning game progress, called key learning indicators (KLIs). Visualizations of these metrics are delivered to designers on a biweekly basis for monitoring of student activity in the system related to learning game performance. KLI metrics include the average number of LOs students are mastering per week, what percent of LOs started on are students mastering, most frequently mastered LOs, most frequently failed games (each corresponding to an LO), and most frequently canceled games. These can be useful to help designers monitor activity and make data-driven decisions to refine design. For example, as new versions of the *MM* system have come out, cancellation rate has been used to evaluate which games need design iteration and/or re-prioritization in the system. For instance, it was discovered a sequencing number game had a high cancellation rate in broad event-stream usage, and further investigation and user testing revealed that forcing kids to listen to lengthy voice-over instructions (before being able to interact with the game) was deterring students from finishing the activity. Subsequent design tweaks to the voice-over rules allowed earlier interaction with the game and fewer resultant cancellations. Recently, based on a visualization of *MM*-based pretest pass rates by learning objective and age, system designers were able to discern which pretests were too difficult for younger students. For example, analysis from an early release of *MM* showed that core LO pretests *counting forward 1–10*, *numeral recognition 11–15*, and *counting out 6–10* had a low pass rate (between 27% and 36%) for the target age of 4 years old, suggesting that these may not be optimal general starting skills in the system. In response, the pretest system was customized to student age upon starting *MM*, which resulted in better engagement and more balanced pass rates in subsequent versions.

In deeper event-stream analyses, EDM methods have been utilized to better understand emergent student patterns, particularly related to learning-supportive behaviors and building learner profiles. These efforts are geared toward more intelligent feedback and personalization, on the premise that if models can be used to identify behavior related to learning in real time, the system can then give specialized feedback and alternative support pathways. For example, using data from a recent pilot study (Thai, Li, & Schachner, 2019), profiles of prior knowledge were built using cluster analysis and then detected for the event-stream data so that students could be automatically classified into prior knowledge level as they played. In this instance, a simple k-means clustering algorithm was chosen, using several event-stream variables around student performance, engagement, and time duration linked to in-game pretest evaluation. We found three groups of prior knowledge level in kindergarten-level students (roughly equivalent to high, medium, and low initial knowledge). Using behavior detection methods, we then used the prior knowledge outcome labels to accurately classify students based on event-stream interaction. Predictive patterns characteristic of each group included a low overall pass rate for students in the lowest prior knowledge group, students with emerging knowledge of number identification in the middle group, and high pass rates on sequencing skills for students in the top knowledge group. This informs potential student profiles based on prior knowledge, enabling better customization of learning pathways, as well as additional investigation into user patterns (particularly in service of the lower prior knowledge students). In deriving prior knowledge profiles and classifying students based on real-time interaction data, the goal is to more deeply understand student needs and be able to intelligently respond in-system for better personalization.

In related EDM-based analyses, the learning analytics team recently leveraged behavior detection methods to build a predictor to track when students are stuck in the system. Essentially, to more intelligently detect students that need extra support outside of existing system scaffolds, the analysis focused on "wheel spinning," a form of unproductive effort, or spending too much time struggling to learn a topic without achieving mastery (Beck & Gong, 2013). In being able to detect wheel-spinning students in the system (as differentiated from those who are productively persistent (e.g., Kai, Almeda, Baker, Heffernan, & Heffernan, 2018) through a predictive model of behavior, we gained the potential to better respond to students who need support and be able to surface these insights to educators as well for additional, interpersonal intervention. Different pathways emerged in relationship to wheel-spinning status, including younger students with very few passed games (stuck on the easiest skills in the system) and students with a few more skills mastered overall but with a very slow rate of progress. Interestingly, these correspond with prior knowledge clustering results, with 80% of stuck students in the lowest prior knowledge cluster. These investigations into emergent student patterns have concretely informed design of the system, resulting in a new section of *MM* being built out just for students who need more exposure and practice with very basic number sense concepts.

In this sense, enabled by the iGBA-based design of the *Mastering Math* system, rich data streams allow a broad range of methods to be applied for better understanding student play patterns and improvement of core design as well as intelligent personalization. Enabling refinement of design—to support students of all prior knowledge levels and pathways through the system—closes the loop of an iteratively refined, data-driven system and opens potential for even better playful learning experiences for students at scale.

12.4 *Mastering Math*: Implications for Classroom Use

In addition to surfacing information for designers, visualization of insights for teachers is important for enabling real-time, interpersonal intervention for students (Mislevy et al., 2016). Principled game design enables salient visualization for teachers for visualization of detailed information about learning performance in play (Ke et al., 2019). The iGBA-based *MM* system is positioned to support teachers with granular visualization of student progress. Indeed, supporting use of *MM* in formal learning environments is central to the team's mission, as Age of Learning offers the core product free for teachers and has shown promising efficacy results in classroom-based research to date.

Grounded in principled design, *MM* has clearly aligned learning objectives (LOs) with specific games and core assessment mechanics, furnishing data on performance at the skill level (which fundamentally drives formative feedback and personalization). This additionally enables the display of information to educators about student performance at a granular LO level (e.g., identifying numerals 1–5, counting forward with numbers 6–10, or numeral decomposition 16–20). With every student taking a different pathway through the personalized learning system, summary visualization of progress becomes even more important. To this end, dashboards are being piloted in which teachers can see student progress on a classroom level through the system (Fig. 12.9). While in nascent stages, these views can serve as a basis for facilitating teacher understanding of student progress for individualized support, student grouping, and even informing of whole classroom lessons tailored to common areas of need. In addition, these visualizations will soon incorporate EDM-based insights—which can detect and flag students stuck in the system, for example, and offer an estimate of student mathematics knowledge coming into the system. This is especially important for teachers in poorly funded schools,

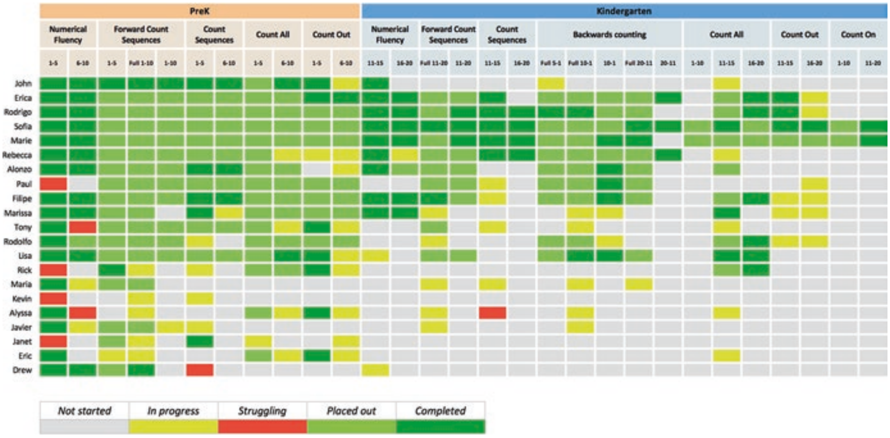


Fig. 12.9 Snapshot of a teacher dashboard prototype (with fictional student names)

which often do not have the resources to pretest mathematics knowledge of students. To continue to support classroom use, ongoing efforts in learning analytics and visualization seek to further develop these dashboards in research-based design iterations.

To better support math education efforts in the classroom, researchers recently ran a pilot study for kindergarten students using *MM*. This randomized controlled study evaluated the impact of *MM* in 20 classrooms, using a validated measure of early mathematics called the TEMA-3 (Ginsburg & Baroody, 2003). Students who were assigned to use *MM* in the classroom showed accelerated mathematics learning after a usage average of 5 hours over 10 weeks (compared to students who did not receive *MM*) as measured by the TEMA (effect size = 0.23; Thai et al., 2019). Results also showed strong alignment of in-system pretest performance and TEMA pretest scores, as well as in-game final assessment and TEMA posttest results (Jacobs et al., 2018). Additionally, evaluation at the item level also revealed that students who played *MM* showed the greatest gains on the most difficult mathematics skills as measured by the external assessment (Thai et al., 2019). Informing future iterations of the system, the data collected in the study also enabled detection of wheel spinning and productive persistence (Owen et al., 2019) and deeper exploration of prior knowledge patterns for better personalization.

12.5 Conclusion and Future Work

Principled GBA can support data-driven systems for game-based learning, enabling engaging, effective playful learning experiences. This leverages the power of games, which can naturally be immersive, interactive vehicles for learning and authentic assessment. For strong foundations in core design as well as data-driven iteration for learning and engagement, this research takes an integrated approach to game-based assessment—grounded in tenets of evidence-based design for embedded assessment mechanics, which serve as touchstones in broader, comprehensive event-stream data collection (consistent with principles of EDM) to enable a wide range of learning insights. This kind of iGBA allows for a data-driven system, with formative feedback and learner-responsive pathways, and iterative, insight-based design on core assessment mechanics as well as emergent play patterns critical to learning and engagement. These insights can then fuel iteration on core design as well as intelligent personalization (e.g., through EDM-based discovery and detection of student patterns) for optimizing learning and engagement.

Implementing these research-based practices in industry-based production environments can be key to leveraging iGBA benefits for impact at scale. *Mastering Math* is just one example of such a system, creating accessible, efficient tools and processes for implementing principled GBA in industry-standard production paradigms. As a result, we can leverage rich data streams for data-driven design, intelligent personalization, and educator insight for student support in system and at the classroom level. Pilot studies show promising results, as *MM* (through broad distribution channels) is positioned to deliver a personalized, game-based learning system to reach students at scale.

This kind of industry-relevant research can support the development of best practices for implementation of principled GBA in large-scale production environments. However, as noted above, getting into a smooth *MM* development process in which all team members were bought in and iGBA tools were synchronized with the development process took some time. A lesson learned from this iGBA implementation in industry is that it takes clear vision, backing from executives, and an initial investment of resources to launch a successful principled design and development process. Exact implementation of an iGBA approach will very likely vary based on production environment and product type, and further study across industry contexts is needed to support the concrete development of best practices. However, the completion of *Mastering Math* with a steady cadence of iGBA-based production practices, and the resultant empirical results in learning outcomes, is a promising example of viably implementing principled learning game development in an industry setting. In *Mastering Math*'s immediate future, larger-scale research studies can bring deeper insight about efficacy, in-game assessment, and emergent student patterns in order to inform iterative design for personalized learning and engagement. As research on teacher visualizations continue, so does the potential to refine dashboards to enable more educator tools—such as specific skill recommendations for next steps, automatic grouping of students (heterogeneously and homogeneously), and intelligent detection tools for identifying and flagging misconceptions as well as specific math patterns recurring on an inter-skill level (e.g., consistent reversal of numbers 6 and 9).

More broadly, as we move more deeply into immersive games for education, future work lies in the development of game-based assessment in the brave new worlds of VR- and XR-based learning environments. Shifting paradigms of technologies call for constantly evolving methodology in leveraging them for educational purposes, including reimagining the kind of data we collect (e.g., in 3D VR space), what learning sciences we can leverage to understand it (e.g., embodied cognition; Wilson, 2002), and carefully investigating what impact these experiences might have on students of different ages and stages.

Most of all, the application of principled game design and iGBA to the educational space can impact those who this work is ultimately built to serve: the students. With principled GBA embedded in immersive learning vehicles, players can have an engaging learning experience tailored to their needs—in which evaluation is embedded seamlessly into core actions of the game and the only necessary assessment is play itself. As we move more deeply into a digital age of education, research-based game development practices have the potential to enable transformative, playful learning experiences across both formal and informal learning environments for impact at scale.

Acknowledgements We wish to thank the entire mastery team at Age of Learning, especially Doug Dohring, Sunil Gunderia, Daniel Jacobs, K.P. Thai, and Vesper Burnett.

References

Aarseth, E. (2007). I fought the law: Transgressive play and the implied player. In *Situated Play. Proceedings of DiGRA* (pp. 24–28).

Asbell-Clarke, J., Rowe, E., & Sylvan, E. (2013). Assessment design for emergent game-based learning. In W. E. MacKay, S. Brewster, & S. Bodker (Eds.), *CHI'13 extended abstracts on human factors in computing systems* (pp. 679–684). Retrieved from http://dl.acm.org/citation.cfm?id=2468476

Baker, E. L., Chung, G. K. W. K., & Delacruz, G. C. (2012). The best and future uses of assessment in games. In *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 227–246). Charlotte, NC: IAP - Information Age Publishing, Inc.

Baker, R. S., & Clarke-Midura, J. (2013). Predicting successful inquiry learning in a virtual performance assessment for science. In S. Carberry, S. Weibelzahl, A. Micarelli, & G. Semeraro (Eds.), *Proceedings of the 21st International Conference on User Modeling, Adaptation, and Personalization* (pp. 203–214). Retrieved from http://www.columbia.edu/~rsb2162/UMAP-2013-BCM-v9.pdf

Baker, R. S., & Siemens, G. (2014). Educational data mining and learning analytics. In K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (2nd ed., pp. 253–274). New York, NY: Cambridge University Press.

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., & Gong, Y. (2013). Wheel-spinning: Students who fail to master a skill. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), Artificial Intelligence in Education (Vol. 7926, pp. 431–440). https://doi.org/10.1007/978-3-642-39112-5_44

Bielaczyc, K., & Kapur, M. (2010). *Playing epistemic games in science and mathematics classrooms*. Retrieved from https://repository.nie.edu.sg/handle/10497/14456

Canossa, A., Badler, J. B., El-Nasr, M. S., Tignor, S., & Colvin, R. C. (2015, June). *In your face (t) impact of personality and context on gameplay behavior*. Presented at the FDG Conference (Foundations of Digital Games), Pacific Grove, CA.

Chung, G. K. W. K. (2015). Guidelines for the design, implementation, and analysis of game telemetry. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: Methodologies for performance measurement, assessment, and improvement* (pp. 59–79). New York, NY: Springer.

Clarke-Midura, J., Code, J., Dede, C., Mayrath, M. C., & Zap, N. (2012). Thinking outside the bubble: Virtual performance assessments for measuring complex learning. In M. C. Mayrath, J. Clarke-Midura, D. H. Robinson, & G. Schraw (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 125–148). Charlotte, NC: Information Age Publishing.

Clements, D. H., & Sarama, J. (2004). Learning trajectories in mathematics education. *Mathematical Thinking and Learning, 6*(2), 81–89. https://doi.org/10.1207/s15327833mtl0602_1

Corrigan, S., DiCerbo, K. E., Frenz, M., Hoffman, E., John, M., & Owen, V. E. (2015). *GlassLab game design handbook*. Retrieved from http://gamedesign.glasslabgames.org/

Danielak, B. (2014). *Analyzing data with ADAGE*. Retrieved from https://capbri.gitbooks.io/makescape-adage-gitbook/

DiCerbo, K. E. (2014). Task models in the digital ocean. *Measurement: Interdisciplinary Research and Perspectives, 12*(1–2), 39–41.

DiCerbo, K. E., Bertling, M., Stephenson, S., Jia, Y., Mislevy, R. J., Bauer, M., & Jackson, T. G. (2015). An application of exploratory data analysis in the development of game-based assessments. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: Methodologies for performance measurement, assessment, and improvement* (pp. 319–342). New York, NY: Springer.

DiCerbo, K. E., & Kidwai, K. (2013). Detecting player goals from game log files. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 314–316). Retrieved from http://www.educationaldatamining.org/EDM2013/proceedings/EDM2013Proceedings.pdf

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

Gee, J. P. (2005). Learning by design: Good video games as learning machines. *E-Learning and Digital Media, 2*(1), 5–16.

Gee, J. P. (2012, September). *Games can drive assessment to a new place*. Retrieved from http://gamesandimpact.org/wp-content/uploads/2012/09/Games-Can-Drive-Assessment-to-a-New-Place.pdf

Ginsburg, H., & Baroody, A. (2003). *TEMA-3 examiners manual* (3rd ed.). Austin, TX: PRO-ED.

Grace, L. D. (2014). A linguistic analysis of mobile games: Verbs and nouns for content estimation. In *Proceedings of the 9th International Conference on the Foundations of Digital Games* (p. 8). FDG.

Groff, J., Clarke-Midura, J., Owen, V. E., Rosenheck, L., & Beall, M. (2015). *Better learning in games: A balanced design lens for a new generation of learning games [white paper]*. Cambridge, MA: MIT Education Arcade and Learning Games Network. Retrieved from https://www.media.mit.edu/publications/better-learning-in-games-a-balanced-design-lens-for-a-new-generation-of-learning-games/

Halverson, R., & Owen, V. E. (2014). Game based assessment: An integrated model for capturing evidence of learning in play. *International Journal of Learning Technology, 9*(2), 111–138. https://doi.org/10.1504/IJLT.2014.064489

Hao, J., Smith, L., Mislevy, R., von Davier, A., & Bauer, M. (2016). *Taming log files from game/simulation-based assessments: Data models and data analysis tools* (pp. 1–17). Princeton, NJ: Educational Testing Service.

Hicks, D., Eagle, M., Rowe, E., Asbell-Clarke, J., Edwards, T., & Barnes, T. (2016). Using game analytics to evaluate puzzle design and level progression in a serious game. In D. Gasevic, G. Lynch, S. Dawson, H. Drachsler, & C. P. Rosé (Eds.), *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 440–448). https://doi.org/10.1145/2883851.2883953

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

Jacobs, D., Thai, K. P., Owen, V. E., Keylor, E., Roy, M.-H., & Burnett, V. (2018, October). *Mastering Math: Comparing an educational video game to an established assessment*. Presented at the Meaningful Play, East Lansing, MI.

Juul, J. (2013). *The art of failure: An essay on the pain of playing video games*. Cambridge, MA: MIT.

Kai, S., Almeda, M. V., Baker, R. S., Heffernan, C., & Heffernan, N. (2018). Decision tree modeling of wheel-spinning and productive persistence in skill builders. *Journal of Educational Data Mining, 10*(1), 36–71.

Kai, S., Paquette, L., Baker, R. S., Bosch, N., D'Mello, S., Ocumpaugh, J., … Ventura, M. (2015). A comparison of video-based and interaction-based affect detectors in physics playground. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, …M. C. Desmarais (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining* (pp. 77–84). Retrieved from http://www.educationaldatamining.org/EDM2015/proceedings/edm2015_proceedings.pdf

Ke, F., Shute, V. J., Clark, K. M., & Erlebacher, G. (2019). Designing dynamic support for game-based learning. In *Interdisciplinary design of game-based learning platforms* (pp. 119–140). New York, NY: Springer.

Kerr, D., & Chung, G. (2012). Identifying key features of student performance in educational video games and simulations through cluster analysis. *Journal of Educational Data Mining, 4*(1), 144–182.

Kevan, J. M., & Ryan, P. R. (2016). Experience API: Flexible, decentralized and activity-centric data collection. *Technology, Knowledge and Learning, 21*(1), 143–149. https://doi.org/10.1007/s10758-015-9260-x

Keylor, E., & Beukers, D. (2018). *DVIT (data validation and integrity tool)*. Glendale, CA: Age of Learning, Inc.

Martinez-Garza, M. M., & Clark, D. B. (2017). Investigating epistemic stances in game play with data mining. *International Journal of Gaming and Computer-Mediated Simulations, 9*(3), 1–40. https://doi.org/10.4018/IJGCMS.2017070101

Mislevy, R., Oranje, A., Bauer, M., von Davier, A., Hao, J., Corrigan, S., … John, M. (2014). *Psychometric considerations in game-based assessment*. Redwood City, CA: GlassLab.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *Research Report-Educational Testing Service Princeton RR, 16*. Retrieved from http://marces.org/EDMS623/Mislevy%20on%20ECD.pdf

Mislevy, R. J., Corrigan, S., Oranje, A., DiCerbo, K. E., Bauer, M., von Davier, A., & John, M. (2016). Psychometrics and game-based assessment. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 23–48). New York, NY: Routledge.

Mislevy, R. J. (2011). *Evidence-centered design for simulation-based assessment (no. 800)*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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), 38.

Norton, D. (2008, June). *A practical model for separating games and simulations*. Presented at the 5th Annual Games+Learning+Society Conference. Retrieved from http://www.previous.glsconference.org/2008/session.html?id=139

Owen, V. E., Anton, G., & Baker, R. S. (2016). Modeling user exploration and boundary testing in digital learning games. In J. Vassileva, J. Blustein, L. Aroyo, & S. D'Mello (Eds.), *Proceedings of the 2016 conference on user modeling adaptation and personalization* (pp. 301–302). New York, NY: ACM.

Owen, V. E., & Baker, R. S. (2018). Fueling prediction of player decisions: Foundations of feature engineering for optimized behavior modeling in serious games. *Technology, Knowledge and Learning, 24*, 1–26. https://doi.org/10.1007/s10758-018-9393-9

Owen, V. E., & Baker, R. S. (2019). Learning analytics for games. In J. L. Plass, R. Meyer, & B. D. Homer (Eds.), *Handbook of game-based learning*. Cambridge, MA: MIT.

Owen, V. E., Roy, M.-H., Thai, K., Burnett, V., Jacobs, D., Keylor, E., & Baker, R. S. (2019, July). *Detecting productive persistence in an adaptive game-based learning system*. Presented at the 12th International Conference on Educational Data Mining, Montreal, Canada.

Owen, V. E., Wills, N., & Halverson, R. (2012). CyberSTEM: A game-based evidence model. In A. Ochsner (Ed.), *Proceedings of the 8th Annual Games, Learning, and Society Conference*. Pittsburgh, PA: ETC.

Plass, J. L., Homer, B. D., Kinzer, C. K., Chang, Y. K., Frye, J., Kaczetow, W., … Perlin, K. (2013). Metrics in simulations and games for learning. In M. Seif El-Nasr, A. Drachen, & A. Canossa (Eds.), *Game analytics* (pp. 697–729). https://doi.org/10.1007/978-1-4471-4769-5_31

Plass, J. L., Homer, B. D., Kinzer, C. K., & Perlin, K. (2012). *Games for learning institute (G4LI) white paper: Ideas for impact games*. Retrieved from http://gamesandimpact.org/wp-content/uploads/2012/09/PlassNYU-Ideas-for-Impact-Games-2.pdf

Ramirez, D. (2016). *How player movement data improves educational game assessments* (Doctoral dissertation, University of Wisconsin-Madison). Retrieved from ProQuest Dissertations and Theses database (14118).

Rieber, L. P. (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.

Rodrigo, M. M. T., & Baker, R. S. (2011). Comparing learners' affect while using an intelligent tutor and an educational game. *Research and Practice in Technology Enhanced Learning, 6*(1), 43–66.

Salen, K., & Zimmerman, E. (2004). *Rules of play: Game design fundamentals*. Cambridge, MA: MIT.

Sarama, J., & Clements, D. H. (2004). Building blocks for early childhood mathematics. *Early Childhood Research Quarterly, 19*(1), 181–189. https://doi.org/10.1016/j.ecresq.2004.01.014

Schell, J. (2008). *The art of game design: A book of lenses*. Burlington, MA: Elsevier.

Serrano-Laguna, A., Martinez-Ortiz, I., Haag, J., Regan, D., Johnson, A., & Fernández-Manjóna, B. (2017). Applying standards to systematize learning analytics in serious games. *Computer Standards & Interfaces, 50*, 116–123. https://doi.org/10.1016/j.csi.2016.09.014

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). Retrieved from http://myweb.fsu.edu/vshute/pdf/shute%20pres_h.pdf

Shute, V. J., D'Mello, S., Baker, R. S., Cho, K., Bosch, N., Ocumpaugh, J., … Almeda, V. (2015). Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. *Computers & Education, 86*, 224–235. https://doi.org/10.1016/j.compedu.2015.08.001

Shute, V. J., & Kim, Y. J. (2014). Formative and stealth assessment. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (pp. 311–321). https://doi.org/10.1007/978-1-4614-3185-5_25

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior, 63*, 106–117. https://doi.org/10.1016/j.chb.2016.05.047

Slater, S., Bowers, A. J., Kai, S., & Shute, V. J. (2017, July). *A typology of players in the game physics playground*. Presented at the Digital Games Research Association (DiGRA), Melbourne, Australia.

Squire, K. (2006). From content to context: Videogames as designed experience. *Educational Researcher, 35*(8), 19–29. https://doi.org/10.3102/0013189X035008019

Squire, K. (2011). Video games and learning. *Teaching and participatory culture in the digital age*. Retrieved from http://eric.ed.gov/?id=ED523599

Steinkuehler, C., & Duncan, S. (2008). Scientific habits of mind in virtual worlds. *Journal of Science Education and Technology, 17*(6), 530–543. https://doi.org/10.1007/s10956-008-9120-8

Stenerson, M. E., Salmon, A., Berland, M., & Squire, K. (2014). Adage: An open API for data collection in educational games. In *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play* (pp. 437–438). New York, NY: ACM.

Stephenson, S., Baker, R. S., & Corrigan, S. (2014, June). *Towards building an automated detector of engaged and disengaged behavior in game- based assessments*. Presented at the 10th Annual Games+Learning+Society Conference, Madison, WI. Retrieved from http://radix.www.upenn.edu/learninganalytics/ryanbaker/GLS_Stephenson_v2.pdf

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.

Thai, K. P., Li, L., & Schachner, A. (2019, April). *Accelerating early math learning with a digital math resource: A cluster randomized controlled trial*. Presented at the 2019 AERA Annual Meeting, Toronto, Canada.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review, 9*(4), 625–636. https://doi.org/10.3758/BF03196322

Zagal, J. P., Mateas, M., Fernández-Vara, C., Hochhalter, B., & Lichti, N. (2005). Towards an ontological language for game analysis. In *Proceedings of Digra 2005 Conference* (p. 13). Authors & Digital Games Research Association DiGRA.

Chapter 13
Effectiveness of Supply Chain Games in Problem-Based Learning Environment

Linda William, Za' Aba Bin Abdul Rahim, Liping Wu, and Robert de Souza

13.1 Introduction

Serious game has been introduced as an interactive educational tool for teaching specific knowledge (Ma, Oikonomou, & Jain, 2011; de Freitas & Liarokapis, 2011). It incorporates non-entertainment elements, such as concepts of SCM and urban logistics, into game environment (Liu, Alexandrova, & Nakajima, 2011). It serves as a pedagogical tool with a learning purpose, moving beyond entertainment to deliver engaging interactive media to support learning (de Freitas, 2006). Serious game is designed to distill specific and complex learning concepts while maintaining the entertainment factors through the student engagement and interaction with information, tools, materials, other students as well as the lecturer/facilitator within the game (Kim, Park, & Baek, 2009). It also provides learning engagement, motivations (Riedel & Hauge, 2011) and constant feedback that help to form the students' skills and knowledge. These characteristics enable serious game to be used as a tool for formative assessment (Delacruz, 2011; Handfield-Jones, Nasmith, Steinert, & Lawn, 1993; Wang, 2008).

Previous studies have identified the benefits of using game-based learning to support teaching and learning (Ma et al., 2011) and to conduct a formative assessment (Delacruz, 2011; Handfield-Jones et al., 1993; Wang, 2008). It includes enhancing engagement and encouraging curiosity, motivation, self-monitoring, self-assessment,

L. William (✉)
School of Informatics and IT, Temasek Polytechnic, Singapore, Singapore
e-mail: lwilliam@tp.edu.sg

Z. A. B. A. Rahim · R. de Souza
The Logistics Institute – Asia Pacific, National University of Singapore, Singapore, Singapore
e-mail: zaaba@nus.edu.sg; rdesouza@nus.edu.sg

L. Wu
School of Engineering, Republic Polytechnic, Singapore, Singapore
e-mail: wu_liping@rp.edu.sg

D. Ifenthaler, Y. J. Kim (eds.), *Game-Based Assessment Revisited*, Advances in Game-Based Learning, https://doi.org/10.1007/978-3-030-15569-8_13

problem-solving and decision making (Knight et al., 2010; Kumar, 2000; Ma et al., 2011; Rieber, 1996). It promotes active participation and interaction as a center of experience to advance the students' understanding (Hou, 2015). Using rich visual and spatial aesthetics, serious game would be able to immerse the students into the game world. The students would voluntarily form new skills and improve their knowledge based on the formative feedback to complete certain tasks and challenges in the game environment.

Although it has been identified that serious game can be used in Problem-based Learning (PBL) environment as a framework for generating and introducing problems to the students (Kim et al., 2009; Kiili, 2005; Burguillo, 2010), there is very limited research on the game's effectiveness as a formative assessment tool in PBL. Different from the traditional learning environment, PBL is a student-centered instructional method which learning is conducted through active learning with a self-directed manner in the collaborative environment for solving the given problem(s) (Hmelo-Silver, 2004). In PBL, an open-ended problem is provided to the students where they can work on it in small collaborative groups to identify the knowledge that they need to solve the problem, apply their knowledge to the problem and evaluate the strategies employed to solve the problem. Serious game can be used to provide a risk-free environment for formative assessment where the students can identify the problems, assess their skill and knowledge and receive feedback based on their actions in the game environment.

Motivated by this insight, this work aims to evaluate the effectiveness of three SCM games as formative assessment tools in the PBL classroom. We design a research model and hypotheses by focusing on the purported benefits of PBL, such as independent learning, greater understanding and lifelong learning skills, as well as essential components of formative assessment. It considers two main criteria, namely: learning objective and game experience. Learning objective criterion assesses the game's ability to initiate the learning and formative assessment process in PBL through problem generation, engagement and feedback. This criterion includes two cognitive components, namely: (1) metacognitive functions and (2) motivation (Surgrue, 1995). While game experience criterion evaluates the games' effectiveness to create an enjoyable experience for voluntary learning.

We use this research model to evaluate three digital SCM games, namely: Beer Distribution game (Sterman, 1989; Wisner, Tan, & Leong, 2014), ACE E-Commerce game (Lindawati, Rahim, & de Souza, 2018) and Disaster Relief game (de Souza, William, Timperio, & Rahim, 2018; The Logistics Institute - Asia Pacific, 2017a, 2017b), as formative assessment tools in the PBL classroom in Republic Polytechnic (RP) (Republic Polytechnic, 2019a, 2019b). We conducted four game sessions from October 2018 to January 2019 with different groups of students for each session. In the first session, we played Beer Distribution game, while in the other three sessions, we played ACE E-Commerce game and Disaster Relief game. At the end of the game sessions, we evaluated the games' effectiveness using a questionnaire derived from the research model. The evaluation results show that these games provide motivation and a positive learning experience that improve students' perceived learning. The results also show that the students were able to absorb the learning objectives and learned about SCM concepts.

The outline of the chapter is as follows: Sect. 13.2 presents the literature review of game-based formative assessment, games for SCM, PBL and typical PBL implementation in RP. Section 13.3 describes the research model and hypotheses for assessing the effectiveness of the games as formative assessment tools in PBL environment. Section 13.4 reviews the research design, including the three SCM games. Section 13.5 describes the evaluation result, and Sect. 13.6 discusses the results and recommendations to use game as a formative assessment tool in a PBL environment. Section 13.7 presents conclusions and future research directions.

13.2 Literature Review

13.2.1 Game-Based Formative Assessment

Learning assessment is commonly divided into two forms, namely: formative and summative assessment. Formative assessment concerns on improving students' competency through integrated interaction and iterative feedback between lecturer and students while their learning is still ongoing (Bell & Cowie, 2001; Black & Wiliam, 1998). It would allow lecturer to analyze the gaps in students' understanding and readiness to perform a certain task (Heritage, 2007; Sadler, 1989). By understanding these gaps, remedial learning activities can be taken to minimize (or even remove) the gaps. Most of the time, the formative assessment does not require scoring and grading. While, the summative assessment focuses with summarizing the students' learning outcomes in terms of scoring or grading to assess total learning effectiveness in the format of reporting at the end of a course of study (Bloom, Hastings, & Madaus, 1971; Sadler, 1989).

There are three essential components for formative assessment (Sadler, 1989; Shepard, 2005), namely: (1) eliciting prior knowledge, (2) providing effective feedback and (3) cultivating students' self-assessment ability. The first component, eliciting prior knowledge, is used to diagnose prior knowledge and experience to build new understanding. Prior knowledge provides a starting point for the students to understand the context that they are learning (Keeley, 2015). Eliciting prior knowledge would help the students to absorb new knowledge by making references to their own relevant knowledge and experience. The second component, feedback, is conducted by assessing the students' performance with some reference level (i.e. model answer) (Orsmond, Merry, & Callaghan, 2004; Sadler, 1989). It informs the students of their current state of learning to understand their strengths and weaknesses, so they can plan remedial actions to improve their knowledge and acquire related skills (Shutte, 2008). This kind of feedback may need to be delivered immediately to help students progress in their learning (Peat & Franklin, 2002). Lastly, the third component, cultivating students' self-assessment ability, focuses on enabling the students to evaluate their own learning at any time. It helps the students to internalize the new knowledge and identify new knowledge and skills to perform better.

Due to its enjoyable nature as well as motivation feature and immediate feedback mechanism, serious game can be used as a tool to conduct a formative assessment (Delacruz, 2011; Handfield-Jones et al., 1993; Wang, 2008). For example, a tic-tac-toe quiz game has been used to conduct formative for a certain subject (Hooshyar et al., 2016; Tsai, Tsai, & Lin, 2015); web-based quiz-game has been implemented on formative assessment in an e-learning environment (Wang, 2008); mathematics game is designed as formative assessment for after-school program (Delacruz, 2011), and game session, adapted from popular television quiz show, is conducted in hospital education to make learning more enjoyable (Howarth-Hockey & Stride, 2002). Game as a formative assessment tool would provide students with opportunities to test their skills and knowledge independently.

13.2.2 Games for SCM

Serious game has been used in SCM teaching and learning for more than seven decades since the Beer Distribution game (Jacobs, 2000; Sterman, 1989; Wisner et al., 2014) is introduced by MIT in the 1960s (Wisner et al., 2014). Beer Distribution game is the most well-known game in SCM and part of many SCM curriculum. Since then, a number of games have been introduced for different concepts of SCM, among which are Lean Leap Logistics Game (Holweg & Bicheno, 2002) and The Chain Game (Muller, Müller, Zedel, Zomer, & Engler, 2015).

In addition to those games, there are several games that introduce specific SCM concepts. Examples of those games are THINKLog (Lindawati, Nugroho, Fredericco, Rahim, & de Souza, 2017; William, Rahim, Souza, Nugroho, & Fredericco, 2018), Online Humanitarian Supply Chain (The Logistics Institute - Asia Pacific, 2017a, 2017b), Disaster Relief game (de Souza et al., 2018; The Logistics Institute - Asia Pacific, 2017a, 2017b; William, Rahim, Boo, & Souza, 2018), LogisticsRush (The Logistics Institute - Asia Pacific, 2017a, 2017b) and ACE E-Commerce game (Lindawati, Rahim, & de Souza, 2018). THINKLog is an extendable board game that can generate different scenarios for various concepts in SCM without changing the basic game structure (William, Rahim, Souza, et al., 2018), while the Online Humanitarian Supply Chain and Disaster Relief game (The Logistics Institute - Asia Pacific, 2017a, 2017b) are digital games focusing on humanitarian logistics which aims to plan and design a coordinated and uninterrupted supply chain of life-saving relief goods to the disaster-affected areas. LogisticsRush is a role-based simulation game that was designed to introduce the concept of in-mall consolidation and loading dock auction as innovative urban logistics solutions (de Souza, et al., 2016) (The Logistics Institute – Asia Pacific, 2014). While ACE E-Commerce game introduces concepts of e-commerce logistics (The Logistics Institute - Asia Pacific, 2016).

The summary and comparison of the Beer Distribution game, ACE E-Commerce game, Disaster Relief game and two other SCM games are presented in Table 13.1. This comparison is based on game categories adapted from (Abdul Jabbar & Felicia, 2015). Although most of these games have been used in SCM teaching and learning, there is no literature that discusses the impacts of these SCM games as formative assessment tools in PBL environment.

13.2.3 Problem-Based Learning

Problem-based Learning (PBL) is a pedagogy that centers on students active learning with a self-directed manner in the collaborative environment (Hmelo-Silver, 2004). The underpinning philosophy of PBL is that learning can be considered as a

Table 13.1 SCM games comparison

Category	Lean leap logistics game	The chain game	Beer distribution game	Disaster relief game	ACE E-commerce game
Game world					
Type	Role-playing simulation	Role-playing simulation	Role-playing simulation	Role-playing simulation	Role-playing simulation
Platform	Digital	Digital	Board, digital, online	Digital	Digital, network game
Technical features	Multi-players with seven core stages in the steel supply chain	Multi-players with five roles	Multi-players with four roles	Single role	Two modes of play: Single player and multi-players with three roles
Game event					
Subject or content areas	Automotive steel supply chain	Supply chain visibility and chain control concepts	Industrial supply chain	Humanitarian logistics	E-commerce logistics
Scenario	One scenario; two products (red and blue) and six stages	One scenario on an international supply chain	One scenario; single product in a four-level supply chain	Three scenarios with three fictitious island maps	One scenario on last mile deliveries of e-commerce goods
Learning objective	Create supply chain awareness and develop and validate improvements' to the steel supply chain	Introduce the collaborative concepts between supply chain partners to increase supply chain visibility	Introduce the basic concepts of the bullwhip effect and the benefits of information sharing	Introduce the importance and complexity of humanitarian relief	Introduce important criteria in last-mile logistics for e-commerce

"constructive, self-directed, collaborative and contextual" activity (Dolmans, De Grave, Wolfhagen, & van der Vleuten, 2005).

In a typical PBL lesson, students are triggered to discover and explore the new knowledge starting with the problem analysis collaboratively with their peers, under the guidance of a lecturer though scaffolding. Formative assessment is done by the end of each lesson with the submission of a reflective journal by each student and feedback by the lecturer to each student. During the problem analysis, students are supposed to connect the concepts with their individual or collective prior knowledge or experience. Students could explore the resources either individually or jointly with a team discussion, with scaffolding and continuous observation from the lecturer during the entire learning process. To close the loop, students need to share their findings as well as their application of new knowledge acquired in solving real industry problems. Peer review and scaffolding are conducted to guide the discussion (Yew & Goh, 2016).

In summary, the main characteristics of PBL are (1) the learning process is triggered by an open-ended problem; (2) students are engaged in learning both independently and collaboratively; (3) lecturer is facilitating the learning process through continuous scaffolding and formative assessment; (4) learning outcomes are presented by the students in the reporting phase; and (5) reflection plays a key role to ensure students to learn effectively for the long term (Schmidt, Rotgans, & Yew, 2011).

13.2.4 PBL Implementation in Republic Polytechnic

PBL has been implemented in Republic Polytechnic (RP) to create an active and engaging classroom setting where students need to analyze problems, think critically and develop solutions while working with a small team (Republic Polytechnic, 2019a, 2019b). A typical one-day PBL implementation requires students to work on a particular problem during three learning phases and two study periods, as summarized in Table 13.2. Typical class size is 20–25 students, and they are further grouped in a team of 4–5. Students work in different groups each day with different lecturers for different modules.

During the first learning phase, a problem is used as the trigger to kick off the lesson, and students work in a group of 4–5 to participate in the learning process by starting with analyzing the problem and learning issues. Throughout the entire first study period, second learning phase and second study period, learning resources are provided for continuous formative assessment to scaffolding the learning. The lecturer facilitates the students' learning by assessing their current skills and knowledge, guiding them to discover the new knowledge to be acquired, and providing regular feedback to ensure that they are on the right track during the learning process. Students are given the opportunity to construct their own knowledge in a team as well as through self-directed learning and self-assessment.

Table 13.2 Typical PBL implementation in RP (Yew & O'Grady, 2012)

	Duration	Key learning activities
First learning phase	1 h	• Exploration and analysis of problem and learning issues
First study period	45 min	• Self-directed research and collaborative learning • Self-assessment
Second learning phase	1.5 h	• Formulation of responding to problems and overcoming of learning obstacles • Lecturer's guidance and feedback
Second study period (including lunch break)	1.5 h	• Consolidation of ideas in a team • Finalization of responses to the problem
Third learning phase	2 h	• Group presentation and critique • Lecturer's feedback and summary of learning issues

During the third learning phase, students present their findings and propose solutions to the class. Critique and feedback from the rest of the class and the lecturer are given during the presentation. By the end of the learning, the lecturer needs to wrap up the lesson by giving the final presentation containing the learning content and recommended solutions for the day.

By the end of the lesson day, students are supposed to submit peer evaluation on their team members' performance throughout the day, self-evaluation of their own performance and a reflective journal to reflect on the daily learning process as well as learning outcomes for formative assessment. Within 3 days from the lesson day, students will be awarded a Continuous Assessment (CA) Grade, which is assessed by the lecturer by taking into account students' learning process, participation in both class and team, team presentation and the quality of reflections. This is a kind of formative assessment that could also be based on students' general behavior (punctuality, attentiveness, effort, teamwork, peer support, etc.) and performance (such as learning process, ability to articulate, explain and defend their solutions).

13.3 Research Model and Hypotheses

To effectively embed serious game as a formative assessment tool in PBL learning process, we need to consider both the pedagogy components (i.e. motivate self-learning and self-assessment) as well as entertainment components (i.e. positive game experience and flow). Hence, the game needs to balance two main criteria, namely: learning objective and game experience. For that purpose, we adopt the game-based learning framework (Van Staalduinen & de Freitas, 2011; William, Rahim, Souza, et al., 2018) to fit the PBL environment, as illustrated in Fig. 13.1.

Our research model and hypotheses (as illustrated in Fig. 13.2) are derived based on this game design framework. For this research model, we review the literature for game-based learning, game-based assessment and PBL to identify important

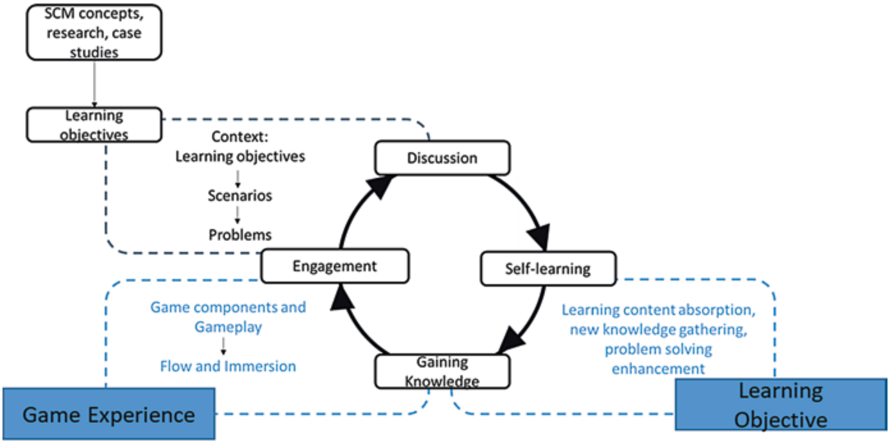


Fig. 13.1 Game design framework for PBL environment

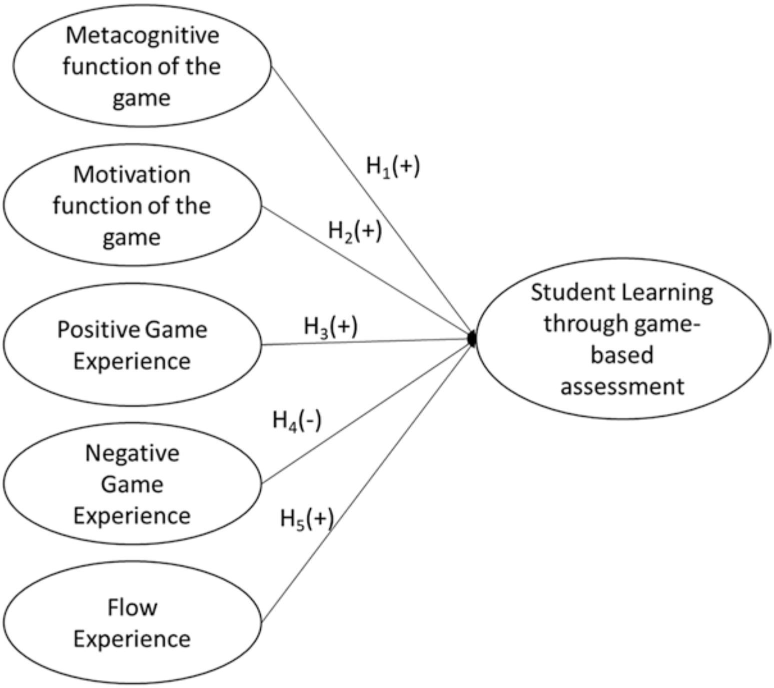


Fig. 13.2 Research model

components for learning objective and game experience criteria. For learning objective criterion, the two identified components are metacognitive function and motivation function. While, for game experience criterion, we identified three components, namely: positive game experience, negative game experience and flow experience.

Metacognitive function, motivation function, positive game experience and flow experience are considered as encouragement components to improve students' perceived learning. While negative game experience would hinder the students from learning.

13.3.1 Learning Objective Criterion

In PBL, serious game would initiate problem-solving and formative assessment processes. Serious games could also be a scaffolding tool to let students experience the real problems in the simulated game environment. Game would provide a risk-free environment for the students to repeatedly apply that knowledge in a risk-free simulation world and to perform self-assessment. Hence, the students' perceived learning using game-based assessment can be evaluated based on the cognitive components of problem-solving, which can be divided into two components, namely: (1) metacognitive functions and (2) motivation (Surgrue, 1995). The components are summarized in Table 13.3. Using these components, we hypothesize the following:

Hypothesis 1: Metacognitive function of the game would improve students' perceived learning through game-based formative assessment

Hypothesis 2: Motivation function of the game would improve students' perceived learning through game-based formative assessment

13.3.2 Game Experience Criterion

Positive game experience would be able to provide an enjoyable experience to improve their skills and knowledge for completing those tasks and challenges in the game environment (Hamari et al., 2016; Hou, 2015). While negative game experience would discourage the students to play and learn from the game. Thus, we posit:

Hypothesis 3: Positive game experience would improve students' perceived learning through game-based formative assessment

Table 13.3 Learning objective evaluation—variables and mechanism

Metacognitive function	Motivation
Variables	
Planning and monitoring	Perceived self-efficacyPerceived attraction of the tasks
Translation in game environment	
Students can control the game and receive feedback in the game environment	Motivation for learning

Hypothesis 4: Negative game experience would hinder the students from learning through game-based formative assessment

Flow experience is described as a student' state of complete absorption in the game (Kiili, 2006). It motivates the students to win the game by achieving new skills or understanding new concepts voluntarily at a greater speed. It is used as a formative assessment to constantly encourage the students to build their understanding or skills to match the goal of the game. Thus, we hypothesize the following:

Hypothesis 5: Flow experience would improve students' perceived learning through game-based formative assessment

13.4 Research Design

13.4.1 Questionnaire Design

Based on the research model and hypotheses, we develop a questionnaire comprising 12 items (Table 13.4). We identify two items for metacognitive function, three items for motivation function, four items for positive experience, two items for negative experience and one item for flow experience. The questionnaire uses the five-point Likert scale (5 = strongly agrees, 1 = strongly disagrees). Other than these items, we also add one item to determine the students' learning that serves as a self-assessment score to the students' perceived learning.

13.4.2 Game Sessions

We conducted four game sessions from October 2018 to January 2019 with different groups of students for each session. In the first session, we played Beer Distribution game, while in the other three sessions, we played ACE E-Commerce game and Disaster Relief game. The first session was an actual PBL class for SCM course which was attended by 119 students. While the three other game sessions were organized specifically to evaluate the effectiveness of ACE E-Commerce game and Disaster Relief game. These three sessions were attended by 30 students. The summary of the game sessions is shown in Table 13.5.

At the end of each game session, we distributed the questionnaire. For the first session, we received 119 complete responses. For the three other sessions, we only managed to get 22 complete responses (out of 30 students) for evaluating both ACE E-Commerce game and Disaster Relief game. In addition to the questionnaire, we also evaluate the students' learning by gathering their learning points. We asked the students to write down their learning points for each game after they have completed the game. Each student is able to list down one to three learning points. We then matched the students' learning points with the intended learning objectives of the games.

Table 13.4 Questionnaire to evaluate the game experience

#	Items	Category	Citation/reference
1	The student has a feeling of control over his/her actions	Metacognitive function—control	Kelly (2010)
2	The student is aware of the impact of his/her actions	Metacognitive function—feedback	Kelly (2010)
3	The student was challenged but believed that his/her skills would allow him/her to meet the challenge	Motivation function	–
4	The student is motivated to ask and discuss the learning concept	Motivation function	IJsselsteijn, de Kort, and Poels (2013)
5	The student thinks that the learning points from the game are interested	Motivation function	–
6	The game is fun and interesting	Positive experience	IJsselsteijn et al. (2013)
7	The student wants to play the game again	Positive experience	IJsselsteijn et al. (2013)
8	The goals of the game were clearly defined	Positive experience	–
9	The student could understand the rules of the game spontaneously and automatically without having to think	Positive experience	
10	The game is boring	Negative experience	IJsselsteijn et al. (2013)
11	The student does not learn anything from the game	Negative experience	IJsselsteijn et al. (2013)
12	The student was deeply involved and engaged in the game	Flow experience	Kelly (2010)

Table 13.5 Summary of the game sessions

#session	Game used	Game duration	Number of participants	Number of complete responses
1	Beer distribution game	60 min	119	119
2	ACE E-commerce game and disaster relief game	60 min (for each game)	10	7
3	ACE E-commerce game and disaster relief game	60 min (for each game)	10	7
4	ACE E-commerce game and disaster relief game	60 min (for each game)	10	8

Brief introductions about these three games and the students' activities during the games are described as follows.

13.4.2.1 Beer Distribution Game

Beer Distribution game is a well-known game in SCM since it was introduced by MIT in the 1960s (Wisner et al., 2014). It has been used and expanded into several variants including the stationary beer distribution game (Chen, 2000), computer

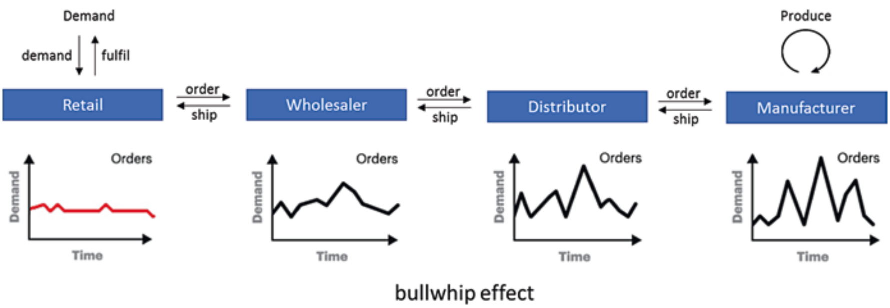


Fig. 13.3 Beer distribution game roles and game flow

simulated beer distribution game (e.g. Kaminsky & Simchi-Levi, 1998; Simchi-Levi, Simchi-Levi, & Kaminsky, 1999) and online beer distribution game (Jacobs, 2000). The Beer Distribution game simulates a single product supply chain, involving a four-level supply chain composed of a retailer, a wholesaler, a distributor and a manufacturer as illustrated in Fig. 13.3. Each supply chain level follows the same set of activities, which are (1) fulfill the demand from customers, (2) order from their supplier and (3) manage the inventories. It introduces the basic concepts of the bullwhip effect in the supply chain due to the variation in customer's demand. The students need to minimize their total supply chain costs.

Beer Distribution game has been adopted in a core module of the SCM course in RP. Hence, the first game session was conducted as part of the actual SCM course. It was the first lesson for this course and served as the first problem in this module. The primary objective of using this game was to get students to recap the major components of a typical supply chain and realize the importance of SCM by letting them experience the supply chain dynamics processes.

In the game session, each student had to take up a role in this game as either a retailer, a wholesaler, a distributor or a manufacturer. They were encouraged to try out any ordering methods that they learned previously, without any communication among the four of them. Direct involvement in the game environment allowed students to experience all the typical things (frustration, overstock, stock-outs, suppliers could not supply in time, etc.) that are real and happening in the industry. During the game session, the students received feedback based on their actions.

After the first round, the lecturer conducted a formal scaffolding process and debriefed the students on what they have experienced in the game, what they have done wrong, what good decisions they have made. It would enable students to perform self-assessment to realize their mistakes and good practices in SCM processes.

Students were to be given a second chance to play the same game with different settings (i.e. shorter lead time, visible Point-Of-Sales info) with the intent to internalize their understanding of the key factors that make differences in supply chain performance. In the second round, students were making their decisions based on their interpretation of information available to them in the game dashboard. Students had also observed their actions, monitored their inventory status closely and interpreted the results and plots to reflect what they have done.

13.4.2.2 ACE E-Commerce Game

ACE E-Commerce is a multi-role based interactive digital game developed to embed the basic concepts of e-commerce logistics (Lindawati, Rahim, & de Souza, 2018). The game provides a fictitious game map (as illustrated in Fig. 13.4) to simulate the processes in e-commerce logistics and test different strategies and e-commerce logistics business models. Learning objectives in ACE E-Commerce game are (1) identifying the process involved in planning, (2) understanding the effective and efficient flow of goods and services from the point of origin to the point of consumption and (3) identifying different criteria between a traditional logistics and an e-commerce logistics.

The game has three playable roles, namely: merchant, e-retailer and Logistics Service Provider (LSP). Each role has its own set of activities and tasks as illustrated in Fig. 13.5. For example, the merchant's main goal is to produce and maintain a healthy level of inventories to meet its customers demand in a timely manner. Depending on the type of request, the merchant may need to deliver the goods to the e-retailers' warehouse or directly to the end-customers (drop shipping). Merchants may also use their own vehicles or engage the service from LSPs for the deliveries. On the other hand, e-retailer receives online orders from end-customers. Depending on the business model selected, e-retailers may need to handle their own inventory, warehouse and/or deliveries, or request the supplier—in this case, the merchant—to do the deliveries on its behalf. While, in this game, LSP needs to provide transportation services (deliveries) for the merchants or the e-retailers. The summary of activities for each role is shown in Table 13.6.

In the three game sessions, we only played ACE E-Commerce game for one round. Each student had to take up a role in this game as either a merchant, an

Fig. 13.4 ACE e-commerce game map

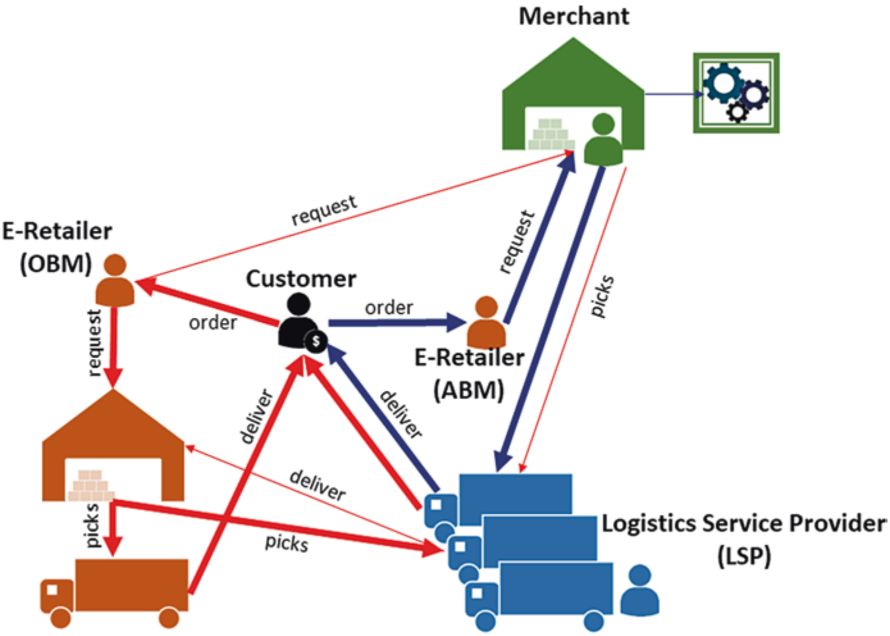


Fig. 13.5 Activity flow in ACE e-commerce game

Table 13.6 Summary of role activities in ACE e-commerce game

Activities	Merchant	LSP	E-retailer
Deliver to end-customer	Yes	Yes	Yes
Create auction for delivery to LSP	Yes	No	No
Sell vehicle services	No	Yes	No
Manufacture products	Yes	No	No
Receive orders from the end-customer	No	No	Yes
Set prices for products	Yes	No	Yes
Purchase vehicle depots	Yes	By default	Yes
Purchase/rent warehouses	Yes	No	Yes
Purchase CDP permit	Yes	Yes	No
Purchase foreign business license	Yes	By default	No
Purchase export permit	Yes	By default	No
Purchase/sell vehicles	Yes	Yes	Yes

e-retailer or an LSP. Similar to the Beer Distribution game session, they were encouraged to try out any ordering methods that they learned previously without any communication with other students. Students experienced the consequences of their strategies in the overall dynamic game environment. These consequences were communicated through continuous feedback in the game environment.

13.4.2.3 Disaster Relief Game

The Disaster Relief Game is developed to enhance the learning experience for humanitarian logistics (de Souza et al., 2018; William, Rahim, Boo, & Souza, 2018). The game was first developed in 2015 based on various case studies in humanitarian logistics (for example: (The Logistics Institute - Asia Pacific, 2013, 2014, 2015)). While the case studies were based on actual disasters and raw data, the game uses fictitious maps as illustrated in Fig. 13.6. The game aims to help students refining their thought process to plan and design a coordinated and uninterrupted supply chain of life-saving relief goods to the affected areas, including cargo and information flow by assessing, sourcing (stocks and procurement), coordinating transportation, warehousing and finally distributing the goods.

There are two phases in this Disaster Relief game, the planning phase and execution phase. Each phase is set to 30 turns. During the planning phase, students are expected to plan various strategies to develop a rescue and resource allocation plan for an imminent disaster that will strike the island for a period of 30 turns. In their plan, students have to choose an appropriate location to transport as many survivors as possible away from the disaster area, as well as the number of relief goods to be sent to this location. They also have to decide on the number and type of transport vehicles in order to deliver these relief goods. All of this needs to be done by considering the limited budget and time period. During the execution phase, a disaster will strike at the expected area, and the game will then execute the plan made by the students in the planning phase. The mission will simulate the effectiveness and resilience of the plan during a disaster. During this phase, random emergency events are also introduced at different turns in the game. This may affect the students' ini-

Fig. 13.6 Fictitious map in Disaster Relief game

tial plan and may require the students to react accordingly. The game ends immediately if there are no more survivors or there are no more funds available during the execution phase.

There are four intended learning objectives for this game. First, it helps the students to understand the importance and complexity in providing an uninterrupted supply of data and goods during humanitarian relief missions. Second, it shows how lead time affects the delivery time of ordering supplies, which can lead to loss of lives. Third, the game helps to understand the humanitarian logistics process, including constraints like working with limited time, budget and resources. And lastly, it represents the urgency of providing supplies as the condition of the incident area affects the mortality rate.

Although it has never been used in PBL environment, this Disaster Relief game has been played in various workshops in Singapore as well as in the South East Asia region. In most of the game sessions, this game was used to teach humanitarian logistics concepts to government officials for humanitarian disaster relief agencies (de Souza et al., 2018).

In the three game sessions, similar to ACE E-Commerce game sessions, we played Disaster Relief game for one round. Each student had to take up a role as a humanitarian agency. In the planning stage, the students were encouraged to plan humanitarian logistics strategies based on what they learned previously. While in the implementation stage, the students experienced the impacts of their plan when a disaster occurs. The continuous feedback based on the students' actions/strategies in the game environment were provided.

13.5 Game Evaluation

13.5.1 Questionnaire Result

Based on the students' responses gathered after the game sessions, we tabulate the summary as shown in Fig. 13.7. The results show that the average scores for both metacognitive function and motivation function for these three games are above 3.31. Although the scores for ACE E-Commerce and Disaster Relief games are lower than the scores for the Beer Distribution game, there is no significant difference between them. The results suggest that the students are both motivated by the game (motivation function) as well as able to plan and control the games (metacognitive function).

For game experience, the results show that the average scores for positive experience and flow experience for these three games are above 3.20. It indicates that the students have good experience with the games and able to engage with the games. The score for negative experience for the ACE E-Commerce game is 2.70 which is

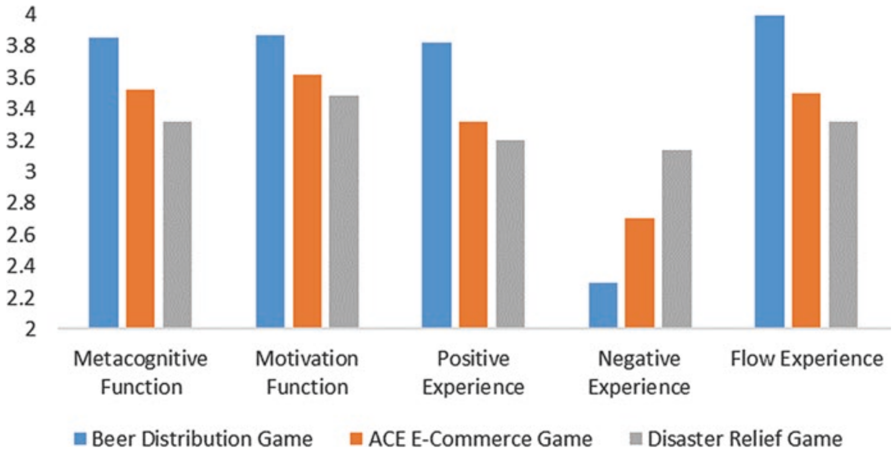


Fig. 13.7 Questionnaire result for each component and each game

above the negative experience score for the Beer Distribution game (score: 2.29). It is considered a good result as it indicates that the students were able to enjoy the game. However, the score for negative experience for Disaster Relief game is rather high (score: 3.13). This can be caused by the inherent complexity in the rules that creates difficulties for the students to understand and play the game (William, Rahim, Boo, & Souza, 2018).

The research model and hypotheses are calculated using linear regression in Ms. Excel 2016. The linear regression result is summarized in Table 13.7. The results show significant relationships between both item 4 and 6 to students' perceived learning ($p < 0.05$ for item 6 and $p < 0.01$ for item 4). While the other items do not have significant influence ($p > 0.05$). This indicates that these games significantly provide motivation and a positive learning experience that improve students' perceived learning. Thus, we conclude that only Hypothesis 2 and 3 are supported.

13.5.2 Students' Learning Points

Other than the questionnaire, we also determine the students' learning by matching the students' learning points, collected after the game sessions, with the intended learning objectives for each game. For most of the students, the learning points are aligned with the intended learning objectives for each game. This indicates that the students are able to absorb the learning objectives by using these games as formative assessment tools. The top five learning points for each game are summarized in Table 13.8.

Table 13.7 Linear regression result

Model	Participation decision	
	Coefficient	Adj. R^2
Constant	−0.0509	0.7011
Item 1: The student has a feeling of control over his/her actions.	0.1458	
Item 2: The student is aware of the impact of his/her actions	0.2257	
Item 3: The student was challenged but believed that his/her skills would allow him/her to meet the challenge.	0.3262	
Item 4: The student is motivated to ask and discuss the learning concept	0.6846**	
Item 5: The student thinks that the learning points from the game are interested	−0.2018	
Item 6: The game is fun and interesting	0.3793*	
Item 7: The student wants to play the game again	0.1015	
Item 8: The goals of the game were clearly defined	0.0736	
Item 9: The student could understand the rules of the game spontaneously and automatically without having to think	−0.2008	
Item 10: The game is boring	−0.1687	
Item 11: The student does not learn anything from the game	0.1827	
Item 12: The student was deeply involved and engaged in the game	−0.1223	

*$p < 0.05$
**$p < 0.01$

13.6 Discussion and Recommendation

The questionnaire results show that there is a significant correlation between two components, namely: motivation and positive game experience with the students' perceived learning. While other components do not show a significant correlation to the students' perceived learning. These results indicate that these games support the learning activities mainly by providing motivation and an enjoyable environment to encourage students to learn specific concepts. These are aligned with the prior studies on serious game benefits (Hou, 2015; Ma et al., 2011). With these motivation function and enjoyable learning environment, games help to provide a seamless formative assessment process using feedback and self-assessment without students' realizing it. It provides a learning environment where students can try their ideas, knowledge, skills and experience and improve their knowledge and skills through constant feedback.

Although the results show that the games have great potential to be incorporated into PBL classroom as formative assessment, a few adjustments in the games may be needed for accommodating the formative assessment purposes in PBL environment. These adjustments aim to support three essential components for formative assessment (i.e. eliciting prior knowledge, providing effective feedback and cultivating students' self-assessment) as mentioned in Sect. 13.2.1.

The first adjustment is on the number of game sessions. We need to conduct at least two game sessions with different scenarios to emphasize the formative

Table 13.8 Students' learning points

Learning points
Beer distribution game
The importance of communication in supply chain network
Bullwhip effect
Backlog management
Teamwork in supply chain network
Inventory management
Disaster relief game
The importance of disaster relief good transportation to save lives
Emergency planning
Demand forecasting
Resource management
Inventory management
ACE E-commerce
Supply chain planning for E-commerce
The importance of demand forecasting
Price comparison
Lead time
The importance of scheduling and routing

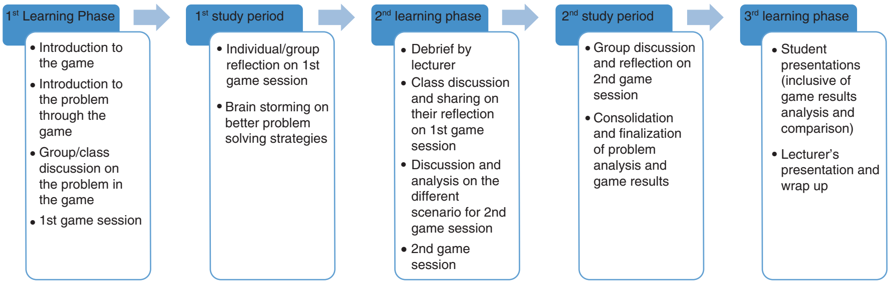


Fig. 13.8 Game setup for PBL environment

assessments at different learning phases, as summarized in Fig. 13.8. It is following the PBL time table in RP as described in Table 13.2. Beer Distribution game has been implemented in RP in this manner. The first session is conducted in the first learning phase. This session would start with the game introduction and exploration. The game would introduce a particular problem based on the real situation, and the students would need to try out any methods that they learned previously or even their gut feeling to solve the problems. The game would encourage the students to elicit their prior knowledge and implement it as their strategies in the game environment. Changes in the game due to the students' actions would serve as effective feedback for the students.

The second game session will be conducted in the second learning phase. In this session, the game should be played with different scenarios to test the students' understanding of the problems and relevant strategies that can be implemented. Different settings (i.e. shorter lead time and visible Point-Of-Sales info) can be used to let students experience the difference and the effect of those settings in the game. In the second session, students would be able to verify their strategies based on the new learning points from the games. They should be able to observe the different performance between the first and second sessions. They can further discuss their strategies and their performance differences in the second study period. It serves as a continuous formative assessment where the students can further assess their own knowledge and skills in a particular game environment. It would cultivate the students' self-assessment ability.

The second adjustment is on the game's feedback mechanism. To provide an effective self-assessment tool, the game should be able to store the students' actions and results when they play the game. It would help them to conduct self-assessment and evaluate their strategies and methods. They can also identify necessary actions for a specific problem. This feature also enables the lecturer to easily spot the mistakes or troubleshoot wherever necessary. In addition to that, the game should also be able to summarize each student's results as well as individual team's results in a different data formats such as plots and graphs, numbers and tables. This would benefit different learners.

The third adjustment is on the lecturer's support functionality. In the PBL classroom, the lecturer would need to guide the students and scaffold their understanding of a particular concept in the game environment. As such, lecturer should be allowed to show the comparison of the results among individual students or teams for an effective debriefing after each game run.

13.7 Summary and Future Works

In this chapter, we focus our work on evaluating the effectiveness of three SCM games as formative assessment tools in a PBL environment. We use the research model and hypotheses based on two criteria, namely: learning objective and game experience. We use this research model to evaluate three digital SCM games, namely: Beer Distribution game, ACE E-Commerce game and Disaster Relief game. The Beer Distribution game introduces a generic flow of product and information in the supply chain, while Disaster Relief game and ACE E-Commerce game focus on more specific SCM topics.

Four game sessions were conducted to evaluate the games. The first session used Beer Distribution game, while the other three sessions used both ACE E-Commerce game and Disaster Relief game. The questionnaire results suggest two components, namely: motivation function and positive game experience significantly help to improve the students' perceived learning. The three games as formative assessment tools are able to form the students' understanding about concepts in SCM through continuous motivation and enjoyable learning environment to enable voluntary learning. In terms of the game experience evaluation, the results indicate that the students are able to enjoy these three games. Although the negative experience for Disaster Relief is relatively high due to the inherent gameplay complexity, we believe that this negative experience can be reduced by providing more time for the students to play the game and enhancing the game visualization by embedding Mixed Reality (MR) into the game (William, Rahim, Boo, & Souza, 2018).

Nonetheless, we see two possible extensions that we would like to study in the near future. First, we would like to test our research model and hypotheses in another learning environment (such as social learning environment) to evaluate the effectiveness of game-based formative assessment in a different learning environment. We would like to evaluate if motivation and positive game experience have significant impacts to improve the students' perceived learning as in the PBL environment. Second, we would like to conduct comprehensive game analytics based on the actions that the students take in the games and feedback that they received. It would help us to understand the students' learning behavior to provide a better formative assessment for them.

References

Abdul Jabbar, A., & Felicia, P. (2015). Gameplay engagement and learning in game-based learning: A systematic review. *Review of Educational Research, 85*(4), 740–779.

Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education, 85*, 536–553.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Bloom, B., Hastings, J., & Madaus, G. (1971). *Handbook on the formative and summative evaluation of student learning*. New York, NY: McGraw-Hill.

Burguillo, J. (2010). Using game theory and competition-based learning to stimulate student motivation and performance. *Computers & Education, 55*(2), 566–575.

Chen, F. A. (2000). The stationary beer game. *Production and Operations Management, 9*(1), 19–30.

de Freitas, S. (2006). Learning in immersive worlds: A review of game-based learning. London: Joint Information Systems Committee.

de Freitas, S., & Liarokapis, F. (2011). Serious Games: A New Paradigm for Education? In Serious Games and Edutainment Applications (pp. 9–21). London: Springer.

de Souza, R., Lau, H., Goh, M., Lindawati, Ng, W., & Tan, P. (2016). Retail precinct management: A case of commercial decentralization in Singapore. *Transportation Research Procedia, 12*, 765–776.

de Souza, R., William, L., Timperio, G., & Rahim, Z. B. (2018). Simulation-based serious gaming in humanitarian logistics. In *Winter Simulation Conference 2018*. Sweden.

Delacruz, G. C. (2011). *Games as formative assessment environments: Examining the impact of explanations of scoring and incentives on math learning, game performance, and help seeking*. Los Angeles: CRESST Report 796. National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Dolmans, D., De Grave, W., Wolfhagen, I., & van der Vleuten, C. (2005). Problem-based learning: Future challenges for educational practice and research. *Medical Education, 39*(7), 732–741.

Hamari, J., Shernoff, D., Rowe, E., Coller, B., Asbell-Clarke, J., & Edwards, T. (2016). Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Computers in Human Behavior, 54*, 170–179.

Handfield-Jones, R., Nasmith, L., Steinert, Y., & Lawn, N. (1993). Creativity in medical education: The use of innovative techniques in clinical teaching. *Medical Teacher, 15*(1), 3–10.

Heritage, M. (2007). Formative assessment: What do teachers need to know and do? *Phi Delta Kappan, 89*(2), 140–145.

Hmelo-Silver, C. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review, 16*(3), 235–266.

Holweg, M., & Bicheno, J. (2002). Supply chain simulation - A tool for education, enhancement and endeavour. *International Journal of Production Economics, 78*(2), 163–175.

Hooshyar, D., Ahmad, R., Yousefi, M., Fathi, M., Horng, S., & Lim, H. (2016). Applying an online game-based formative assessment in a flowchart-based intelligent tutoring system for improving problem-solving skills. *Computers & Education, 94*, 18–36.

Hou, H. (2015). Integrating cluster and sequential analysis to explore learners' flow and behavioral patterns in a simulation game with situated-learning context for science courses: A video-based process exploration. *Computers in Human Behavior, 48*, 424–435.

Howarth-Hockey, G., & Stride, P. (2002). Can medical education be fun as well as educational? *British Medical Journal, 325*, 1453–1454.

IJsselsteijn, W., de Kort, Y., & Poels, K. (2013). *The game experience questionnaire*. Eindhoven, Netherlands: Technische Universiteit Eindhoven.

Jacobs, F. (2000). Playing the beer distribution game over the internet. *Production and Operations Management, 9*(1), 31–39.

Kaminsky, P., & Simchi-Levi, D. (1998). A new computerized beer game: A tool for teaching the value of integrated supply chain management. *Global Supply Chain and Technology Management, 1*(1), 216–225.

Keeley, P. (2015). *Science formative assessment: 75 practical strategies for linking assessment, instruction and learning*. Thousand Oaks, CA: Corwin Press.

Kelly, H. (2010). How games and simulations can help meet America's challenges in science mathematics and technology education. In *Serious game design and development: Technologies for training and learning* (pp. 117–133). Hershey, PA: IGI Global.

Kiili, K. (2005). Digital game-based learning: Towards an experiential gaming model. *The Internet and higher education, 8*(1), 13–24.

Kiili, K. (2006). Evaluations of an experiential gaming model. *Human Technology, 2*(2), 187–201.

Kim, B., Park, H., & Baek, Y. (2009). Not just fun, but serious strategies: Using meta-cognitive strategies in game-based learning. *Computers & Education, 52*(4), 800–810.

Knight, J., Carly, S., Tregunna, B., Jarvis, S., Smithies, R., de Freitas, S., … Dunwell, I. (2010). Serious gaming technology in major incident triage training: A pragmatic controlled trial. *Resuscitation, 81*(9), 1174–1179.

Kumar, D. (2000). Pedagogical dimensions of game playing. *ACM Intelligence Magazine, 10*(10), 9–10.

Lindawati, Nugroho, E., Fredericco, R., Rahim, Z. B., & de Souza, R. (2017). ThinkLog: Interactive learning for supply chain management. In *IEEE 6th International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*. Hong Kong.

Lindawati, Rahim, Z. B., & de Souza, R. (2018). Using Interactive Digital Gaming to Facilitate Learning in e-Commerce Logistics. Conference on *The Fourteenth International Conference on eLearning for Knowledge-Based Society*. Bangkok, Thailand.

Liu, Y., Alexandrova, T., & Nakajima, T. (2011). Gamifying intelligent environments. In *Proceedings of the 2011 International ACM Workshop on Ubiquitous Meta User Interfaces*. Scottsdale, AZ.

Ma, M., Oikonomou, A., & Jain, L. (2011). Innovations in serious games for future learning. In *Serious games and edutainment applications* (pp. 3–7). London, UK: Springer.

Muller, T., Müller, R., Zedel, K., Zomer, G., & Engler, M. (2015). Enhancing awareness on the benefits of supply chain visibility through serious gaming. In *Conference on e-Business, e-Services and e-Society* (pp. 503–512). Cham, Switzerland: Springer.

Orsmond, P., Merry, S., & Callaghan, A. (2004). Implementation of a formative assessment model incorporating peer and self-assessment. *Innovations in Education and Teaching International, 41*(3), 273–290.

Peat, M., & Franklin, S. (2002). Supporting student learning: The use of computer–based formative assessment modules. *British Journal of Educational Technology, 33*(5), 515–523.

Republic Polytechnic. (2019a). Author. Retrieved January 03, 2019, from https://www.rp.edu.sg/discover/

Republic Polytechnic. (2019b). *The RP experience*. Author. Retrieved January 03, 2019, from https://www.rp.edu.sg/about-us/who-we-are/the-rp-experience

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.

Riedel, J., & Hauge, J. (2011). State of the art of serious games for business and industry. In *2011 17th International Conference on Concurrent Enterprising (ICE)*.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144.

Schmidt, H., Rotgans, J., & Yew, E. (2011). The process of problem-based learning: What works and why. *Medical Education, 45*(8), 792–806.

Shepard, L. A. (2005). Linking formative assessment to scaffolding. *Educational Leadership, 63*(3), 66–70.

Shutte, V. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.

Simchi-Levi, D., Simchi-Levi, E., & Kaminsky, P. (1999). *Designing and managing the supply chain: Concepts, strategies, and cases*. New York, NY: McGraw-Hill.

Sterman, J. (1989). Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Science, 35*(3), 321–339.

Surgrue, B. (1995). A theory-based framework for assessing domain-specific problem solving ability. *Educational Measurement: Issues and Practice, 14*(3), 29–36.

The Logistics Institute - Asia Pacific. (2013). *Emergency preparedness and performance measurement in humanitarian organizations in Asia*. Singapore, Singapore: Author.

The Logistics Institute - Asia Pacific. (2014). *Typhoon Yolanda relief response report*. Singapore, Singapore: Author.

The Logistics Institute - Asia Pacific. (2015). *Disaster relief supply chains: Addressing challenges in robustness and resilience to enable efficiency and effectiveness in humanitarian response*. Singapore, Singapore: Author.

The Logistics Institute - Asia Pacific. (2016). *E-commerce trends and challenges: A logistics and supply chain perspective*. Singapore, Singapore: Author and National University of Singapore.

The Logistics Institute - Asia Pacific. (2017a, 11). *Resource allocation and management in disaster relief - Innovative learning through serious games series*. Singapore, Singapore: Author and National University of Singapore. Retrieved 12 26, 2018, from http://www.tliap.nus.edu.sg/PDF/whitepapers/Vol17-Nov-SG_Res%20Alloc_n_Mgmt%20in%20DR.pdf

The Logistics Institute - Asia Pacific. (2017b). *Serious game video trailer*. Retrieved December 26, 2018, from http://www.tliap.nus.edu.sg/outreach/sg_trailer.html

The Logistics Institute – Asia Pacific. (2014). *Challenges and opportunities in clustered urban logistics: The Jurong Gateway precinct*. Singapore, Singapore: Author and National University of Singapore.

Tsai, F. H., Tsai, C. C., & Lin, K. Y. (2015). The evaluation of different gaming modes and feedback types on game-based formative assessment in an online learning environment. *Computers & Education, 81*, 259.

Van Staalduinen, J., & de Freitas, S. (2011). A game-based learning framework: Linking game design and learning. In M. Khine (Ed.), *Learning to play: Exploring the future of education with video games* (pp. 29–54). New York, NY: Peter Lang.

Wang, T. H. (2008). Web-based quiz-game-like formative assessment: Development and evaluation. *Computers & Education, 51*(3), 1247–1263.

William, L., Rahim, Z., Souza, R., Nugroho, E., & Fredericco, R. (2018). Extendable board game to facilitate learning in supply chain management. *Advances in Science, Technology and Engineering Systems Journal, 3*(4), 99–111.

William, L., Rahim, Z. B., Boo, I., & de Souza, R. (2018). Embedding mixed reality in humanitarian logistics gaming. In *IEEE International Conference on Teaching, Assessment, and Learning for Engineering (IEEE TALE)*. Wollongong, NSW.

Wisner, J., Tan, K., & Leong, G. (2014). *Principles of supply chain management: A balanced approach*. Boston, MA: Cengage Learning.

Yew, E., & Goh, K. (2016). Problem-Based Learning: An Overview of its Process and Impact on Learning. *Health Professions Education, 2*(2), 75–79.

Yew, E., & O'Grady, G. (2012). One-day, one-problem at Republic Polytechnic. In One-day, one problem: An approach to problem-based learning (pp. 1–16). Singapore: Springer-Verlag.

Chapter 14
What Does Exploration Look Like? Painting a Picture of Learning Pathways Using Learning Analytics

José A. Ruipérez-Valiente, Louisa Rosenheck, and Yoon Jeon Kim

14.1 Introduction

To prepare students for success in our ever-changing knowledge economy, learning and teaching is moving toward valuing future-ready skills, also called twenty-first century skills or soft skills, including skills like problem-solving, interpreting information, and communication. Educational games and simulations are one important tool for building these kinds of skills. They can provide open-ended but scaffolded experiences in which students can test out ideas in a low-stakes environment, retrying levels or challenges until they succeed. However, one challenge in teaching interpersonal skills in general, and teaching them through games in particular, is that these skills are much harder to assess than the content knowledge and procedural skills that have been valued by schools in the past. While digital games for learning do have the affordance of being able to collect large amounts of very nuanced activity data, the most common types of analysis and reporting have most often focused on things like content-specific failures and successes, percent of correct attempts, and progress through the game. Dashboards or tools that count these achievements for each student and output certain metrics for teachers to track progress are a helpful starting place, and certainly convenient for teachers. However, richer metrics of game-based learning have the potential to not only show what students understand and can do, but also provide much deeper insights into patterns in how they are approaching problems and illuminate the different learning pathways they take. Learning analytics techniques that use generalizable methods but that are tailored to the specific game mechanics and assessment mechanics of a given game can play a key role in mapping out the learning pathways that students take and characterize the ways they are engaging with the game. As such, this chapter will use examples of measurement and data analysis methods from *The Radix*

J. A. Ruipérez-Valiente (✉) · L. Rosenheck · Y. J. Kim
Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: jruipere@mit.edu; louisa@mit.edu; yjk7@mit.edu

D. Ifenthaler, Y. J. Kim (eds.), *Game-Based Assessment Revisited*, Advances in Game-Based Learning, https://doi.org/10.1007/978-3-030-15569-8_14

Endeavor (shortened as *Radix* from hereon), a multiplayer online STEM game, to demonstrate ways that we can understand not only *what* students are learning through games, but also *how* they are going about learning it. The main research questions for this study are:

1. How can we use activity data metrics to characterize exploration in digital learning games?
2. What specific methods of analysis can we implement to understand students' learning processes?

In this chapter we will start with a literature review of learning games and learning analytics to provide background for this study. We will describe the Radix game, pilot implementation, and data collection. We will then present three metrics that were designed and implemented—quest progression linearity, quest event focus, and time per quest—and describe the patterns that were found by applying these metrics to Radix pilot data. Finally, we will explain potential ways these metrics could be interpreted and used in a classroom setting, and discuss the broader impact that richer data on learning pathways could have on game-based learning.

14.2 Literature Review

Learning does not only involve acquiring content knowledge but also involves the development of supporting motivation and the establishment of skills and ways of thinking. In the past, this has not been easy to realize in traditional classroom settings due to physical and time constraints. However, research over the past decade has revealed that digital educational games can support meaningful and authentic learning, in deeper ways than more conventional forms of teaching (Connolly, Boyle, Macarthur, Hainey, & Boyle, 2012; Papastergiou, 2009; Vogel et al., 2006; Wouters, Van Nimwegen, Van Oostendorp, & Van Der Spek, 2013). When they play video games, people practice a set of twenty-first century skills that can be applied to studying and working in the real world (Prensky, 2006). Gee (2003) also argues that gaming has the potential to increase the impact and effectiveness of the work of individuals by bringing about synchronized intelligence, where humans and digital tools complement each other's abilities in order to achieve new goals. In addition, the National Research Council has reported that in the field of science learning in particular, simulations and games have the potential to advance multiple learning goals including conceptual understanding, science process skills, and discourse and argumentation (Honey & Hilton, 2011). Research into learning games has also revealed some of the game elements that best enable content also inspire interest, creativity, and social interaction (Squire, 2011). These value-added features that are specifically designed to support learning have in fact been found to magnify learning (Clark, Tanner-Smith, & Killingsworth, 2016). All of this evidence explains why the Joan Ganz Cooney Center's Level Up Learning survey reported that 74%

of teachers are currently using digital games for instructional purposes with their students (Takeuchi & Vaala, 2014).

One tool that can help make sense of players' actions in open-ended digital games for learning is learning analytics (Berland, Baker, & Blikstein, 2014). Over the last decade the production of data has expanded at a stunning fast pace. In education, multiple virtual learning environments have been emerging, such as MOOC platforms, games for learning, intelligent tutoring systems, and more. To analyze all these data we need a combination of theory, design, and data mining techniques, and in order to fulfill these requisites the field of Learning Analytics (LA), an intersection between data science and learning sciences (Gašević, Kovanović, & Joksimović, 2017), has been gaining a lot of attention over the last years. The data analysis of these huge data samples has immense potential for the field. Nonetheless, LA should focus on the learning process, and therefore it also should be situated within the existing framework of educational research (Gašević, Dawson, & Siemens, 2015). In the field of education, learning analytics has explored numerous questions, such as course attrition (Kloft, Stiehler, Zheng, & Pinkwart, 2014), predicting the success of a student in a degree for admission purposes (Nghe, Janecek, & Haddawy, 2007), to predict if students are going to surpass a course or not (Calvo-Flores, Galindo, Jiménez, & Pineiro, 2006), to generate recommendations about learning resources in educational systems (Salehi & Nakhai Kamalabadi, 2013), or to predict the final grade of a student in a test (Pardos, Gowda, Ryan, & Heffernan, 2010).

Learning analytics can be especially useful in open-ended environments (like games) where the freedom of interaction is much higher (Blikstein, 2011). Therefore, the self-regulated strategies of learners and methods to measure those start to become more important in such environments (Segedy, Kinnebrew, & Biswas, 2015), and the use of optional activities in self-regulated environments has been found to have a relationship with learning outcomes (Ruiperez-Valiente et al., 2016). In the context of learning analytics applied to games, there have been new approaches in the past years, showing the challenge of developing new psychometric models for those environments (Gibson & Clarke-Midura, 2015), and one of the major goals is being able to use learning analytics for a trustworthy assessment of students' knowledge (Serrano-Laguna, Torrente, Moreno-Ger, & Fernández-Manjón, 2012). Some previous work has shared ideas similar to ours in other contexts such as to measure the focus on actions to earn badges (Ruipérez-Valiente, Muñoz-Merino, & Delgado Kloos, 2017) or the linearity of students' following the recommendations of a system in online learning (Ruipérez-Valiente, Muñoz-Merino, Leony, & Delgado Kloos, 2015).

14.3 Background

The Radix Endeavor is an inquiry-based online game for STEM learning developed at the MIT Education Arcade. It is an MMO-style game set in a virtual multiplayer world that is fairly open-ended and exploratory but that has set sequences of tasks for players to work through. The Radix world contains embedded biological and mathematical systems that involve the world's realistic but fictional flora, fauna, and civilizations. Players take on game tasks, or quests, that guide them to probe the game's systems and develop a firsthand understanding of math and biology concepts in a variety of topic areas. The game is exploratory, leaving a lot of experimenting and problem-solving up to the players. It incorporates a wide variety of content as well as STEM practices and soft skills. It is a long-form game, meant to be played over the course of a semester and revisited during each relevant curricular unit. In addition, it presents opportunities for players to collaborate both in and outside of the game, leading to a unique deep learning experience.

When players enter the game for the first time, they begin a sequence of tutorial quests designed to get players used to moving around the world, using tools, and collecting data about their environment. Upon completion of the tutorial quest line, an array of topical quest lines is unlocked, including four in biology: genetics, ecology, evolution, and human body systems; and three in math: geometry, algebra, and statistics. While the quests are sequenced within a topic area, players are free to switch between quest lines according to their interests throughout their play sessions. Each quest line may have anywhere from four to ten quests within it, and each quest is made up of multiple smaller tasks which provide some scaffolding to players. The quest content is aligned with curriculum standards, and the tasks are specific to the domain. For instance, in one of the genetics quests players must figure out how dominant and recessive traits work in order to breed non-toxic *glumbugs* for a chef to use in his cooking. In the algebra quest line, players explore a marketplace where they must barter with vendors who offer different rates of trade, using unit conversion concepts to maximize the *zorbits* they earn on the exchange (see Fig. 14.1). In order to make sense of the in-game systems and complete their tasks, players have a number of tools at their disposal. Some tools are useful across quest lines, and some are domain-specific, but all tools are accessible at all times, regardless of the quest a player is currently working on. This design means that one of the skills players are practicing is selecting the tool that will be the most helpful or efficient to solve a problem. For example, the trait examiner and trait decoder let players identify the phenotypes and genotypes of a species,

Fig. 14.1 Screen shots of tools used in Radix quests

and the breeding station lets them breed plants and animals. These are most useful in genetics challenges, whereas the data library, which lets them do simple analyses of means and distributions, can be useful in a number of math and biology quests.

The specific interactions and problems presented in the quests are unique to a topic area in order to provide an environment where players are engaging in authentic inquiry. At the same time, there are elements of quest design that are consistent across quest lines and that are important to the game's pedagogical approach. Quests are introduced in context, to present an authentic problem in the fictional world. Players know generally what they need to do, but they are not told exactly what steps to take to solve the problem or which tools to use. They need to experiment with the systems to build some content understanding, usually iterating on their strategy based on the feedback they get from the game. When they turn in a quest, or present the solution, they are asked to not only hand in a game object or artifact, but also explain their reasoning or back up their claims. For example, along with the non-toxic bugs they must also create a Punnett square that shows which parents will breed the desired offspring. There is no penalty or disincentive for submitting incorrect solutions. Rather, players get some feedback and are invited to continue experimenting or try a new approach. This type of quest design is meant to create an inquiry experience where players explore and build their own knowledge in a low-stakes environment. This provides an opportunity for players to build and demonstrate skills such as creative problem-solving, experimentation, and supporting claims with evidence. It also provides an opportunity for designers and educators to recognize those skills and assess progress in their development.

The designing of specific game elements in Radix with the goal of generating evidence of learning was one of the project's goals and research questions from the start. We aimed to create a digital environment for inquiry learning that could provide feedback to both players and teachers about how players are approaching problems, using tools, and building conceptual knowledge in math and biology. These are skills that are difficult to measure with traditional tests, and we wanted to research how well a digital game could collect telemetry data for an embedded assessment approach. For example, quest tasks were designed to provide opportunities for players to build and demonstrate their inquiry skills. Game data was collected for actions relevant to quests and exploration, and that data was interpreted to provide teachers with feedback on what their students were struggling with. The feedback mechanisms were only built out for an initial level of two quest lines. In this study, we apply learning analytics techniques to form the basis of the next level of measurements that could be conducted around Radix gameplay to provide insights for how players approached the quests rather than simply describing what they were able to achieve or perform in the game.

14.4 Method

14.4.1 Pilot Description and Context

Radix launched as a free tool available across the US and internationally in late January 2014 and has been played in all 50 states and at least seven different countries. The dataset used in this study was collected during the pilot period which ran through August 2015. While the game was designed with high school math and biology teachers in mind, Radix has been used by upper elementary, middle, and high school teachers as well as by a few instructors at community colleges and universities. Outside of the formal school environment, the game has also been picked up by various after school groups, enrichment programs, and the homeschool community who are using it with a wide variety of ages. During the pilot period, informal marketing and outreach was done to recruit teachers to participate in the pilot at various levels. This included reaching out to local and national teacher networks to publicize the game, as well as a number of press articles and blog posts showcasing the project and its opportunities for participation. Teachers created accounts for their students to play, but players who heard about the game via other channels were also able to create player accounts not associated with a school or teacher. Participating teachers were provided with some professional development opportunities and implementation resources, but they were encouraged to tailor their implementations and use the game as they saw fit in their classroom. Most of them had their students play relevant quest lines at the time they were covering a given topic area in their class. Outside of school players naturally played as much or little as they chose to, working through quest lines according to their interests.

14.4.2 Data Collection

We used the data from the pilot study that we described in previous Sect. 14.4.1. The design of Radix emphasized a rich data infrastructure that could allow researchers to perform detailed analytics of students' interactions. Radix has a relational database with more than 20 tables that collects most of the interactions of students with the game, such as player metadata, tool usage, quest related events, or social interactions. As part of this study, we develop an algorithmic machinery that processes such data to create interpretable information such as the metrics that we present.

The data set includes over 14,000 Radix accounts; however, some of these accounts were not activated or barely interacted with the game. We therefore included only those accounts that were active within the game for at least 1 h. With this filter, the number decreases to 5493 accounts with 5532 virtual characters. From the total, 4841 (87.5%) of the characters were student accounts created as part of the pilot studies in schools, and 691 (12.5%) of the characters were created by other online users. Some of these characters used Radix for over 22,000 h, generated more than one million events, completed more than 68,000 quests, and sent more than 60,000 social chat messages.

14.4.3 Data Analysis and Metrics

This study focuses on the processes students use to solve quests within the game and then ultimately connects these metrics with an in-game learning outcome such as the percentage of correct responses. We have defined three brand-new metrics based on process mining techniques to investigate how students are interacting with a learning environment that is very open-ended and presents numerous possibilities and choices within the learning process of each student. The three metrics that we define are as follows:

- *Quest progression linearity*: This metric takes advantage of the multiple quest chains available in Radix as described in Sect. 14.3. Since students are free to jump from one quest chain to another, we investigated this issue by computing a percentage of quest chain changes by each student when they are still able to progress further within the current quest chain. For instance, if a student finishes quest GN1.1, from the Genetics topic which is part of the GN1 quest chain, and then completes ST1.1 which is part of Statistics ST1 quest chain, that would count as a quest chain change. However, if the student finishes GN1.8, which is the last quest of GN1 quest chain, and then completes ST.1.1, that would not count as a quest chain change since GN1.8 was the last quest of that quest chain and the student is forced to switch to a new one. Then, we computed a percentage as follows:

$$100 * (\text{number of quest chain changes}) / (\text{number of quests completed})$$

- *Quest action focus*: Each of the quests of Radix is designed to be solved using experimental approaches by using specific tools to answer questions. Often, students will need to explore a bit before they are able to understand the requirements of the quest, what tools they need to use, and how. In this metric, we explore the percentage of events that each student completes before solving a quest, which of these are strictly related to quest events, and which of them were not explicitly necessary to solve the quest (such as other tool events or social actions). Then, we computed a quest action focus percentage as follows:

$$100 * (\text{action events related to quest}) / (\text{total events before quest})$$

- *Time per quest and average time difference per quest*: Since the path to solve each quest is not obvious once they receive the task, it might need exploration, experimentation, and extra time depending on the strategy and knowledge of each student. Additionally, each quest might have a potentially different difficulty or length. Therefore, exploring the time required to solve each quest provides the potential to understand students' process and game dynamics. We computed the quest completion time between the acceptance of the quest and the quest being completed, omitting any times when the user was not interacting with the game. Since each student might resolve different quests, and each quest might potentially need different efforts, computing an average time per quest for each student would be biased by the quests that they completed. Therefore, to generate an informative per student metric, we computed the average time per quest and then used the time spent by student in that quest to calculate the difference and compute an average time difference per quest for each student. This way, we can finally present an average number per student that indicates how many minutes faster or slower they are solving quests compared to the rest of students:

$$\sum_{i=1}^{Q} t_{j,i} - a_i$$

where Q is the number of quests in Radix, $t_{j,i}$ would be the time t to complete quest i by player j, and a_i would be the average time for all players to complete quest i.

Since the Radix world is so open, we acknowledge that we cannot be completely accurate about measuring if student actions are devoted to finishing one quest or not; therefore some of these metrics represent an approximation of the ground truth. We explore these metrics at a student level, but also as global dynamics defining the Radix ecosystem, which can be useful for game design and understanding complex behaviors in open-ended game environments.

14.5 Results

The three first subsections of the results describe the global dynamics of each metric while the fourth subsection connects together the three metrics at a student level with joint visualizations and correlations.

14.5.1 Quest Progression Linearity

To illustrate the dynamics of the quest system in Radix, we created a global graph of the most typical quest pathways followed by students, using Gephi (Fig. 14.2). The graph was constructed by creating edges between quest completions. For example, if 10,000 students completed quest TUT1.1 and then went on to complete TUT1.2, that value would represent the weight between nodes TUT.1.1 and TUT.1.2. Then, we used the thickness of the edge (line) to encode this weight, showing the

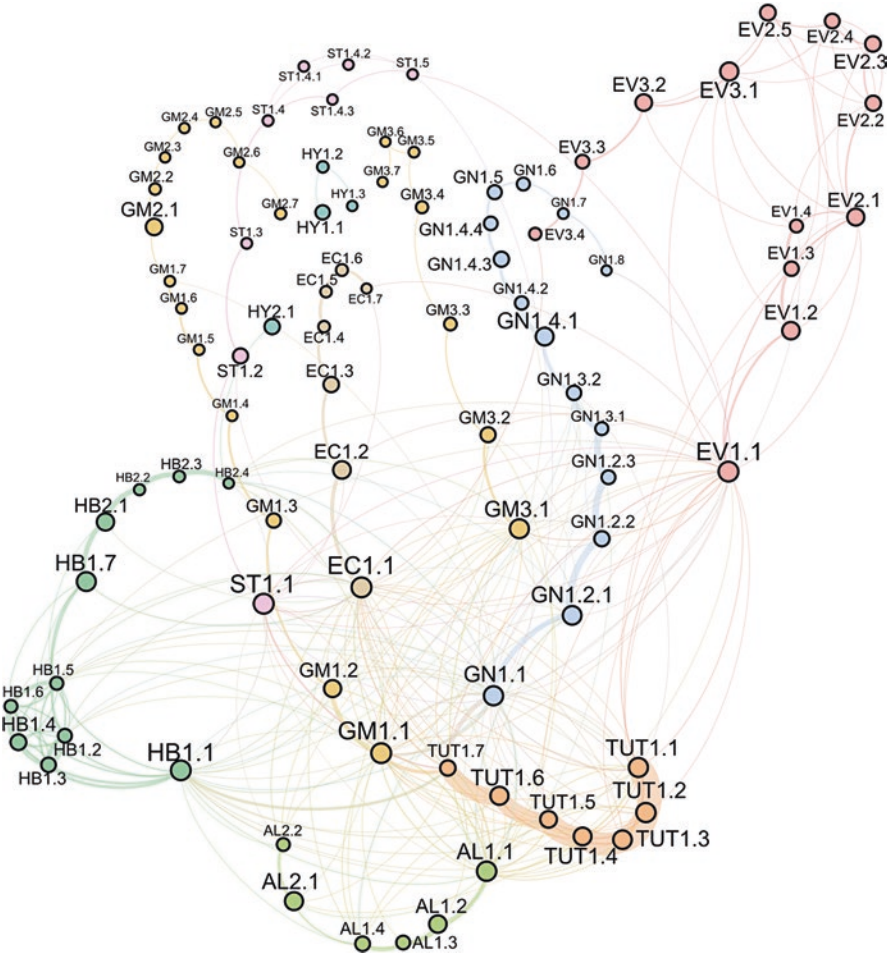


Fig. 14.2 Graph network represent the global dynamics of the quest system

frequency at which students followed this path. Additionally, we used the size of the node/label to codify the centrality of the node within the network. Finally, the label represents the quest ID, and the color represents the topic of the quest.

The global dynamics of the quest ecosystem are very clear in Fig. 14.2. For example, we can see a high centrality for TUT1.1 quest, since is the first quest available in Radix, and then for EV1.1, GN1.1, EC.1.1, GM1.1, ST1.1, AL1.1 and HB1.1, as they are the first quest in each quest chain (or topic) and unlock after players solve the first tutorial quests. Moreover, we can see how the layout algorithm has grouped quests from the same topic close together based on the weight influence, which denotes that students usually solve quests from the same topic without jumping around. Additionally, we can see thicker edges between consecutive tasks of a quest chain, for example, TUT1.1, TUT1.2 until TUT1.7 which indicates that students generally solve consecutive quests from the same quest chain. These results are tightly coupled with the design and implementation of the game. The way the quests are presented in the game leads players sequentially through a quest line, although it doesn't force them to complete tasks in the given order. In addition, many teachers who used Radix in class specified a particular quest line, encouraging students to focus on that topic area.

While Fig. 14.2 explores the global dynamics, the individual learning path of each student can be completely different. Therefore, to illustrate this idea we present in Fig. 14.3 two student examples, one that follows a highly linear quest progression and a second one that has performed frequent quest chain jumps during his/her learning process. Student A represents a very linear quest progression: the student completes consecutive quests from each quest chain and only changes to a new quest chain after finishing the current one, changing quest chains only 4% of time upon completing an individual quest. On the opposite end of the spectrum, Student B advances by frequently jumping between quest chains; more exactly, upon completing a quest, 60% of time they changed quest chains. We will delve into the significance of how these different behaviors and strategies might influence learning outcomes and other metrics later in this paper.

14.5.2 Quest Action Focus

Each quest of Radix is designed to require some degree of inquiry and exploration with the environment to understand the task and complete the requirements of the quest. Each quest chain focuses on specific STEM content that students will learn through experimentation with the environment by using specific scientific tools. Therefore, each quest chain is associated with a set of tools that allow students to

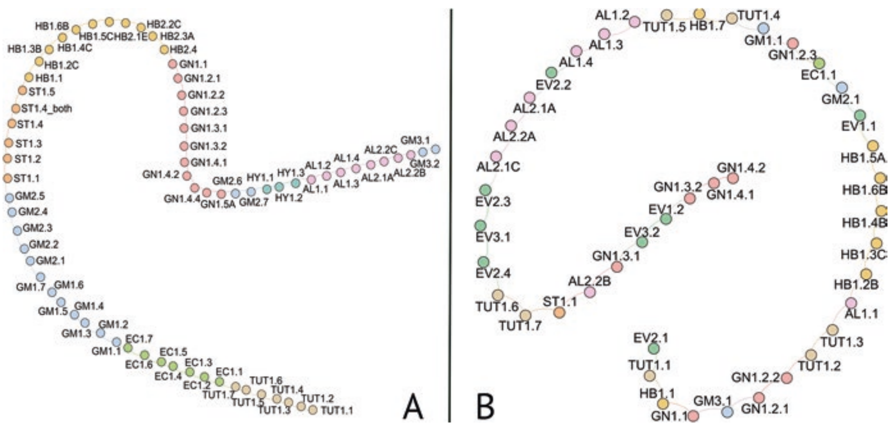


Fig. 14.3 Graphs representing one student with a high level of progression linearity (**a**) and one with a low (**b**)

gather the evidence to complete the task. However, students might need to experiment with the different tools before they have a clear sense of which tools are appropriate to the task and how to solve it. For example, in the GN1 (Genetics) quest chain students interact with the flora and fauna of Radix to learn how to breed for certain phenotypes, and then apply those skills to help villagers to cure diseases and more. Students have to identify the relevant genotypes and use Punnett squares to teach villages how to breed the traits of the flora and fauna themselves. To accomplish these tasks, they need to use four tools: the trait decoder and examiner, the Punnett square, and the breeding station. The rest of the available tools or actions are not necessary for the GN1 quests, but students might experiment with them while they work to solve the quest.

That is why, as we can imagine, the dynamics in terms of events and actions of each quest should be different, and we use that fact to compute this quest action focus measure for five of the main quest chains AL1, GN1, GM2, EV1, and ST1, and Fig. 14.4 shows a graph with the dynamics of four of them. Analogously to the previous section, we generated graphs to explore the dynamics of each quest chain separately, by creating an edge between each event generated by students before completing any of the quests from its chain. Again, the thickness of the edge encodes its weight, and the size of the node/label the centrality of the node within the network. The color encodes the type of node, and we use green for the use of tools related to quest requirements, orange for other tools, yellow for quest events, blue for party events, and white for chatting events. Note that which tools are coded as a

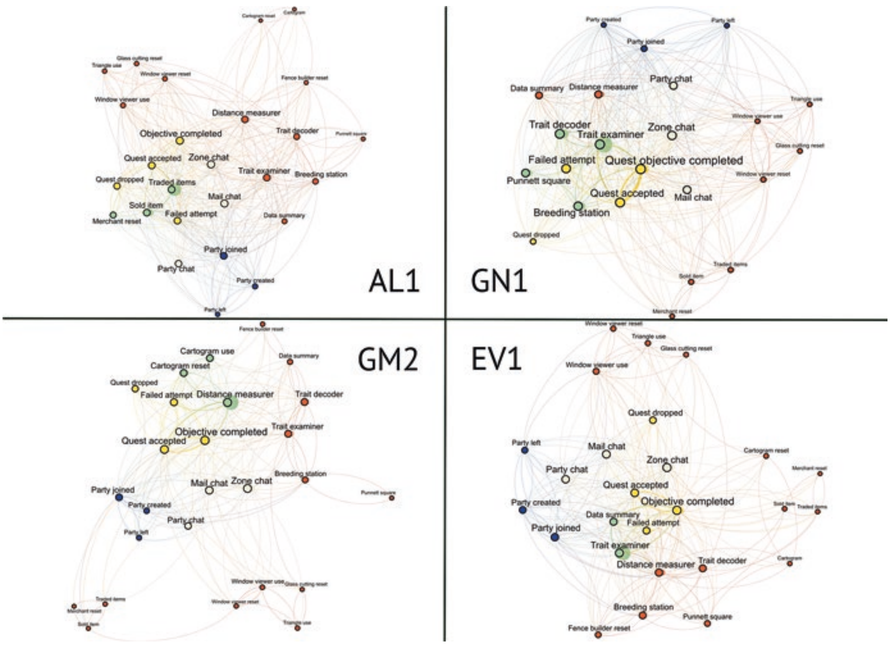


Fig. 14.4 Graphs of action dynamics per quest chain

green or orange node should change from quest chain to quest chain according to the requirements of each quest chain.

There are a number of interesting things to note from the dynamics of each quest chain in Fig. 14.4. First, we can see how the most central part of the network are always those actions related to the quest (green) and quest events (yellow), whereas other action events not related to the quest (orange) still show in the network but form subgroups as part of periphery. Interestingly, the periphery groups are similar in each graph; for example, we see the subgroup formed by the events "triangle use," "glass cutting reset," "window viewer use," and "window viewer reset," which represents the set of actions required to complete GM1. This might represent the behavior of jumping from one quest chain to another, and that is why the global dynamics of each quest chain capture these subgroups as well. The outer glow that some of the nodes have, for example traded items in AL1, represents the self-loop degree of an event, hence the thick self-loop of traded items in AL1 would mean that trading items consecutively was a very common two-gram sequence. This allows us to identify the main tools of each one of the quest chains. Finally, the high degree of centrality of chat events (white) that show that the social component is highly interspersed between quest actions.

14.5.3 Time per Quest

The last metric that we have proposed targets the time to resolve each quest of Radix. The algorithm gathered all events with timestamps that before the quest completion event, back to the last quest completion event and computes the effective active time to complete each quest. We can analyze this metric at both the quest and student level. At the student level, we can see a student's "efficiency" in quest completion by comparing the amount of time or events needed to that needed by other students. At the quest level, we can see approximate the amount of effort required to solve a quest by looking at whether it requires on average more or less time or events than other quests. Additionally, the percentage of correct solutions suggests the degree of difficulty of a quest.

In Fig. 14.5, a boxplot visualization with the distribution of the time and percentage of correct responses per quest appears on the left. As a summary, we can see that the median of percentage correct is around 62% with a high variance, which denotes that there are some quests with very low correct ratios. The median time required per quest is 8.8 min, so generally quests do not take much time to complete, but we can see numerous outliers in the upper (more time) part of the distribution. For example, the quest EV1.1. (Evolution)—in which students need to explore to find the typical characteristics of *menjis* (a type of animal in the world of Radix)—required, on average, 110 min, but had a high correct response rate of 80%. Another example of the same quest chain would be EV1.4, where students have to make sense of *menjis* characteristics by responding to some questions, took on average 4.8 min but had a correct response rate of only 25%.

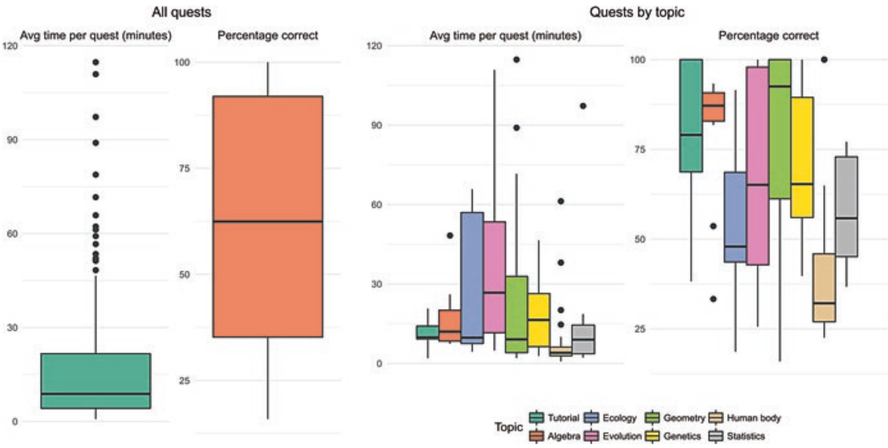


Fig. 14.5 Boxplot distribution of the metrics by quest and by topic

These dynamics can change a lot from one quest topic to another, so in Fig. 14.5, each topic is broken out in an analogous visualization on the right. Geometry and Algebra have the highest, and Human Body has the lowest correct percentage ratio and lowest time per quest, which is likely due to the fact that the response method of this quest chain consisted of multiple choice questions, so students might have been using trial-and-error to guess the correct response. The quest topic that required the highest average time is Evolution, which is likely due to that domain requiring travel to different zones and data collection from a number of animals.

14.5.4 Distribution of Metrics by Student and Correlations

The previous subsection presented some global dynamics of the metrics that we have explored, and this subsection reports the distribution of the individual metrics at the student level with correlations. Figure 14.6 shows a boxplot with the distribution of each one of the metrics per student. The first item shows the quest progression linearity with a median of 13.4% and mean of 16% of quest chain changes. Therefore, we see that generally speaking, students carry out a very linear learning path, following quest chains and infrequently switching between quest chains. We do see some outliers, with more than a 50% change of quest chains as discussed in Sect. 14.5.1, but this behavior does not represent the norm. The quest action focus has a median of 80% and mean of 65%, which represents that when working on a quest, the average student shows a high focus on quest-related events. Surprisingly, for the events that do not belong to quest actions, an average of 14% of events were “other action events” (i.e. tools not related to the quest), and 20% were social events (e.g. sending any kind of chat messages). However, as we saw in the global dynamics, many of the students are socializing within the game while resolving the quests.

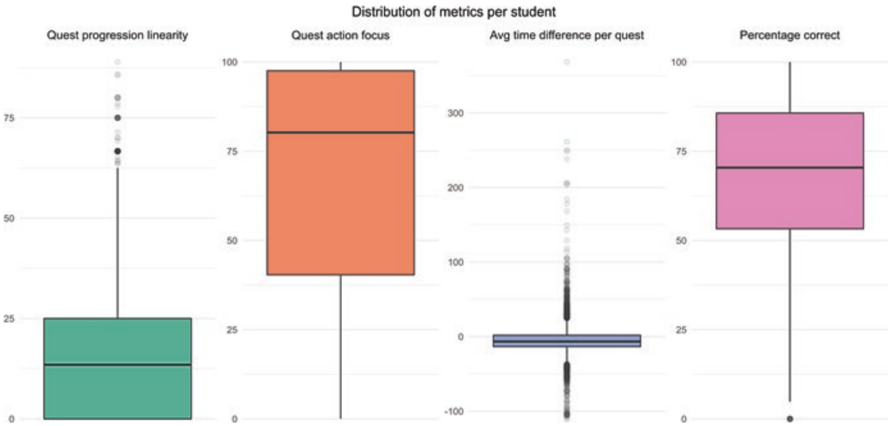


Fig. 14.6 Boxplot distribution of the metrics by student

Table 14.1 Average value of metrics split in two cohorts by type of account (all *t*-tests are significant with *p*-value below 0.01)

School student	Quest progression linearity (%)	Quest action focus (%)	Avg time difference per quest (min)	Percentage correct (%)
No	5.98	73.99	−1.77	80.95
Yes	17.47	64.23	−5.08	67.52

The third metric represents the time difference per quest and is thus a measure of how fast students solve quests in comparison to their peers. The median is −6.5 and mean −4.6 min, but more importantly, it shows numerous outliers that indicate some students solving quests much faster or slower than the average. Finally, the percentage of correct responses has a mean value of 70%, and again we see a moderate variance that represents students with a higher or lower percentage of correct responses.

Comparing differences in metrics related to students' learning paths between cohorts yields interesting results. In this case, we wanted to compare two cohorts, based on different types of accounts. First, those accounts that were part of the pilot study and were created by students in school, and second, those accounts that were independently created online by any interested individual. Table 14.1 shows the average for each metric and cohort, where all of the *t*-tests show a statistically significant difference between the means of the two cohorts. The quest progression linearity varies significantly, with school students following a less linear path (17% quest chain changes) than the online learners (5.98%). On the other hand, online learners engaged in fewer chatting events than school students, resulting in a higher quest focus for online learners; 64.2% of school students' events were quest focused versus 75% for online learners. We do not have a clear hypothesis that can explain these differences, but one possibility is that school students explored more in the game (hence with more quest chain jumps and less focus) and that online learners were more serious about advancing with the game. The difference in chatting events may have been facilitated by classroom use, with students playing Radix together in class. Finally, we can also see that school students were a few minutes slower in finishing quests and had a much lower percentage of correct responses.

Table 14.2 shows the correlation between the metrics. To remove the possibility of spurious correlations or diminished effects, we computed the set of correlations only for those students that interacted with Radix for at least 5 h ($N = 1397$). We find a few interesting insights. First, if we look at the correlations of percentage correct to quest progression linearity, we find a low-moderate negative correlation of −0.26, which might indicate that students who are jumping between quest chains will have more failed quest attempts. Second, we also find a low positive correlation of 0.2 when comparing percentage correct with the quest action focus, indicating that students who create a higher proportion of events related to the quest tools are more likely to correctly solve the quest. While common-sense might suggest these results, more work is needed to understand the influence on learning. Finally, we find a low-moderate correlation of −0.32 between quest progression linearity and quest action focus, which indicates that if students are switching quest chains frequently, they are likely to have a lower quest action focus. This may be because they are jumping from one set of quest tools to another due to the frequent changes in quest lines.

14.6 Discussion

The metrics presented here, describing quest progression linearity, quest action focus, and time per quest, help tell the story of how students are learning and exploring in a game like Radix. Specific learning outcomes are only one aspect of what a student gets out of playing an inquiry-based game, whereas the experience of exploration and discovery is an important part of a student's learning experience in the game. This pathway may vary from student to student depending on their personality, interests, and ways of thinking.

When we looked at quest progression linearity, we see that there is quite a bit of variation in the sequences in which students completed quests. Some students were more focused on one quest line at a time, while others jumped around, completing quests in different topic areas. Radix was designed for students to have this choice

Table 14.2 Correlations between the metrics (∗indicates a *p*-value below 0.01)

	Quest progression linearity	Quest action focus	Avg time per quest	Percentage correct
Quest progression linearity	1	−0.32∗	0.09	−0.26∗
Quest action focus	−0.32∗	1	−0.14	0.2∗
Avg time per quest	0.09∗	−0.14∗	1	−0.07
Percentage correct	−0.26∗	0.2∗	−0.07	1

to allow students to follow their own interests, an important aspect of inquiry learning. It is, however, important to note that neither type of behavior is inherently better than any other, though these differences provide a source of rich information for teachers to understand what their students are doing in the game, what content they have explored, and what they are interested in. Once teachers understand this metric, they can use the information in their classroom context. For example, a teacher who has asked students to explore the world in an open-ended game-based learning lesson might be very interested to know which students dug deeply into a specific topic area and developed a deep interest in that domain, so that she could support their learning and offer resources for continued study. In addition, this teacher might like to know which students jumped around the most, because it might provide useful evidence of either a lack of focus or an independent motivation, depending on the student. While the quest progression linearity metric alone doesn't tell us what students are learning, it helps describe students' patterns of interaction and allows some inferences into their interest level, which can inform how a teacher guides their learning.

The quest action focus metric provides another type of insight into how players explore the game world. In Radix, while a player solves quests they can also engage in more interest-based activities. Providing this information to a teacher could be helpful in order to understand how players are engaging with the tools of the game. If a number of students are using the trait examiner tool during a geometry quest, for example, teachers may want to bring that up in a class discussion, finding out what players were trying to do or what they were interested in, in order to support student interests and tie those into the curriculum more tightly. Perhaps those students were sitting near each other, noticed a rare species that happened to live in the area they were walking through, and all decided to find out more about its traits. The teacher might decide to have some discussion about how that experience connected to their science class, or pose the question of why that rare species lives in a particular biome. Similarly, if students seem to be using the chat feature more than usual during a particular quest line, teachers may recognize that there is something of note there—whether it be a challenge that students need to work through together, something they are excited about, or an indication that students were off-task (itself something worth probing further). While the metrics themselves don't tell teachers what exactly is going on, they give a sign that there is something going on that may be worth discussing, thereby tightening up the feedback loop between lesson plans, student experiences, and teacher feedback.

Lastly, being able to compare time spent on quests across quest lines and across students in relation to either a class, school, or larger community of players can also give teachers a richer picture of how students are spending time in the game. If many students spent more time on evolution than on ecosystems for example, and had more failed attempts or demonstrated misconceptions in other science lab activities, a teacher may realize that students need more review on evolution concepts. As another example, if a student spends less time completing the algebra quest line than they did to complete any of the geometry quests (in relation to class averages), but keeps coming back to algebra tools during other quest lines, a teacher might decide to give that student deeper challenges in the area of algebra, where they have an interest and ability.

In addition to providing information just for teachers, if built into the game, all of these metrics could facilitate valuable self-reflection for students not only on *what* they learned, but *how* they learned. Students can explore questions of how efficient they are in their learning, whether they are spending time on the areas they are most interested in, and where they might need to focus their efforts in and outside of the game. These tools can deepen learners' engagement in self-assessment processes, supporting independent learning habits. For all stakeholders, understanding an array of factors about how students are learning and the variety of learning pathways present within one activity can emphasize the richness of the learning experience and the importance of assessing and characterizing more than content knowledge and skills. Providing richer metrics that help us see differences in learning pathways and open reflective conversations can encourage the community to value constructs that are traditionally harder to measure. This approach fits well with formative assessment practices, in that it paints a richer picture that informs what activities to do more or less of, new strategies to try, and connections that can be made between concepts. At a higher level, these three metrics represent a new way to look at measurement and feedback in open-ended digital learning environments. These metrics and others using a similar approach could be applied to different learning games and digital interactives to expand our understanding of learning experiences and refine the way digital learning is embedded into teaching practices.

Metrics and analytics like the ones presented here enable us to measure patterns of exploration. This type of measurement is not unrelated to educational assessment, but it's important to note their distinct goals and uses. The aim of educational assessment is primarily measuring learning outcomes and collecting empirical evidence of learning gains. Our work on measurement of patterns of exploration does not, however, let us make specific claims about student learning. Rather, it uses clickstream data to discover analytics that can tell us how people are engaging with a game or digital learning experience. We can identify and categorize patterns of interaction and engagement that may vary across individual learners and their learning contexts. These two approaches—educational assessment and analytics of engagement—are complementary because together they provide a complete picture of both what students are learning as well as how they are learning it. In order to guide students along a learning pathway that is productive, educators and learning designers need to know what knowledge and skills their students are building, and also have some understanding of the mechanisms being used to get those results. Either one without the other does not fully explain a student's learning. By combining learning analytics and educational assessment together in games like Radix, games can provide more robust and actionable interpretations of how students are learning from playing games.

14.7 Broader Impact and Future Work

The way we have defined patterns of exploration and begun to measure them, as described in this chapter, can be applied to other learning games as well. We have presented some examples of how these metrics could be used by teachers to inform instruction, and by students to enable self-reflection in the context of Radix. Building these approaches into an educational game can expand the game itself into a more complete game-based learning system. Continually updated analytics mean that patterns of exploration identified for a given student or class can feed back into the game in the form of adaptive leveling or customized scaffolding. In addition, these informative analytics can be communicated to teachers who can make meaning out of it to provide personalized feedback and support on an ongoing basis. By recognizing these patterns in this way, either the teacher or the game itself can provide valuable data-driven scaffolding and feedback, thereby making the learning experience more relevant to students. All users and stakeholders can more easily recognize the multiple pathways that lead to learning, and celebrate the variety of ways students choose to explore concepts. Our long-term goals in game design and learning analytics are to inspire and assist other designers to incorporate measurement of patterns of exploration into their games, simulations, and digital interactives for learning. We believe that measuring these patterns will not only provide informative data to teachers and students, but that it can reveal the types of exploration actually happening in learning games. Moreover, making the patterns visible can push designers to shape their game environments to be more inquiry-based, student-centered, and constructivist, incorporating more progressive pedagogies that support deep learning and the building of future-ready skills.

The possibilities for future work are far-reaching. We have seen these three metrics exert a small influence on the percentage of correct solutions of a student, which can be seen as a learning outcome within the game environment. However, there is low fidelity of implementation within this study, since students who are working together might be sharing solutions, and because teachers varied widely in how they asked students to use the game. In the future, to more closely connect patterns of exploration with learning outcomes, we might use the pre- and post-tests performed as part of the pilot study in Radix to compute learning gains and find potential relationships with the process metrics we have described. One of the problematic areas of learning analytics in general, and in a game-based learning system specifically, is that the models and metrics are hardly generalizable due to important differences in context. Therefore, we would like to work on developing more general process mining methods that can be applied in multiple game-based learning systems with minimal adaptation. We would like to apply those methods to a variety of learning games designed in our research group and from other organizations. This would enable us to research how patterns of exploration interact with the learning of both content and skills. Through working with practitioners, these learning analytics methods would also provide a rich environment to understand how teachers and students can use the measurement data to become aware of the exploration they are engaging in and more effectively guide their learning journeys. We believe this approach and the ability to better use patterns of exploration represent one of the cornerstones in open-ended and complex environments for learning.

References

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. https://doi.org/10.1007/s10758-014-9223-7

Blikstein, P. (2011). Using learning analytics to assess students ' behavior in open-ended programming tasks. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (pp. 110–116). https://doi.org/10.1145/2090116.2090132

Calvo-Flores, M. D., Galindo, E. G., Jiménez, M. C. P., & Pineiro, O. P. (2006). Predicting students' marks from Moodle logs using neural network models. *Current Developments in Technology-Assisted Education, 1*(2), 586–590.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research, 86*(1), 79–122.

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. https://doi.org/10.1016/j.compedu.2012.03.004

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends, 59*(1), 64–71.

Gašević, D., Kovanović, V., & Joksimović, S. (2017). Piecing the learning analytics puzzle: A consolidated model of a field of research and practice. *Learning: Research and Practice, 3*(1), 63–78.

Gee, J. P. (2003). What video games have to teach us about learning and literacy. *Computers in Entertainment (CIE), 1*(1), 20.

Gibson, D., & Clarke-Midura, J. (2015). Some psychometric and design implications of game-based learning analytics. In *E-learning systems, environments and approaches* (pp. 247–261). Cham, Switzerland: Springer.

Honey, M. A., & Hilton, M. (2011). *Learning science through computer games and simulations. Committee on science learning: Computer games, simulations, and education*. Washington, DC: Board on Science Education, National Research Council, The National Academies Press.

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014). Predicting MOOC dropout over weeks using machine learning methods. In *Workshop on Analysis of Large Scale Social Interaction in MOOCs* (pp. 60–65). Retrieved from http://www.aclweb.org/anthology/W/W14/W14-41.pdf#page=67

Nghe, N. T., Janecek, P., & Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. In *2007 37th Annual Frontiers in Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports. FIE'07*. (p. T2G-7).

Papastergiou, M. (2009). Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation. *Computers & Education, 52*(1), 1–12. https://doi.org/10.1016/j.compedu.2008.06.004

Pardos, Z. A., Gowda, S. M., Ryan, S. J. D., & Heffernan, N. T. (2010). Ensembling predictions of student post-test scores for an intelligent tutoring system. In *Proceedings of the 4th International Conference on Educational Data Mining* (pp. 189–198).

Prensky, M. (2006). *Don't bother me, Mom, I'm learning! How computer and video games are preparing your kids for 21st century success and how you can help!* St. Paul, MN: Paragon House.

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., & Delgado Kloos, C. (2017). Detecting and clustering students by their gamification behavior with badges: A case study in engineering education. *International Journal of Engineering Education, 33*(2–B), 816–830.

Ruiperez-Valiente, J. A., Munoz-Merino, P. J., Delgado Kloos, C., Niemann, K., Scheffel, M., & Wolpers, M. (2016). Analyzing the impact of using optional activities in self-regulated learning. *IEEE Transactions on Learning Technologies, 9*(3), 231–243. https://doi.org/10.1109/TLT.2016.2518172

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., Leony, D., & Delgado Kloos, C. (2015). ALAS-KA: A learning analytics extension for better understanding the learning process in the Khan Academy platform. *Computers in Human Behavior, 47*, 139–148. https://doi.org/10.1016/j.chb.2014.07.002

Salehi, M., & Nakhai Kamalabadi, I. (2013). Hybrid recommendation approach for learning material based on sequential pattern of the accessed material and the learner's preference tree. *Knowledge-Based Systems, 48*, 57–69. https://doi.org/10.1016/j.knosys.2013.04.012

Segedy, J. R., Kinnebrew, J. S., & Biswas, G. (2015). Using coherence analysis to characterize self-regulated learning behaviours in open-ended learning environments. *Journal of Learning Analytics, 2*(1), 13–48.

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. https://doi.org/10.1016/j.procs.2012.10.072

Squire, K. (2011). *Video games and learning: Teaching and participatory culture in the digital age. Technology, education—Connections (the TEC series)*. New York, NY: Teachers College Press.

Takeuchi, L. M., & Vaala, S. (2014). *Level up learning: A national survey on teaching with digital games*. New York, NY: Joan Ganz Cooney Center.

Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research, 34*(3), 229–243.

Wouters, P., Van Nimwegen, C., Van Oostendorp, H., & Van Der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology, 105*(2), 249.

Chapter 15
Making a Game of Troublesome Threshold Concepts

Kayleen Wood

15.1 Introduction

The body of existing research on gamification in educational settings compares traditional pedagogy to gamification pedagogy. There is general consensus that it has a positive effect on learning outcomes and engagement (Kapp, 2016). Technology enabled curriculum affords the inclusion of a gamified plot driven narrative—storytelling, with the learner in a role-playing, first person perspective (de Villiers & Hess, 2019)—and provides adaptive, self-paced, and self-directed learning. Combined with learner scaffolds (the feedback and support mechanisms) gamification pedagogy holistically addresses cognitive load theory. However, simply transferring the conventional lesson plans used in the classroom to the online learning environment is not the answer. But what gamification features lead to learning and what conditions support gamification for learning (Kapp, 2016)? Further, how can game-based assessment (GBA) be integrated into the learning experience to ensure learning outcomes are met?

We commonly see GBA as part of a capstone unit where learners are required to bring all their accumulated learning into a virtual simulation. However, with this type of critical thinking identified as a desirable attribute for graduates, accounting educators should be embedding opportunities to develop these skills at the introductory unit level. To address this, an active gamified learning pedagogy for the time value of money concept has the capacity to insert the passive theory and concept textbook-based learning into a plot-driven narrative "promot[ing] commercial realism and connect[ing] discipline-based knowledge with practical situations" (Siriwardane, 2014, p. 97).

The learning game in total is the GBA, and therefore the whole design model and mapping of the pedagogy to the game is key to success. This book chapter presents

K. Wood (✉)
Southern Cross University, East Lismore, NSW, Australia
e-mail: kayleen.wood@scu.edu.au

D. Ifenthaler, Y. J. Kim (eds.), *Game-Based Assessment Revisited*, Advances in Game-Based Learning, https://doi.org/10.1007/978-3-030-15569-8_15

the Gamification Alignment Table and the Gamification Alignment Model for GBA, through the constructive alignment of a gamification pedagogy for threshold concepts. This is done using the example of the time value of money gamified learning experience. The theory, the process, the design, and the evaluation from this research is reported.

15.2 Learning and Curriculum Theories

Research on game-based learning and curriculum has explored learner engagement and curriculum design, to discover if learning can happen in games. Thomas and Seely Brown (2011) recognised that games have uncertainty built into them via the choices and decisions available, and that there is not always one correct answer. Instead the curriculum, or game world, is an arena where the content is bounded and a synergy of learning takes place as *collective indwelling* in a fluid way: "the feeling and belief that group members share a tacit understanding of one another, their environment, and the practices necessary to complete their task" (Thomas & Seely Brown, 2011, Loc 1621 of 2399). The educational designer sets the boundaries or parameters of the learning space, and exit points are multiple expressions of learning as assessments. In fact, there may be no finite destination answer, just a progression of increasingly complicated questions or tasks. This structure parallels the rubric used in assessment grading where levels of achievement or performance are situated in a matrix. The output of one learning game play activity, becoming input for the next higher level, as the learner reflects, self-assesses, and receives peer assessment from other members of the gamified learning experience group. All these formative assessment instances combine for a total *assessment as learning* experience evaluated by the teacher.

15.2.1 Pedagogically Sound

The challenge for GBA to be pedagogically sound (Arnab et al., 2013) requires balancing curriculum and game development, to ensure the GBA is not technology as an add on, but a resource to enhance the learning. Duplicating conventional content and assessment used in the classroom to the online learning environment is not the answer. Learners interact with material differently when online and face to face. Activities that work in the classroom typically cannot directly transfer to online, nor do online activities conveniently adapt to conventional classroom delivery (Tennant, McMullen, & Kaczynski, 2010). Constructively aligned curriculum with learning outcomes matched to assessment items through learning resources is still central.

15.2.1.1 Constructivism

Using Crotty's (1998) perspective of theory and practice, constructionism (noun) is the epistemology, and constructivism (verb) is the theoretical perspective that describes the actions the participants take. Extending this to social constructivism, as an approach to teaching and learning, we have the pedagogical practice of using culture, language, and context to create meaning.

> Constructivists emphasise the instrumental and practical function of the theory [of] constructionism and knowing. This constructivism is primarily an individualistic understanding of the constructionist position. Social constructivism [as a subgroup emphasises] not focusing on the individual mind but outward to the social constructions of meaning and knowledge, as a more adequate description of knowledge created in the process of social exchange. (Crotty, 1998, pp. 57–58)

The constructivist developmental processes in a gamified learning experience are the feedback mechanisms. When feedback is intrinsic to the active learning, that is, it is the consequence of a choice relative to the intended goal, the learner can resolve and construct their own learning without extrinsic teacher intervention (Laurillard, 2002, 2016).

15.2.1.2 Constructive Alignment

Is it possible to construct a gamified learning experience which incorporates GBA, that is going to meet everyone's needs? Biggs' (2003) principle of constructive alignment as a system of interrelated, sequential items, gives us a robust repeatable model for outcomes-based curriculum design.

The constructive alignment model of curriculum is derived from both the teaching and learning perspectives, matching learning objectives to assessment tasks. Thus, good curriculum design is a transitional sequence of activities supported by linked items of assessment (Nulty, personal communication, 3 October 2011). Constructive alignment (Biggs, 2003) allows learners to construct meaning via accumulation of knowledge, and teachers deliberately align learning outcomes with learning activities measured by appropriate assessment and feedback. This targeted approach fits easily into the gamified learning experience, where learners explore and engage with various content activities and assessment as part of the narrative thread of the game.

15.2.1.3 Social Constructivism

Social constructivism as an approach to teaching and learning is a pedagogical practice using culture, language, and meaning to give context. "The social constructionist perspective opens up the possibility to look at the interaction between the individual construction and the culture within which it exists" (Crotty, 1998, p. 63).

In this way it "offer[s] rich alternatives for understanding the processes of learning and education, knowledge and truth, and experience and culture … [presenting] a perspective on the world of action and interaction" (Hickman, Neubert, & Reich, 2009, p. vii). From an education design perspective, Millwood (2014) takes an expressive constructionist stance which is compatible with GBA to:

> [S]upport creative decision making in the design of learner-centred, technology-enhanced education, as a design practitioner in technology-enhanced learning ... [seeking] analytical and descriptive means to improve designs through effective design and development processes. (p. 3)

15.2.1.4 Cognitive Load Theory

Cognitive load theory holds that working memory has finite cognitive capacity. Applied to an education setting, if a learning task requires too much capacity, learning and related assessment outcomes will be hampered (de Jong, 2010). "Gamification is an approach that can make immersion easier, lessening the cognitive load of the students in [e-learning] environments and aiming for an enjoyable experience" (Simon, 2016, p. 208).

Pertaining to learning and assessment, cognitive load theory identifies how learning occurs for the specific functional elements of data processing, and the types and limits of memory used. Optimising total cognitive load results in improved learning efficiency with less stress. Using cognitive load theory, Mostyn (2012) showed that procedural efficiency, not motivational methods, was more efficient for novice learners in his accounting case study. He proposed the application of cognitive load theory to introductory accounting starting with optimising intrinsic load. The teacher exercises control over the complexity of content and offers supplementary material. This is done via the *chunking principle*: separating and sequencing the interactive elements of a topic to deal with diverse learners with various base level knowledge and schema progression. Attention is paid to reduce the extraneous load of supplemental materials, by using worked examples as opposed to multiple instances of problem-solving tasks. The design and presentation of the educational content is considered and formatted using vignettes, white space, consistent formatting, reader friendly text, and icons. This chunking of content maps directly to the levels of learning and GBA experienced in game play.

The implications for game software as a learning and assessment resource in accounting education—creating awareness, interest, and achieving learning outcomes through applied teaching methods—are strengthened through research grounded in cognitive load theory. For example, the cognitive load theory research of Mason, Seton, and Cooper (2016) demonstrated the effective of the use early achievement. A GBA early achievement will take the form of a reward or positive progress feedback to the learner in first few actions or decision-making stages to promote self-efficacy, that is the belief in their ability and capacity to achieve their goals, and motivate them to continue. The early achievement provides evidence to the learner that their investment in the process or narrative of the gamified learning experience was worthwhile. It is demonstrable proof of their ability to achieve, and this propels them further into the gamified learning experience in a seeking and inquisitive way.

15.2.2 Balancing Assessment and Game Development

For this research a game was developed for the accounting and finance technical threshold concept of time value of money. Time value of money was chosen because it is a standalone accounting threshold concept that can be encapsulated and developed for teaching, as a discrete piece of learning, to develop, contrast, and evaluate the effectiveness of the gamification of teaching and GBA. It also represented an achievable project within the researcher's constraints of time, money, technical expertise, and available technology.

The traditional time value of money pedagogy methods, using tables and derivation of formulae, are categorised as passive learning: the students' role being the recipient of knowledge transfer. To the contrary, active learning methods foster critical thinking (Biggs, 1987), with the ancillary effects of enhancing student motivation and engagement, and encouraging self-learning (Healy & McCutcheon, 2008). GBA is by design an active learning process of developing and supporting metacognition through formative assessment and constructivism. The student is the connector between assessment and learning—they are active, engaged, and a critical assessor. They make sense of information, relate it to prior knowledge, and use it for new learning. This *classroom* assessment is a vehicle to help students develop, practice, reflect upon, and analyse their own learning.

The challenge for the GBA designer is to use the mechanics and structures of game design to create assessment that is seamless in the flow of the narrative of the game and at the same time effective in measuring learning objectives. The resultant game developed for this research took the form of a non-linear, iSpring enhanced PowerPoint, published as a SCORM file in the university learning management system. Students enter the GBA classroom, and the plot-driven narrative begins. They receive instructions and scaffolding via multiple means of representation, traverse four levels of the time value of money gamified learning experience, make choices, repeat, and revisit at will. The game is self-timed. The difficulty escalates, and they build knowledge as they proceed through the levels, with Level 4 being the GBA. The GBA takes the form of an embedded quiz with feedback on each question, results, and review opportunity at the end. The data (including time on task, pathway, and student details, as well as GBA results) are sent to the teacher for intervention and feedback.

15.2.3 Learner Scaffolds: Formative Assessment

The learner is supported by a teacher within a game through the use of hard and soft scaffolds. Hard scaffolds are planned and programmed static supports based upon typical student difficulties with a task. These scaffolds are embedded at fixed points into the game learning space to assist learning. They raise awareness of learning objectives to enable connection from the game world to real world, directing conceptual understanding from simplistic reasoning to complex reasoning, and assist in maximising learning and knowledge transfer. Soft scaffolds are also planned and programmed into the game learning space by the teacher, but these scaffolds are dynamic, situation-specific aids that may or may not be encountered by the student as they move through the game. They direct learners to unexplored areas, pose troublesome questions, and encourage multiple perspectives. Both hard and soft scaffolds provide feedback to the student who becomes a *knowledge consolidator* performing formative GBA tasks during game play. Scaffolding equates to the levelling up that learner's experience in a learning game as they progress through the plot-driven narrative to more difficult concepts. During this learning in the game, they experience learning by doing and receive feedback from formative GBA, which is critical to constructing knowledge and advancing through the game to higher cognitive tasks.

15.3 The Classroom

No longer exclusively a physical space, the digital classroom is any interface where students access and interact with content. The digital learning environment facilitates new ways of representing content as well as allowing educational designers to guide the learner's interaction with the learning environment (Schrader & McCreery, 2012). In a gamified learning experience the educational designer creates and supports the activity at the interface. Assessment as learning delivered in a GBA provides a unique affordance of situated learning environments to support higher order thinking skills and problem solving.

15.3.1 Assessment as Learning

In the gamified learner centred environment, GBA is assessment as learning (Earl, 2006). The learner is situated as the connector between learning and assessment: active, engaged, and using formative classroom assessment to develop, practice, and analyse their own learning. This is a paradigm shift in the way the educationalists think about learning and assessment, for example, embedding assessment into eEnhanced texts and apps. "Games can also embed assessment into the learning process without disrupting the game flow" (Ghergulescu & Muntean, 2012, p. 356) of the learning experience—flow being the essential element to maintain engagement and motivation (Csikszentmihalyi, 2008). Activities that are assessment (Earl, 2013) are performed within the gamified learning experience as learning—both as ongoing self-assessment by the learner, and via the monitoring by the teacher, providing feedback.

Kapp's (2012) guidelines for sound pedagogical integration of GBAs into curriculum are defensible to higher education.

- Embed games into instructional programmes: introduce and explain, play, debrief learning and how game events supported learning.
- Align game objectives with learning objectives.
- Include instructional support in the form of learning scaffolds.
- Build in choice and system response for interactivity to engage, as opposed to passive conveyance of content.
- Entertainment influences instructional effectiveness, but learner's engagement with content makes learning more likely to occur.
- Provide unlimited access and encourage extended and repeated game play.

All of these criteria resonate with *gamification as learning* being analogous to assessment as learning. Students are encouraged to monitor and reflect on their own learning through activities and assessment items, adjusting their game play to achieve deeper understanding and embed learning. This is supported by continuous real time assessment and feedback—transparent information for both learners and teachers to facilitate self-paced and self-directed learning. The gamified learning experience provides information on student progress by allowing them to review levels or content already completed through replay, allowing for new choices, and seeking alternate outcomes (Kapp, 2012).

15.3.2 Gaming Features that Lead to Learning

Scaffolding and interactivity work together to afford learning for GBA. A gamified learning experience can assist leaners to construct a new knowledge base through a motivating process of actively bridging formal and informal knowledge, but when game tasks are too difficult motivation and engagement decline. To maintain confidence once the student is in the game, scaffolding that supports learning activities fosters motivation (Chen & Law, 2016); however it is the interface that determines the strength of the initial engagement with the game. Eseryel, Guo, and Law (2012) proposed an interactivity design and assessment framework to promote motivation and complex problem-solving skills.

> [I]n order to design an educational game we need to pay special attention to the functionality, game play, referentially, social, and pedagogical issues [to] target learners' motivation and complex problem-solving skills, we take the explicit interactions between players and games as a persistent cycle of making choices through the game play. (p. 260)

Eseryel et al. (2012) held that interactivity was a function of three levels: (1) interface, the physical experience and use of the game; (2) narrative, the cognitive immersion in the story; and (3) social, the collaborative opportunities with other players. Of these, the functional interactivity of play action at the interface (Level 1) is the essence of the game, allowing for suspension of disbelief (Csikszentmihalyi, 2008). The player is then immersed in the first-person perspective of the plot-driven narrative (Level 2), which motivates them to continue through until the end of the game and the integrated GBA.

15.4 The Model

15.4.1 Gamification Alignment Table

Incorporating games and GBA into a new culture of learning (Thomas & Seely Brown, 2011) involves coupling game designs, learning principles, learner engagement, and learning outcomes, by means of *gamification alignment*, that is mapping the elements and language of gaming against curriculum components. The gamification alignment table (Table 15.1) developed for this research provides a lexicon of pedagogical terms and their corresponding gaming labels. This table allows GBA designers to see at a glance the game elements they need to employ to activate the various aspects of assessment and feedback. For example, the learner performs various activities using resources and learning tools, receiving formative assessment, to complete a learning outcome, and achieve a grade. The GBA learner embodies an

Table 15.1 Gamification alignment table

Pedagogical lexicon	Gaming lexicon
Unit/course description	Story
Curriculum	Game map
Learner	Avatar
Learning outcome	Mission
Successful completion of unit	Goal
Activity	Challenge
Resources/learning tools	Artefacts
Peer/team-based learning	Team
Formative assessment	Lives
Assessment	Quest
Marks	Points/experience points
Grade	Score
Student ranking	Leaderboard
Extra activities	Side quests
HD opportunities	Bonuses
Discussion board	Chat

avatar, which faces challenges with the help of artefacts, and multiple lives, to complete a mission and receive a score.

15.4.2 Gamification Alignment Model

The integration of the pedagogical and gaming lexicons into GBA to construct a gamified curriculum requires the cooperation of the content expert, the educational design, curriculum designer, and teacher (time and money resources often meaning these are encapsulated in the one person), plus the digital learner, to "[design] an effective educational game to fulfil the unique affordances of a situated learning environment to support higher order thinking skills and problem solving skill acquisition while maintaining high student motivation" (Eseryel et al., 2012, p. 282). To achieve this, a gamification alignment model (Fig. 15.1) is proposed, paralleling the thinking skills elucidated in Bloom's (1956) taxonomy of learning with Allen's (online instructional eLearning game designers) (http://www.alleninteractions.com/about) taxonomy of gaming, to be populated with concepts and pedagogical verbs for use by educators and educational designers in planning and designing gamified learning experiences. To further illustrate the investment in time and practice spent at each level, Bergmann's (2016) flipped learning model has been incorporated. The lower order thinking skills of remembering, understanding, and applying equated to recall and memory, judgement, and consequence games, while important foundations for learning require and should demand less activity time to embed and master than higher order thinking skills of analysing, evaluating, and creating, equated to

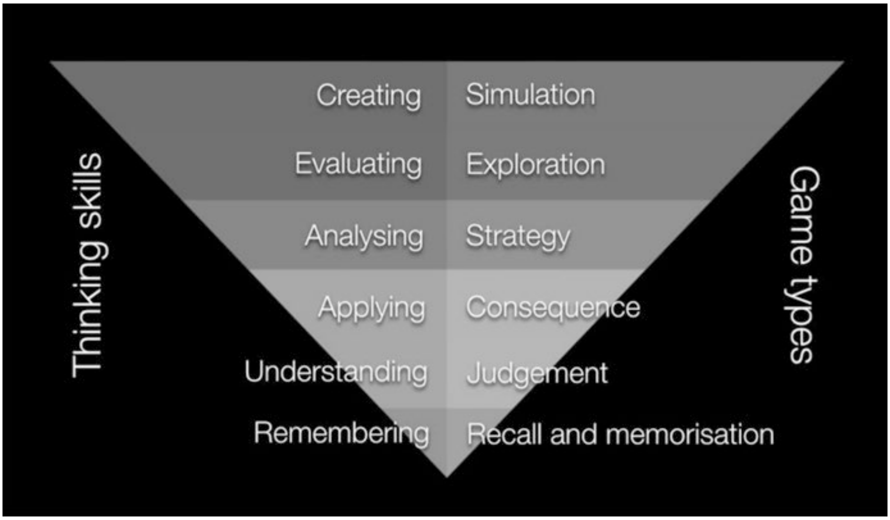


Fig. 15.1 Gamification alignment model (adapted from Allen Interactions http://www.alleninteractions.com/about; Bergmann, 2016; Bloom, 1956)

games of strategy, exploration, and simulation. The construction of the model illustrates the cumulative nature of the thinking skills. Each progressively higher level builds on and incorporates the level/s below so that at any time during higher level activities, lower levels are still being called upon.

A threshold concept game, like that created for this research, targets the gamification alignment model levels of understanding/judgement, applying/consequence, analysing/strategy, and some evaluating/exploration. As the learner moves through the levels of the game, the constructs being assessed move up the learning taxonomy levels in tandem, providing both summative and formative GBA opportunities. The formative GBA occurs in the form of incidental and purposive, embedded artefacts and clues available for collection during the game play. Experience points and scoring rules within the game are matched to learning objectives, but not just for the sake of making it game-like. Because a "raw score cannot indicate how much of the construct the student has" (Belland, 2012, p. 39), the use of formative feedback for verification and elaboration supports learners in their search for, and accumulation of, knowledge. They are not merely directed by a teacher with a summative assessment focus. Development is spontaneous and occurs in the zone of proximal development (Vygotsky, 1978)—the threshold where the learner is not yet competent and able to complete a task independently. However, "designing formative feedback that is effective in guiding students' learning, while still creating an engaging game, is difficult" (Belland, 2012, p. 31). Students are canny consumers of games and not easily misled by a flimsy façade. Engagement and motivation can be achieved through the application of the plot-driven narrative promoting student mindfulness and providing context for the learning.

Within the design phase consideration of the learning objectives to be assessed is always the core driver. A conceptual design map of the GBA for this research was created; however upon starting the actual designing, and with reference to the constructive nature of the content, an additional level was included: the first or top level under the character—Simple interest/Compound interest (Fig. 15.2). This level scaffolds the first terminology and concept that the learner must understand in order to successfully progress through the time value of money game. From here they have the tools to move to the next level, future value (building, or deriving, the time value of money formula), and the subsequent level, present value (manipulating the time value of money formula and recognising the properties and use of the discount factor).

15.4.3 Conversational Framework Methodology

Laurillard (2002) encapsulated the constructivist, social constructivist, and cognitive learning theories and presented a *conversational framework* against which to evaluate the effective use of learning technologies in higher education. The refined conversational framework adapted for this research (Fig. 15.3) situates the teacher and learner around these learning theories to illustrate where the constructed

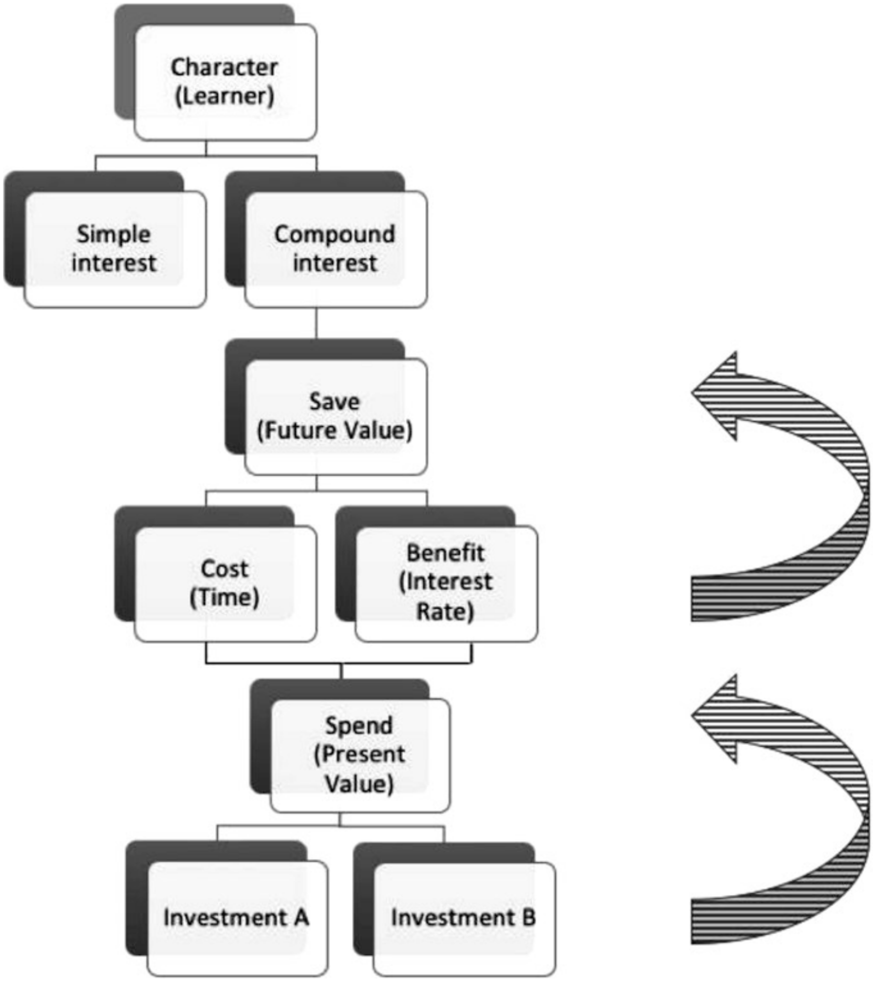


Fig. 15.2 Design map of episodic GBA with multiple repeats of the levels, as shown by the looping arrows

learning environment is described, engaged with, reflected upon, adapted, and re-described in a continuous flow of perception and knowledge construction (Laurillard, 2002). These are the interfaces: the meeting places for interaction and communication possibilities that inform the design phase and future iterations of the GBA.

This research employed the empirically validated survey instrument eGameFlow of Fu, Su, and Yu (2009) (Appendix) to examine the interface relationships between teacher and student, and teachers and students with the gamified learning experience (Laurillard, 2002), through the factors of: concentration, goal clarity, feedback, challenge, autonomy, immersion, social interaction, and knowledge improvement. Across the top of the framework is the discursive interface between

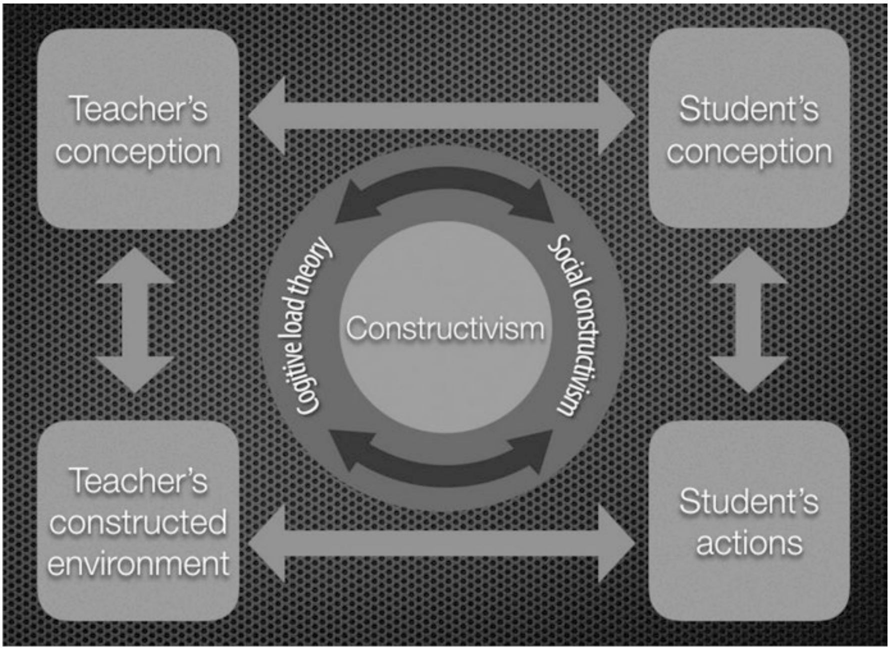


Fig. 15.3 The refined conversational framework (adapted from Laurillard, 2002)

the teacher and the student, where theories and ideas are transposed into the GBA. At this interface Hypothesis 1: Gamification of the learning experience enhances learner engagement is measured with the eGameFlow factor question sets of engagement, challenge, autonomy, and immersion. Down the right side is the adaptive and reflective interface of the student with the content, where students consider their understanding and experience with the GBA. At this interface is Hypothesis 2: Gamification of the learning experience enhances perception of self-efficacy is measured with the factor question sets of concentration, goal clarity, feedback, challenge, autonomy, and knowledge improvement. At the bottom of the framework is the interactive interface between the student and the non-human content, where learning objectives, learner actions, and GBA and feedback take place. This is the interface for Hypothesis 3: Gamification of the learning experience enhances learner performance and is measured by the GBA, in conjunction with the factors question sets of goal clarity, feedback, challenge, and knowledge improvement. Finally, on the left side is the adaptive and reflective interface where the teacher interacts with the non-human content to examines students' actions, GBA outcomes, and needs. This interface of Hypothesis 4: Gamification of the learning experience enhances education design will be described by discussion board comments and future focus group analysis.

15.5 The Game

15.5.1 Gamified Learning Experience for Accounting and Finance Technical Threshold Concepts

To demonstrate GBA as learning, a gamified learning experience with choices, repeats, and scaffolds to match the lexicons of the gamification alignment table was created. Ideally the data gathered from the students' game play would enable the educator to trace movement and time within the gamified learning experience to examine how the results of the GBA matched to the learning objectives. This gamified learning experience is at the beginning of the learners' higher education with the aim to embed and assess technical threshold concepts. It employs a gamification curriculum design of accounting and finance technical threshold concepts to improve user engagement, learning, assessment outcomes, demonstrated through GBA.

15.5.2 Time Value of Money GBA

To design a GBA for the threshold concept of time value of money, each of Meyer and Land's (2006) five characteristics of threshold concepts was taken into consideration (Table 15.2).

In Table 15.2, the GBA column gives generic descriptors which can be applied to any gamified threshold concept. The GBA for the time value of money column is specific to the game developed for this research. The following Figs. 15.4, 15.5, 15.6, 15.7, and 15.8 are taken from the time value of money game and illustrate the translation and depiction of each of these characteristics.

15.5.3 GBA Performance Outcomes and More

The last level of the game is the summative GBA where students are asked to apply all their constructed knowledge of the threshold concept to a similar scenario to those experienced within the game (eight application of learning questions). There are also six multiple-choice theory questions. For each of the 14 answers students provide, immediate feedback is given (Fig. 15.9). At the end of the GBA their results are presented to them as shown in Fig. 15.10. They are able to review their answers, the correct answers, and comprehensive explanations and feedback.

Forty ($N = 40$) volunteer students completed the prototype GBA. The sample group consisted of equal numbers of males and females. Average age was just over 32 years ($M = 32.18$). There were 27 on campus students and 17 online students, and 23 were undergraduate and 17 post graduate. Twenty-three of the students in the

Table 15.2 Characteristics of threshold concepts matched to the GBA for the time value of money

Characteristic	Threshold concept	GBA	The GBA for the time value of money
Transformative	• Requiring a significant shift in thinking or world view	• Immersion in the GLE as a character or avatar with different scripts, backgrounds, and sets of attributes	• Choice of two avatars to represent learner
Irreversible	• Cannot be unlearned (they are part of semantic memory)	• Artefacts, improvements, and attributes are carried through to future missions	• Information is presented via artefacts and can be revisited
Integrative	• Merged seamlessly into the existing schema	• Progression through the GLE is cumulative	• Learners collect more information by moving through the story
Represent a boundary	• Movement into a new level of understanding	• Progression through the worlds or scenes of the GLE with ever increasing levels of complexity, understanding, and skill required	• Each level builds on the previous learning
Troublesome	• Counter-intuitive, incoherent to current way of thinking and knowing	• Progression requires alternate approaches to a task to achieve a favourable outcome	• Learner is asked to make choices and consider outcomes

sample were domestic, and 17 were international. Twenty-one students reported they had studied the threshold concept time value of money before and 19 reported they had not. The data collected from the volunteer students GBA and the eGame-Flow survey were analysed using SPSS Statistics software. A one-way ANOVA was performed to determine if there were any statistical differences between the mean total scores of the students in any of these sub-groups. The average total score for all students who completed the prototype was 10/14 or 71.43% (M = 10). There were no significant differences found at the 0.05 level in the mean score achieved within any of the sub-groups. The results are reported in Table 15.3.

Typically, the students' first encounter with time value of money occurs in financial accounting followed up with financial and managerial accounting, and then advanced accounting or corporate finance units. Despite this repeated exposure, students find this threshold concept challenging (Siriwardane, 2014). Student assessment results in introductory accounting subjects consistently return high failure rates of 35–45% (e.g., Doran, Bouillon, & Smith, 1991; Kealy, Holland, & Watson, 2005). The GBA mean score of well above 50% for a technical threshold concept and the consistency of scores among the sub-groups is promising for Hypothesis 3, that gamification of the learning experience enhances learner performance.

The eGameFlow survey (Fu et al., 2009), administered after the GBA, required students to rate their experience on a five-point Likert scale (with 0 = strongly disagree, 1 = somewhat disagree, 2 = neutral, 3 = somewhat agree, 4 = strongly agree)

Fig. 15.4 Transformative: First scene—avatar choice

via agreement to a number of statements for each of the factors of: concentration (activities support concentration while minimising cognitive overload), goal clarity (task are clearly explained at the start), feedback (scaffolds learning and movement to higher levels), challenge (activities are matched to skill level and increase in difficulty), autonomy (opportunities are provided for initiative and control), immersion (investment in the story and outcome), and knowledge improvement (increases knowledge and skills in line with curriculum). The factors have varying numbers of items; therefore for comparison, scales were created using the mean of the items to represent each factor. Mean values, standard deviations, and coefficients of variation (CV) were calculated for each factor. CVs were all less than 1 and indicated low

Fig. 15.5 Irreversible: Hard scaffold active object "Clue" example

variation of data. Cronbach's alpha was used to test for factor reliability, and all items were retained. The results are shown in Table 15.4.

Mean values for all factors rated favourably between 2.74 (immersion) and 3.59 (feedback). Hypothesis 3 is therefore supported by the encouraging factor scores related to performance of learning, specifically those of goal clarity, feedback, challenge, and knowledge improvement. Feedback factor had the highest mean coupled with the lowest standard deviation and coefficient of variation, revealing the least discrepancy between opinions of participants. This was closely followed by challenge and goal clarity in high mean score/low standard deviation/coefficient of variation.

Pearson correlation was used to determine the strength of any relationship between the eGameFlow factors and total correct answers (Table 15.5). Of all the factors, feedback showed a significant correlation with total correct ($r = 0.386$), followed by concentration and immersion. This is perhaps the strongest support of Hypothesis 3, gamification of the learning processes leading to enhanced learning outcomes, because arguably the major attribute of the gamified curriculum is the inclusion of learning scaffolds during the individual learner's progress through the experience. The significance of the feedback factor lends weight to the success of the GBA. Of note, although autonomy showed a slight negative correlation ($r = -0.085$, not significant) an examination of the statements that made up this factor and the data collected from participants revealed that this finding was likely due to most reporting they had not taken the opportunity to repeat stages of the game (Appendix, Item A3).

Fig. 15.6 Integrative: From the question prompt, clicking on the piggy bank artefact, the hard scaffold fixed information emerges

However, the measurement of performance outcomes alone does not definitively show if "change occurs as a direct result of experience [within] a game" (Schrader & McCreery, 2012, p. 13). To ascertain causation it is imperative to test for other non-assessment metrics including engagement and self-efficacy as postulated in Hypothesis 1: Gamification of the learning experience enhances learner engagement, and Hypothesis 2: Gamification of the learning experience enhances perception of self-efficacy. Hypothesis 1 is supported overall by the mean scores (Table 15.4) for the factors of concentration, challenge, autonomy, and immersion, which describe and measure engagement. Hypothesis 2 is likewise supported by the mean scores (Table 15.4) for the factors of concentration, goal clarity, feedback, challenge, autonomy, and knowledge improvement, which describe and measure efficacy.

In consideration of Hypothesis 4: Gamification of the learning experience enhances education design, Belland (2012) acknowledged that the collection of quality data from GBA to facilitate designing appropriate assessment was central to designing games and called for measurement of learning during game play. Building on this and using an iterative process of design, evaluate, and reflect, the GBA in this research provides proof of concept for GBA in accounting education against the traditional use of tables and derivation of formulas. Further evidence

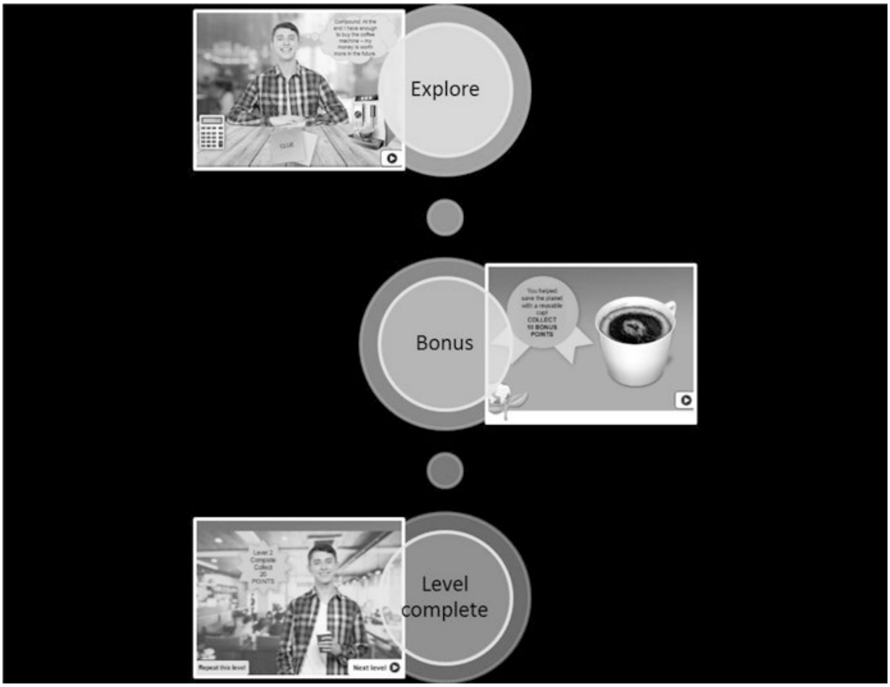


Fig. 15.7 Represent a boundary: Finding additional information and points through exploration

was gathered through the discussion board on the learning management site. Students commented:

> The quiz was very informative, and the layout and instructions were easy to follow. The content was presented clearly and was interactive, and the problem math questions were creative and short. The maths concepts were challenging however I feel confident to apply these to future finances. (Student MM)

> I am glad that finally someone [is] asking about what makes it easier to learn or how to present materials. It was very informative, short and succinct. As an international student with English as my second or even third language, I personally prefer pictures and diagram to learn new concepts. It is quite challenging and time consuming to go through the whole text, we have to convert sometimes to our own language and then figure out what is the message of text, whereas, material presented by diagram, chart, etc. We would like to learn something that leads us to final answer. (Student AM)

15.6 Beyond the Limits

The game was published as a SCORM file into the university's learning management system and provided some non-invigilated time on task data for inferential discussion, but pathways and repeats (part or full), game score accumulation, and

Fig. 15.8 Troublesome: Opportunity to reflect on the choice of "money today or at the end of the year"

bonuses found were beyond the scope of the budget for this research. Repeats, improvement, and perceived in-game experience, that is, how the learners responded to feedback and scaffolds embedded in the GBA, are all metrics in need of further investigation.

15.7 Contribution

Advances in technology have enabled more sophisticated GBAs to be designed and created by education practitioners without reference to costly external specialists. This research has produced a validated instructional design model: a replicable model for mapping and creating technical threshold concepts, specifically in the accounting and finance. It will readily translate further into other business threshold concepts, and those of other disciplines, with potential to move into broader andragogy and heutagogy: financial advisors and banks, for client awareness and education. Interest has already been shown for inclusion in eEnhanced texts.

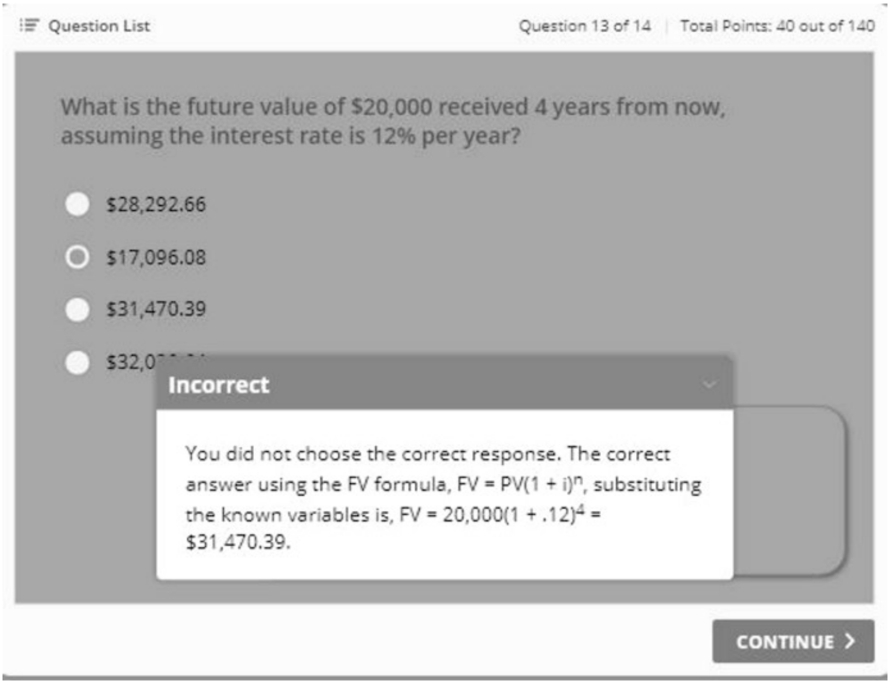


Fig. 15.9 Progressive feedback provided for each answer

Appendix: eGameFlow Survey

(Five-point Likert scale responses with 1 = completely disagree and 5 = completely agree)

Factor	Item number	Content
Concentration	C1	The gaming activities are related to the learning task
	C2	I remained focused on the game
	C3	I was not distracted from the learning task
	C4	I was not burdened by tasks that seemed unrelated
	C5	The workload of the game is adequate
Goal clarity	G1	Game goals were presented at the beginning of the game
	G2	Game goals were clear
	G3	Intermediate goals were presented at the beginning of each scene
	G4	Intermediate goals were clear
Feedback	F1	I received feedback on my progress in the game
	F2	I received immediate feedback on my actions
	F3	I was notified of new tasks immediately

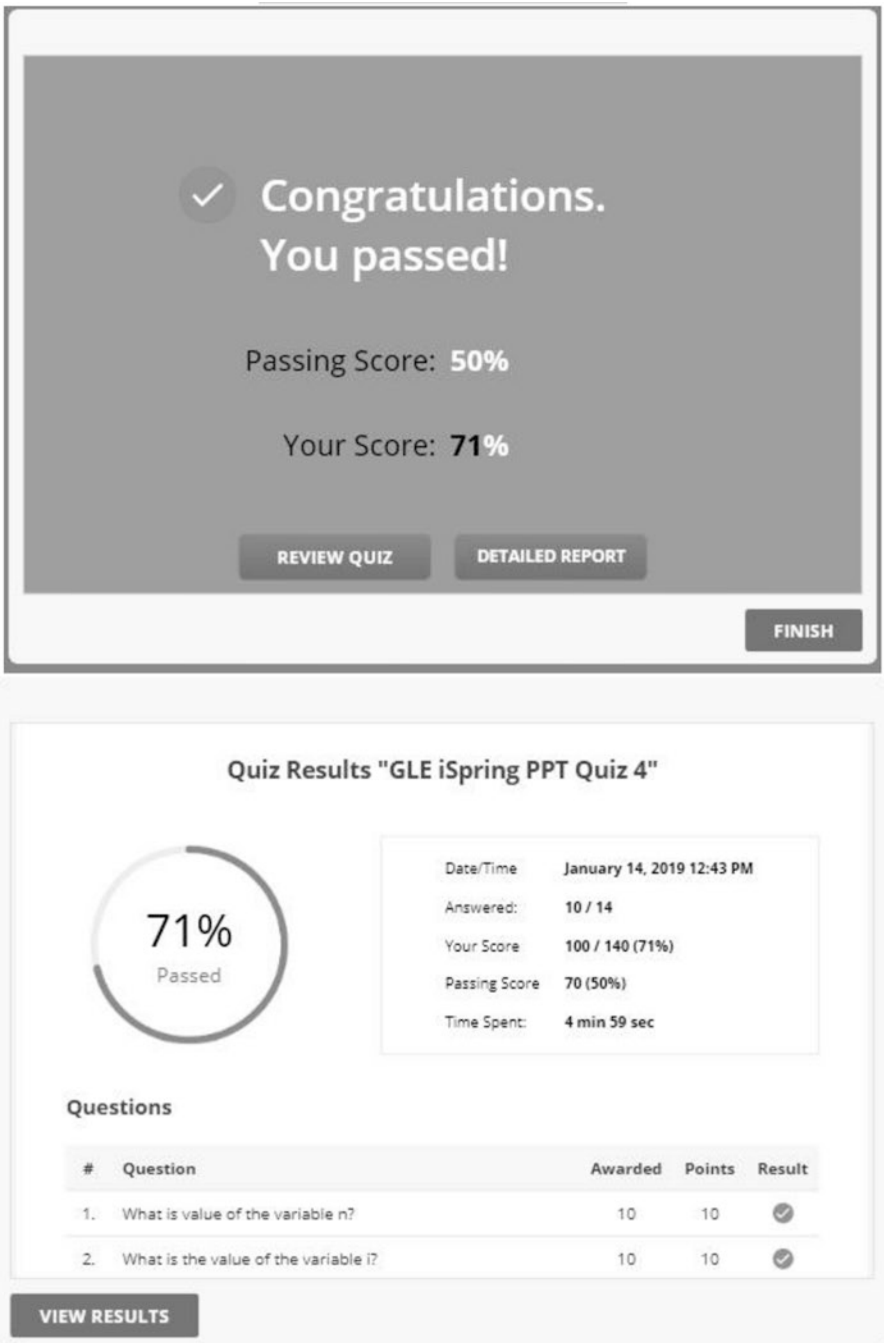


Fig. 15.10 Student view of GBA results

Table 15.3 Total scores achieved on GBA, full sample and by sub-group

Group	*N*	*M*	SD	Sig.
Sample	40	10	2.428	
Male	20	10.35	1.981	0.183
Female	20	9.65	2.815	
On campus	27	9.74	2.740	0.642
Online	13	10.54	1.561	
Undergraduate	23	10.57	2.465	0.695
Post graduate	17	9.24	2.223	
Domestic	23	10.04	2.458	0.578
International	17	9.94	2.461	
Prior study	21	9.86	2.632	0.901
No prior study	19	10.16	2.243	

Table 15.4 Mean values of sample ($N = 40$) eGameFlow factors and corresponding hypotheses

Factor	*M*	SD	CV	Cronbach's alpha	H1	H2	H3
Concentration	3.15	1.001	0.32	0.830	✓	✓	
Goal clarity	3.33	0.83	0.25	0.845		✓	✓
Feedback	3.59	0.761	0.21	0.811		✓	✓
Challenge	3.32	0.82	0.25	0.910	✓	✓	✓
Autonomy	2.76	1.126	0.41	0.615	✓	✓	
Immersion	2.74	1.077	0.39	0.892	✓		
Knowledge improvement	3.29	1.608	0.49	0.846		✓	✓

Factor	Item number	Content
Challenge	H1 H2 H3 H4 H5	The game provided hints that helped me with the challenges The game provided other supports to help me with the challenges The difficulty of challenges increased as my knowledge improved The game provided new challenges at an appropriate pace The game provided different levels of challenges tailored to my needs
Autonomy	A1 A2 A3	I felt a sense of control and impact over the game I understood the stages of the game I used the opportunity to repeat stages of the game
Immersion	I1 I2 I3 I4 I5	I forgot about time passing while I played the game I became unaware of my surroundings while I played the game I temporarily forgot about other things while playing the game I became involved in the game I felt emotionally involved in the game

Table 15.5 Correlation between GBA total correct answers and eGameFlow factors

Factor	Pearson correlation (r) Total correct
Total correct	1
Concentration	0.310
Goal clarity	0.071
Feedback	0.386[a]
Challenge	0.157
Autonomy	−0.085
Immersion	0.248
Knowledge improvement	0.188

[a]Correlation is significant at the $p \leq 0.05$ level

Factor	Item number	Content
Knowledge improvement	K1 K2 K3 K4 K5	The game increased my knowledge I understood the basic idea of the game straight away I applied my knowledge within the game The game motivated me to integrate my knowledge straight away I want to know more about the concept taught in the game

Note. Adapted from Fu et al.'s (2009) post validity and reliability tested instrument

References

Arnab, S., Brown, K., Clarke, S., Dunwell, I., Lim, T., Suttie, N., … de Frietas, S. (2013). The development approach of a pedagogically-driven serious game to support relationship and sex education (RSE) within a classroom setting. *Computers & Education, 69*, 15–30. https://doi.org/10.1016/j.compedu.2013.06.013

Belland, B. (2012). The role of construct definition in the creation of formative assessments in game-based learning. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 29–42). New York, NY: Springer. https://doi.org/10.1007/978-1-4614-3546-4_3

Bergmann, J. (2016, October). *Flipped learning network*. Paper presented at the Flipped Learning Symposium, Adelaide.

Biggs, J. B. (1987). *Student approaches to learning and studying*. Camberwell, UK: Australian Council for Educational Research.

Biggs, J. B. (2003). *Aligning teaching for constructing learning*. Heslington, UK: Higher Education Academy.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals*. London, UK: Longman.

Chen, C.-H., & Law, V. (2016). Scaffolding individual and collaborative game-based learning in learning performance and intrinsic motivation. *Computers in Human Behavior, 55*, 1201–1212. https://doi.org/10.1016/j.chb.2015.03.010

Crotty, M. (1998). *The foundations of social research: Meaning and perspective in the research process*. Thousand Oaks, CA: Sage.

Csikszentmihalyi, M. (2008). *Flow: The psychology of optimal experience*. New York, NY: Harper Collins.

de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science, 38*(2), 105–134. https://doi.org/10.1007/s11251-009-9110-0.

De Villiers, R., & Hess, A. (2019). *The magic cloak: Releasing the creative genie in your workplace*.

Doran, B. M., Bouillon, M. L., & Smith, C. G. (1991). Determinants of student performance in accounting principles I and II. *Issues in Accounting Education, 6*(1), 74–84.

Earl, L. M. (2006). Assessment as learning. In W. D. Hawley (Ed.), *The keys to effective schools: Educational reform as continuous improvement* (pp. 85–98). Thousand Oaks, CA: Sage.

Earl, L. M. (2013). Assessment for learning, assessment as learning: Changing practices means changing beliefs. *Assessment and Learning, 2*, 1–5.

Eseryel, D., Guo, Y., & Law, V. (2012). Interactivity design and assessment framework for educational games to promote motivation and complex problem-solving skills. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 257–285). New York, NY: Springer.

Fu, F.-L., Su, R.-C., & Yu, S.-C. (2009). EGameFlow: A scale to measure learners' enjoyment of e-learning games. *Computers & Education, 52*(1), 101–112. https://doi.org/10.1016/j.compedu.2008.07.004

Ghergulescu, I., & Muntean, C. H. (2012). Measurement and analysis of learner's motivation in game-based e-learning. In D. Ifenthaler (Ed.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 355–378). New York, NY: Springer.

Healy, M., & McCutcheon, M. (2008). Engagement with active learning: Reflections on the experiences of Irish accounting students. *Irish Accounting Review, 15*(1), 31–49.

Hickman, L. A., Neubert, S., & Reich, K. (2009). J*ohn Dewey between pragmatism and constructivism*. Oxford: Fordham University Press.

Kapp, K. (2012). *The gamification of learning and instruction*. Somerset, UK: Centre for Creative Leadership.

Kapp, K. (2016). Choose your level: Using games and gamification to create personalized instruction. In M. Murphy, S. Redding, & J. Twyman (Eds.), *Handbook on personalized learning for states, districts, and schools* (pp. 131–143). Philadelphia, PA: Temple University, Center on Innovations in Learning. Retrieved from http://www.centeril.org/2016Handbook/resources/Cover_Kapp_web.pdf

Kealy, B. T., Holland, J., & Watson, M. (2005). Preliminary evidence on the association between critical thinking and performance in principles of accounting. *Issues in Accounting Education, 20*(1), 33–49. https://doi.org/10.2308/iace.2005.20.1.33

Laurillard, D. (2002). *Rethinking university teaching: A framework for the effective use of learning technologies* (2nd ed.). Abingdon, UK: RoutledgeFalmer.

Laurillard, D. (2016). Learning number sense through digital games with intrinsic feedback. *Australasian Journal of Educational Technology, 32*(6), 32–44. https://doi.org/10.14742/ajet.3116

Mason, R., Seton, C., & Cooper, G. (2016). Applying cognitive load theory to the redesign of a conventional database systems course. *Computer Science Education, 26*(1), 68–87. https://doi.org/10.1080/02336408.2016.1160597

Meyer, J. H. F., & Land, R. (2006). *Overcoming barriers to student understanding: Threshold concepts and troublesome knowledge*. Abingdon, UK: Routledge.

Millwood, R. (2014). *The design of learner-centred, technology-enhanced education (Doctoral dissertation)*. Boston, MA: University of Boston.

Mostyn, G. R. (2012). Cognitive load theory: What is it, why it's important for accounting instruction and research. *Issues in Accounting Education, 27*(1), 227–245. https://doi.org/10.2308/iace-50099

Schrader, P. G., & McCreery, M. (2012). Are all games the same? In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 10–28). New York, NY: Springer.
Simon, K. (2016). Where is e-learning headed? Relating simulation and gamification to a possible shift in education. *Argumentum, 12*, 201–215.
Siriwardane, H. (2014). Kiono Bag Boutique: A serial case for introductory financial accounting. *Accounting Education, 23*(1), 95–113. https://doi.org/10.1080/09639284.2013.847307
Tennant, M., McMullen, C., & Kaczynski, D. (2010). *Teaching, learning and research in higher education: A critical approach*. New York, NY: Routledge.
Thomas, D. T., & Seely Brown, J. (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.

Chapter 16
Emerging Practices in Game-Based Assessment

Vipin Verma, Tyler Baron, Ajay Bansal, and Ashish Amresh

16.1 Introduction

Serious and educational games have been a subject of research for a long time. They usually have game mechanics, game content, and content assessment all tied together to make a specialized game intended to impart knowledge of the associated content to its players (Van Eck, 2006). While this approach is good for developing games for teaching highly specific topics, it consumes a lot of time and money. Being able to re-use the same mechanics and assessment methods for creating games that teach different contents would lead to a lot of savings in terms of time and money. The Content Agnostic Game Engineering (CAGE) Architecture mitigates the problem by disengaging the content from game mechanics (Baron, 2017). Moreover, the content assessment in games is often quite explicit in the sense that it interrupts the flow of the players and thus hampers the learning process, as it is not integrated into the game flow. Stealth assessment can be beneficial in such cases to keep the player engagement intact while assessing them at the same time (Shute, 2011). Integrating stealth assessment into the CAGE framework in a content-agnostic way will increase its usability while also decreasing the time and cost of developing in-game assessment.

The word "agnostic" has Greek origin which translates to "not known". The word content agnostic in the context of an educational video game emphasizes the fact that the game mechanics are independent of the target content domain of the game. In the following sections, this chapter will dive into the theory of motivation, followed by the definition of game mechanics, content, and assessment. Then the emerging need for content-agnostic assessment will be discussed, and how the motivation can help in effective learning. It will be followed by the approaches to make the assessment unobtrusive and then methods to quantify the learning gains.

V. Verma (✉) · T. Baron · A. Bansal · A. Amresh
Arizona State University, Chandler, AZ, USA
e-mail: vverma9@asu.edu; tjbaron@asu.edu; ajay.bansal@asu.edu; amresh@asu.edu

D. Ifenthaler, Y. J. Kim (eds.), *Game-Based Assessment Revisited*, Advances in Game-Based Learning, https://doi.org/10.1007/978-3-030-15569-8_16

16.2 Self-Determination Theory

Motivation is to be moved to do something and can be categorized as intrinsic or extrinsic (Ryan & Deci, 2000). Intrinsic motivation involves an innate desire to achieve an outcome while extrinsic motivation uses external rewards to drive a person towards the desired outcome. Since people learn better while acting on their natural tendencies, intrinsic motivation can actuate better and higher-quality learning (Ryan, LaGuardia, & Rawsthorne, 2005). Inherent interactivity, challenge, fantasy, and curiosity in the video games help in sustaining the intrinsic motivation of the players during the game-play (Freire et al., 2016; Malone, 1981). Avatar customization in the game *Zombie Apocalypse* is an example of intrinsic motivation (Birk, Mandryk, & Atkins, 2016). Extrinsic motivation such as a grade, on the other hand, can be detrimental to learning.

Self-Determination theory (SDT) specifies the degree to which a person is intrinsically motivated to improve themselves (Chatzisarantis, Biddle, & Meek, 1997). Unfamiliar gaming environment motivates players to master the environment and learn new skills in the process. As shown in Fig. 16.1, it has three components: autonomy, relatedness, and competence. The need for autonomy is related to the sense of control over one's surroundings (Deci & Vansteenkiste, 2004). Video games present autonomy by providing its players with a set of choices and allowing its players to follow their own path towards an objective. Customization of player avatar in *Second Life* (Linden Labs, 2003) and branching narratives in *Dragon Age: Origins* (BioWare, 2009) are some examples of autonomy manipulation within games. The need for relatedness revolves around a person's desire to have a sense of belongingness among their peers, competitors, and instructors. Multiplayer games allow the need for relatedness to be fulfilled by allowing a person to play with others. Multiplayer group (clan) play in *League of Legends* (Riot Games, 2009) and

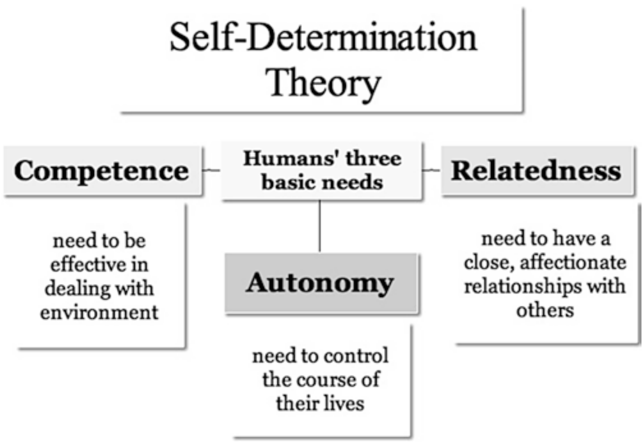


Fig. 16.1 The three components of SDT (Deci & Ryan, 2008)

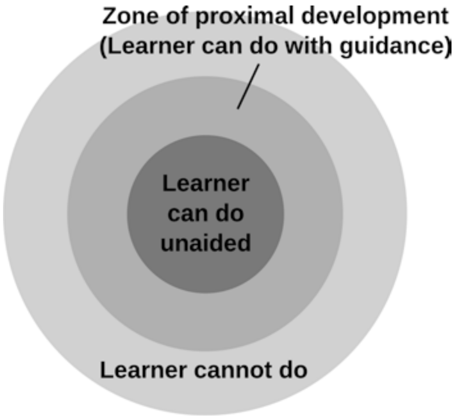


Fig. 16.2 Zone of proximal development (Vygotsky, 1978)

match-making algorithms in online multiplayer games like *Brawl Stars* (Supercell, 2017) are some ways to keep relatedness intact. The need for competence relates to a person's ability to attain learning objectives. Video games can promote competence by providing incremental objectives with an increasing difficulty level. Ryan, Rigby, and Przybylski (2006) used SDT to explain the motivation pull in video games. Their experiments suggested that a video game which is autonomy-friendly, relatedness-invoking, and competence-evoking could help sustain the motivation levels in a video game (Sørebø & Hæhre, 2012). Situations that thwart these three needs undermine the intrinsic motivation of an individual. However, high autonomy makes it difficult to compare the evidence gathered from two different players, and increased relatedness can lead to construct-irrelevant variances, thus thwarting the assessment process. So, a delicate balance is required to keep both engagement and assessment intact, simultaneously.

Tasks that lead a learner to the cusp of their abilities affect their engagement and motivation positively (Gee, 2003) and help them remain in flow (Csikszentmihalyi, 1975). Vygotsky (1978) used the term called the zone of proximal development to describe this edge of abilities. The zone of proximal development is the difference between what the learners can do without any assistance and what they cannot do even if they had help. This zone contains the skills that the learner can attain when guided properly. A learner with high skill level when presented a low-level challenge will get bored, while introducing a difficult task to an unskilled learner will make them anxious or frustrated. Thus, it is advisable to keep the learner in the zone of proximal development by keeping the optimum level of challenge suitable for their current skill level (Fig. 16.2).

16.3 Game Mechanics, Content, and Assessment

Sicart (2008) defined the term game mechanics as the ways in which players interact with the gaming environment. A game mechanic can be understood as a verb, for example, climbing, running, whistling, shooting, grabbing, and switching weapons

(Järvinen, 2008). Mechanics are a means to overcome the challenges encountered during the game-play or any desired outcome that requires an effort (Sicart, 2008). For example, stabbing is a basic mechanic found in the game *Shadow of the Colossus* which involves plunging a weapon into the body of the colossus to injure them (Team Ico, 2005).

The content domain of a game is the topic which the game is trying to teach its players (Baron & Amresh, 2015). For example, consider a game designed to teach encryption methods to its players. The content domain for this game would be Cryptography. Unlike game mechanics, which are important pieces in any video games, content domain is defined only in educational video games. Commercial entertainment games are not meant for teaching purposes; hence they do not need to define a content domain. Defining a content domain is a crucial part in the design of an educational video game because its aim is to impart skills pertaining to that domain. It is thus a common practice to specify a content domain and then design the educational game around it.

Assessment is a process which uses data to determine if the learning goals are met (Chin, Dukes, & Gamson, 2009). Consider the game from the previous example in which the content domain is Cryptography. Then the purpose of the in-game assessment would be to find out if the player has learned how to use basic encryption mechanisms taught by the game such as the Caesar cipher. Assessment is critical to the growth of serious games and the quantification and validation of learning so that their benefits can be justified over other instructional strategies (Ritterfeld, Cody, & Vorderer, 2009). Assessment and learning should happen simultaneously in an educational game so that the players are aware of their current skill and can progress towards the learning outcome accordingly. Setting up the assessment is equally important as defining the content and mechanics for an educational video game. In level-based games, the level progression will be governed by the assessment, as players will be allowed to progress further in the game only if they demonstrate the ability to clear the previous set of challenges. In the absence of an assessment, the level of game progress will not be an indicator of the skill level of the player.

The two most pertinent questions while designing any assessments are: what and why (Plass et al., 2013). That is, what variables need to be measured and why they need to be measured in order to accurately assess student progress. In educational games, learning outcomes are the variables that are measured to gauge the effectiveness of learning employed in the game. Three categories of variables exist during an educational assessment: general trait variables, general state variables, and situation-specific variables. Trait variables (such as executive functions, verbal ability, and spatial abilities) are relatively stable and are usually not targeted in educational video games. Typically, the aim of educational games is to improve the state variables (such as subject-specific knowledge and meta-cognition) while keeping the situational variables at their optimal level for maximum learning to occur. Situational variables (such as emotional state, engagement, and cognitive load) will change as a result of the player's interactions with the gaming environment. Game design affects the situational variables to a greater extent, and thus it is important to follow game design principles that optimize these variables to keep the player in a zone of proximal development.

Confounding results may occur during an assessment procedure due to several reasons (Plass et al., 2013). Motor skills, content irrelevant skills, and emotions are several potential confounding variables. For example, a game that requires its learners to tilt a tablet device in order to guide a ball to the correct answer could lead to an incorrect observation if the learner tilted the device too quickly and guided the ball to the wrong location despite having the required skill to answer it correctly. Similarly, a game which involves chemical equation balancing may be confounded by the need to know about basic algebra. Further, situations that lead to different results when people respond differently under different emotional states could present a potential confound to the assessment process. It is important that these variables be taken care of during the assessment process. It is problematic if a student is answering incorrectly because of these reasons despite having the required level of competency.

16.4 Disconnecting the Mechanics and Assessment from Content

Previously, commercial games have been used for educational purposes (Van Eck, 2006). Using commercial off-the-shelf (COTS) games for learning is cost-effective and thus gaining acceptance owing to its practicality. However, they pose various challenges as commercial games were not designed for learning. Very limited topics can be taught using COTS games, which might be neither complete nor accurate. These games may cover a large range of content, as a breadth approach, or they may focus on a narrow and specialized topic, as a depth approach. Games that take a depth approach to the content may have missing contents, while the games that take the breadth approach may have missing topics within the content. The depth approach focuses on few topics with lots of detail, while the breadth approach focuses on several topics generally. However, the absence of relevant topics and contents causes a state of cognitive disequilibrium which promotes the thinking and learning of its players in order to attain equilibrium (Kibler, 2011). This persistent cycle of cognitive disequilibrium and equilibrium helps the players engage to the game-play and maintain flow (Csikszentmihalyi, 1975). However, the missing content needs to be addressed using either the traditional classroom activities or through the game itself. But the flow will be interrupted if the players are asked to stop the game to be educated on the missing content. Thus, COTS-based games are detrimental to the flow experience of the player (Van Eck, 2006). This suggests that the ideal solution is to link the game content domain with the game mechanics in order to obtain an optimal flow experience. However, linking the two may cause another problem. For example, imagine that you developed an educational video game which is designed to teach chemical equation balancing with an embedded assessment to evaluate the learning progress. Over time, a developer may decide to create a new game to teach basic cryptographic encryptions. The problem that you will find is that if you can use the same game to teach encryption as well, it would be really difficult to teach and assess the learning of encryptions using it. You may need to make many modifications to the game to teach the encryptions which would need a substantial amount of time and effort. As an alternative, you can also develop an entirely new game from scratch, which after a certain point may be easier than trying to modify the original game.

To mitigate this problem, one can design game mechanics which are content-agnostic, i.e. mechanics which are independent of the content being taught by the game. However, this may cause several other problems. The first problem is the same which is encountered when using COTS games for learning, as it can lead to inaccurate and incomplete content (Van Eck, 2006). However, this problem can be reduced if the learning and assessment strategies are taken into consideration during the early stages of game design. Baron (2017) has provided a game development framework called CAGE which helps in creating a content-agnostic game. The second problem that may arise is the issue of generalizability. It may be boring to play multiple games for learning different contents, all of which employ the same game mechanics, as the mechanics will become difficult to enjoy after a while. Further, there exist some specialized skills which require highly specific training that could be very difficult to fit to other content. So, it is difficult to create a single game which can address multiple content domains. However, this should be kept in mind while developing a game and accommodated using the adaptive game design and feedback capabilities to palliate this problem to a considerable extent. Moreover, it will be better over the current state where a specific game is required for each type of content and assessment.

16.5 Stealth Assessment

There are three types of assessments depending on the time when assessment takes place (West & Bleiberg, 2013). They are diagnostic, formative, and summative. Diagnostic assessments occur prior to delivering instruction to measure the prior knowledge of a student. It can be used to design the delivery of information before a student starts learning. Formative assessments monitor the student's understanding during the learning and can be used to plan the subsequent learning strategy according to the changing level of the player. Based on the continuous evaluation of the student, it can be used to provide ongoing feedback, remediate misconceptions, and dynamically adapt the learning as the learning progresses. Its purpose is to improve student learning by keeping them in the zone of proximal development. Summative assessment occurs after the learning process to evaluate overall achievement summary of a student's performance. Summative assessments inform whether the student has attained the required knowledge or not. Summative assessments are usually high stakes and answer questions such as whether the employee should be promoted, should a player be allowed to progress to the next level, or what grade or SAT score should be assigned to a student. Formative assessments provide an opportunity to rectify mistakes without any grave penalties, while summative assessments do not give a chance to correct errors.

Christel Moors, head of a middle school in Atheneum, Bree, dreams of a school devoid of grades (Renard, 2016a). Her school has removed all the exams and is striving towards a system free from grades and tests, which helps reduce the stress and anxiety levels of students. They believe in formative assessments instead of the grades calculated via summative assessments. The school also thinks that self-determination theory is the way to implement it, and they only talk about a student in terms of his/her strength and weakness instead of grades. To achieve autonomy for students, the instructional strategy needs to move from traditional methods to interactive ones with choices (Renard, 2016a). Students should be allowed to be themselves with the learning activities that fit their world. By doing this, students will be more engaged to the learning material, as they own their learning process. The process involves many challenges for students to accomplish their goals, and they are free to decide which pathway to follow at their own pace. A student should feel connected to his/her peers and teachers in order to be able to make mistakes and learn from them, which follows the principle of relatedness. Further, every student should have a positive self-image and feel competent enough to take on new challenges to obtain satisfactory results. This way each student will have their own success story with a boost in self-confidence. A student who is self-driven, connected with peers, and confident will be better motivated to learn (Renard, 2016a).

Bellotti, Kapralos, Lee, Moreno-Ger, and Berta (2013) suggest incorporating the assessment into the game itself, known as stealth assessment which aims to remove the demarcation between learning and assessment (Moreno-Ger, Martinez-Ortiz, Freire, Manero, & Fernandez-Manjon, 2014). Also, Shute and Ventura (2013) proposed learning games as an alternative to traditional learning with a benefit of adjusting the learning to the level of a struggling student with the help of an embedded stealth assessment. They argued that the classroom learning progresses at its own pace with little regard to a single struggling student. However, student interaction with the gaming world can be analysed at run-time or later to quantify the learning gains. Run-time analysis can be used for personalizing the learning of an individual student by augmenting the game with the help of dynamic adaptation and actionable feedback to improve learning. Formative stealth assessment improves the accessibility for the customized learning to happen (Renard, 2016b). It helps in obtaining the current standing of the student and the objective that they are working towards while helping them thrive towards it.

Stealth assessment is based on Evidence-centred design (ECD), which itself consists of five layers where the assessment design decisions take place (Mislevy, Almond, & Lukas, 2003). Information about the content domain of interest is gathered in the first layer, called the *Domain Analysis* layer. Thus, information is then used to build assessment arguments in the second layer, which is the *Domain Modelling* layer. These assessment arguments are converted into the specific tasks in the third layer, called the *Conceptual Assessment Framework* layer. In the fourth layer, which is the *Assessment Implementation* layer, the tasks are presented to the students, and their responses are analysed. *Assessment Delivery* layer is the last one where the assessment is reported. All these layers are guided by the third layer of *Conceptual Assessment Framework*, which consists of three models: a competency model, a task model, and an evidence model. The competency model, also called the student model, composes of variables representing student skills and knowledge that need to be assessed (Mislevy, Behrens, Dicerbo, & Levy, 2012). The task model consists of the situations and scenarios used to elicit the behaviours that can reveal the skills under observation. It usually relates the unobservable skills with the observable missions in games (Shute & Spector, 2008). The evidence model is responsible for updating the competency model on the basis of evidence gathered from the task model and is the bridge between the two models (Conrad, Clarke-Midura, & Klopfer, 2014).

16.5.1 Stealth Assessment Techniques

There are various ways a stealth assessment can be incorporated in a video game. Some of them are mouse-tracking (Rheem, Verma, & Becker, 2017), emotion tracking, log analysis, Bayesian modelling, along with several other Educational Data Mining techniques (Baker et al., 2012). The strength of all these techniques is that they provide rich information without the use of any expensive intrusive equipment, such as eye-tracker, galvanic skin response sensor, EEG, and other biometric instruments.

16.5.1.1 Mouse-Tracking

Educational video games that involve the use of a computer mouse or a touchscreen device can use mouse or touch-tracking as a stealth measure to assess situational specific variables, such as cognitive load (Rheem et al., 2017). Figure 16.3 shows a sample mouse-tracking plot depicting the trajectories for mouse-movement from the start location to the target. The process involves tracking the mouse-coordinates with time, and it is used to make inferences about the state or intent of the player. Mouse-tracking has been used in the past for inferring positive and negative emotions (Yamauchi & Xiao, 2018), memory strength (Papesh & Goldinger, 2012), gender stereotypes (Freeman & Ambady, 2009), numerical representation (Faulkenberry, 2016), perceptual decision making (Lepora & Pezzulo, 2015), and cognitive load (Rheem, Verma, & Becker, 2018). The inferences can then be used to alter the game-play to suit the player. For example, if it is observed that the player is experiencing a high cognitive load, then relevant steps should be taken to reduce the extraneous load by adapting the game in a suitable manner. While mouse-tracking is beneficial, collecting mouse-tracking data is a resource-intensive process and may demand extensive computer memory depending on the required temporal

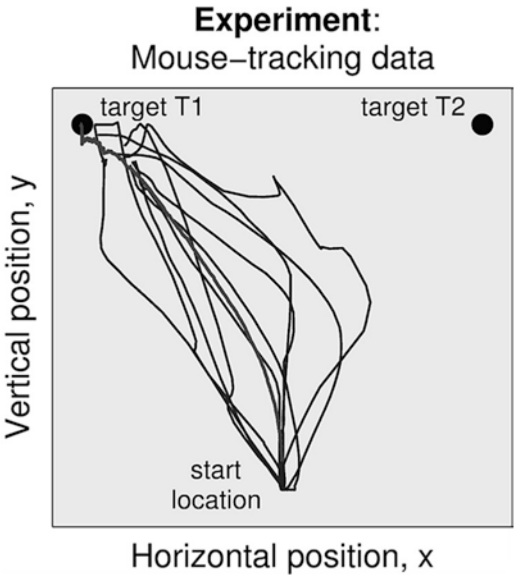


Fig. 16.3 Sample plot showing mouse-trajectory data (Lepora & Pezzulo, 2015)

resolution. For example, tracking mouse coordinates every 200 ms is less expensive compared to collecting it every 50 ms.

16.5.1.2 Emotion-Tracking

Emotion tracking involves tracking the mood of the player during the game-play so that it can be used to adjust the game for an optimal experience. A person might get bored if the game difficulty is too low, or they may get frustrated if it is too high. Thus, the game difficulty should be kept at such a level that keeps them in a state of flow (Csikszentmihalyi, 1975). The process requires facial tracking to detect the mood of the player. There are various methods available for the affect detection using facial tracking that use the Facial Action Coding System (Ekman & Friesen, 1978). Visage|SDK (Visage) from Visage Technologies and Affdex (Affectiva) from Affectiva are two software development kits which can be embedded in a video game for affect detection. Figure 16.4 shows the facial tracking snapshot, with the action units highlighted using white dots.

16.5.1.3 Log Analysis

Player data such as the number of lives remaining, number of player deaths, player level, time spent on a level and during a task, hint usage, quiz responses, score, and anything else that can be assigned to an observable variable can be collected and stored in a log file. A sample log file shown in Fig. 16.5 can then be analysed later

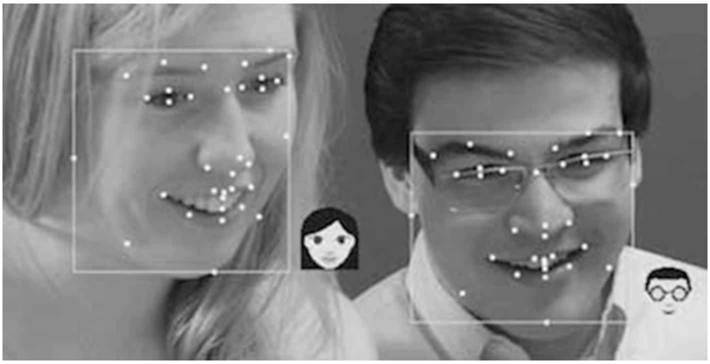


Fig. 16.4 Snapshot of emotion tracking using Affectiva (Metrics, 2019)

for a summative assessment or used for a runtime formative analysis. Wang, Shute, and Moore (2015) has incorporated the best practices to be used for a logging system. In short, they suggested to keep the log files customizable, manageable, well organized, usable, and include only the relevant data in it.

16.5.1.4 Bayesian Modelling

Bayesian modelling is a probabilistic approach to model the conditional dependence of a variable on several other variables (Friedman, Geiger, & Goldszmidt, 1997). García, Amandi, Schiaffino, and Campo (2007) used a Bayesian network to predict the learning styles of students in a web-based learning system. Figure 16.6 depicts a simple Bayesian network called knowledge tracing for a two-quiz sequence that incorporates the four performance parameters called prior knowledge P(L), guess rate P(G), slip rate P(S), and learn rate P(T) (Corbett & Anderson, 1994). Prior knowledge is obtained using diagnostic assessment and probabilistically influences all the other parameters. Guess rate is to account for the correct answers despite not having the knowledge required to do so, while slip rate is for the incorrect response by a skilled student. Learn rate is the probability that the learning will occur in the second quiz based on the response of the previous quiz. Bayesian networks can be used to model complex student models and will be discussed in more detail in the following sections.

16.5.1.5 Educational Data Mining

Educational Data Mining (EDM) consists of methods which are used to discover patterns in high volumes of educational data gathered during the student game-play interactions (Scheuer & McLaren, 2012). As a non-stealth measure, EDM has been used by D'Mello and Graesser (2010) to predict the affective states of students

```
3.483282 timeSpentInMenu 3.483282
4.441414 FoundTheFace
13.10056 CorrectAnswersDuringPreTest [1,2,5]
13.10056 timeSpentInPreTest 8.844877
219.8173 timeSpentInReading 206.7167
220.789 FoundTheFace
229.12 StateDetected Boredom
269.58 CollectedEverything
277.4767 timeSinceLastDeath 87.34433
296.36 P(skill(t=0)|evidence 0.321
309.1737 selfReportedFrustration
400.3962 timeOnDiffLevel1 235.5336
400.3962 SettingTheDifficultyLevel Two
422.8518 cumulativeTimeSpentOnLevels 260.8243
429.7115 timeOnUESSurvey 5.928741
434.5058 timeOnDemoSurvey 4.794312
447.4459 CorrectAnswersDuringPostTest [1,4]
447.4459 timeSpentOnPostTest 12.94009
447.4459 totalTimeSinceStart 508.0759
447.4459 totalHelpCount 0
447.4459 totalCoinsPickedUp 69
447.4459 totalLivesPickedUp 1
447.4459 totalDistractorsPickedUp 3
447.4459 enemyCollisionCount 11
447.4459 hazardCollisionCount 0
447.4459 timesDied 4
447.4459 skillLevel 0.9858083
```

Fig. 16.5 A sample log file

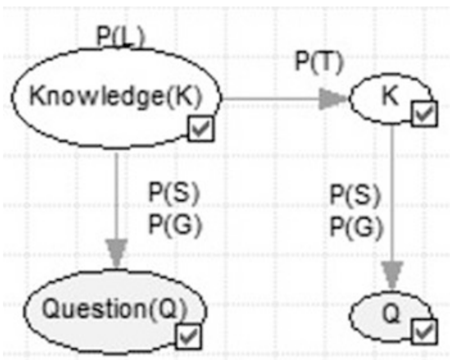


Fig. 16.6 Bayesian knowledge tracing (Pardos & Heffernan, 2010)

while they were sitting on a chair. They investigated the affective states and posture patterns of 28 students while they were learning with the help of an interactive tutoring system. Application of binary logistic regression associated the leaning back on a seat with boredom and disengagement and leaning forwards to frustration or delight depending on the angle of inclination while leaning forward. As a stealth measure, EDM was used by Baker and colleagues (2012) to predict the affective states of players using interaction logs and obtained a better than chance performance. EDM has also been used in the past to measure the degree of agency with which a student exerts control over their choice patterns (Snow, Jacovina, Varner, Dai, & McNamara, 2014). There is a wide array of EDM methods available such as, clustering, classification, regression, support vector machine, and reinforcement learning. Hence a great deal of care should be taken to pick the right one. Further, all the assumptions (if any) should be kept in mind while using that method.

16.5.2 Student Model

There are various aspects of a student that may need modelling while they are interacting with an educational video game. It can comprise trait variables, state variables, situation-specific variables, or any combination of them. The student model is a representation of the corresponding student assessment variable(s) at any point in time during the assessment. The student model can be potentially used to personalize the student learning to keep them in the zone of proximal development and provide necessary remediation if required.

Figure 16.7 above shows an example of a student model for an educational video game which uses the Dynamic Bayesian Network of knowledge tracing adapted from Pardos and Heffernan (2010). It is similar to the network in Fig. 16.6, except it is more complex and dynamic. The network shown in Fig. 16.6 consists of two nodes: a student node (S), a knowledge node (K), and a question node (Q). The prior knowledge parameter P(L) depicts the initial skill level of a student. The knowledge node corresponds to the state of the student knowledge, i.e. whether the skill has been attained or not. While the question node depicts whether they answered the quiz correctly or incorrectly. Figure 16.6 contains more nodes such as student node (S) and Distractor nodes (D). The student node represents an individual student. The arc below the Knowledge node depicts the conditional dependence of Knowledge at time step $t + 1$ on the Knowledge at previous time step t. This is shown clearly in the unrolled Dynamic Bayesian Network in Fig. 16.8.

Consider a game which is designed to teach encryption methods to its players using the basic Caeser cipher. The aim of any level in the game is to encrypt a plain text using a given key. To achieve this, the player is tasked with collecting the letters which appear in the resultant cipher-text when plain text is encrypted using the given key. For example, in Fig. 16.9, the resultant cipher-text for the given plain text “ATTACK AT DAWN” using the encryption key 2 will be “CVVCEM CV FCYP”. So the task of the player is to collect the letters ‘C’,‘V’,‘V’,‘C’,‘E’,‘M’,‘C

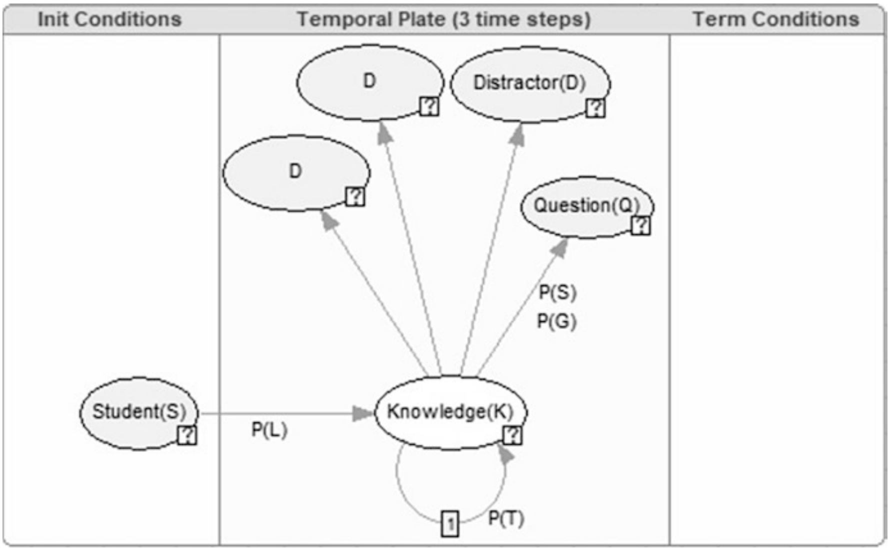


Fig. 16.7 An example of a dynamic Bayesian network

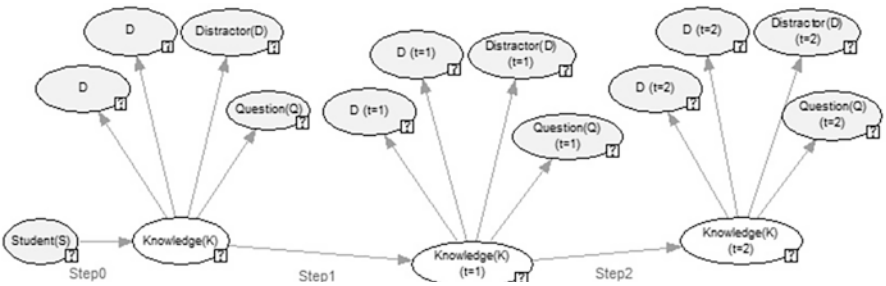


Fig. 16.8 Unrolled version of dynamic Bayesian network from Fig. 16.7

','V','F','C','Y','P' which are scattered throughout the level. The student node, in this case, would represent an individual player and their initial knowledge about encoding text using the Caeser cipher. Knowledge node would correspond to the state of their encoding skill, and question node would represent whether they achieved the task successfully or not. In addition to these three, Figs. 16.7 and 16.8 have several other nodes called distractor (D) which represents various distractors laid out around the level to check student skill and potential guessing. In the example game shown above, a distractor could be a letter which does not appear in the resultant cipher-text and therefore not supposed to be collected. Figure 16.4 displays a distractor letter 'H' which does not appear in the resultant cipher-text "CVVCEM CV FCYP". Collecting these distractors while not having the required skill could suggest guessing. All the performance parameters which represent the conditional probabilities at various nodes can be used for Bayesian inference while

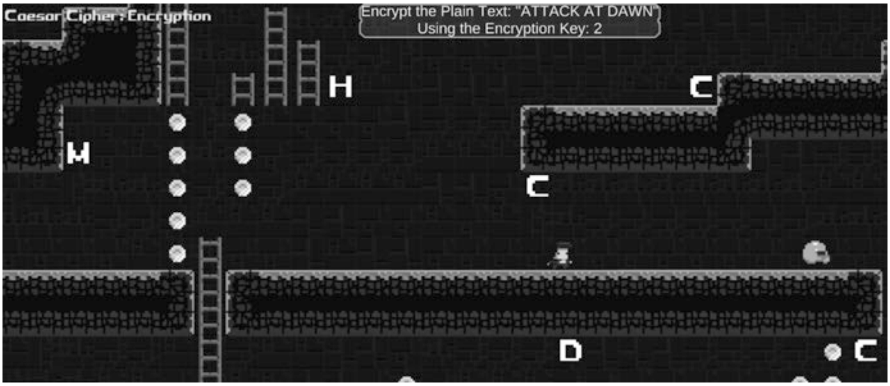


Fig. 16.9 An example of a game for student modelling

the game-play is in progress. The inference can be used to gauge the current skill level of student given various pieces of evidence. This, in turn, can serve as a formative assessment of the student skill and can be used for personalizing the learning of an individual by taking appropriate measures in accordance with the student model.

16.6 Content Agnostic Game Engineering

Educational video games have been shown to be effective for learning, but the learning gains are not generalizable (Cheng, Rosenheck, Lin, & Klopfer, 2017; Fletcher & Tobias, 2011; Freeman & Higgins, 2016). The results are often limited to the games used for research, and they are not content-agnostic. CAGE is an architecture for designing educational video games and assessment in which the game mechanics are independent of the game content while keeping the educational value of the game intact (Baron, 2017). It follows a game design approach and helps keep the players engaged to the game-play and learning. Being content-agnostic, it facilitates making the subsequent versions of the game and thus accelerating the development process. Only the first game will require the full-scale expenses; the following games will need some minor changes to accommodate the new content leading to reduced time and cost requirements.

CAGE has been proven to be effective in reducing the time spent while developing subsequent versions of the same game for different content (Baron, 2017). Baron (2017) did a study based on 11 students from a game-based learning class in Arizona State University. Participants were asked to make two games using the CAGE framework. On an average, they reported writing 70% lesser code and spending about 55% less time in developing the second game when using the CAGE framework. The results also indicated that the participants perceived the CAGE framework to be helpful in speeding up the game development process. However, it led to a decrease in cognitive load and engagement for players, when playing the second content right after the first one. For the first version of the game, the mechanics are new to the player and need to be learned. However, for the second version, the mechanics are the same and thus not required to be learned, hence the expected decrease in cognitive load and engagement.

The CAGE model depicted in Fig. 16.10 essentially consists of a one-way loop which begins with the player input to the game (Baron, 2017). The input is passed from the system hardware to the mechanics component which converts them into in-game action. The actions are then analysed by the content component, evaluating the action and passing the evaluation to student model which then accumulates the evaluation and passes the feedback to the player. Player then incorporates the feedback in their subsequent action.

CAGE architecture is component based and consists of the mechanics component, the content component, the student model, and the framework which binds them all together.

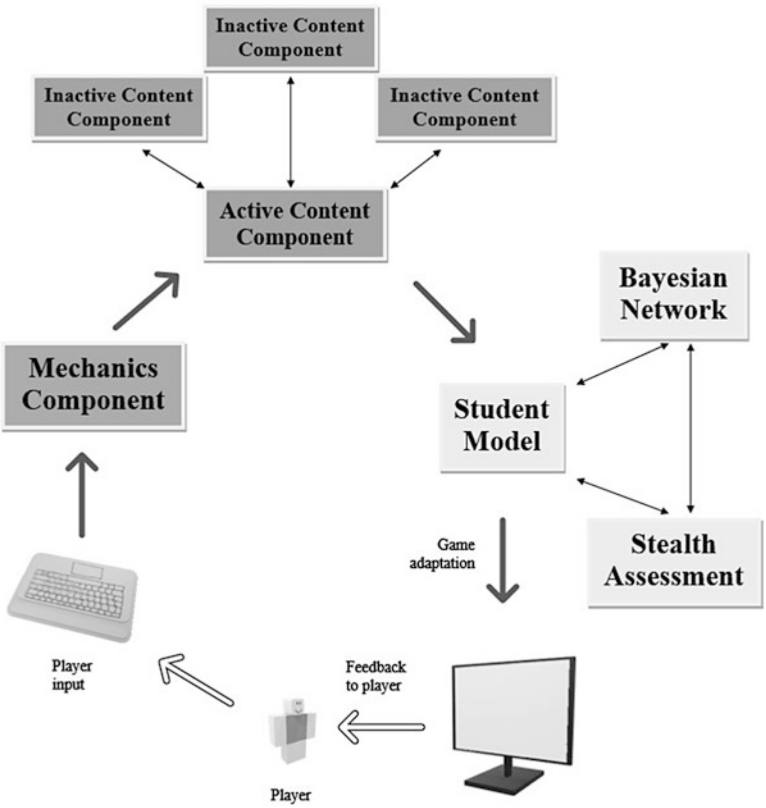


Fig. 16.10 The CAGE model (Baron, 2017)

16.7 Cage Architecture

The architecture framework is built in Unity game engine built by Unity Technologies (Baron, 2017). The framework utilizes generic messages called Hooks which are activated during the game-play events invoked by the player input. These Hooks are passed to the content component and processed if they are relevant to the content domain being played; otherwise they are ignored. This allows the mechanics component to send out the hooks to content component without knowing which content is active at present. The content component selectively implements the relevant Hooks. If an unknown Hook is received by the content component, it is ignored, and the player action is marked as invalid by the content component for that Hook.

16.7.1 Framework

The Framework is the skeleton that keeps all the components tied together (Baron, 2017). It connects the external input of the player to the game mechanics. The evaluation of the input is passed to the content component, and then to student model, which returns the feedback to the player via the framework part of the architecture. The player then incorporates the feedback into their next action, and the cycle is repeated. The Framework is static and consistent across all the version of the game developed using the architecture.

16.7.2 Mechanics Component

This component processes the input received from the player and converts it into in-game action. In CAGE architecture, this component is designed to be content-agnostic (Baron, 2017). Usually, game mechanics and content domain are either deeply connected as in traditional games, or poorly connected when using COTS games (Van Eck, 2006). However, in CAGE architecture they will be independent of each other and thus facilitate the mechanics to be content-agnostic.

16.7.3 Content Component

In CAGE this component is designed to be dynamic and easily replaceable with another content, being independent of the game mechanics (Baron, 2017). It evaluates the player action for their knowledge and skill level in that domain and passes the evaluation results to the student model to update the state of the student model.

16.7.4 Student Model

Student model represents the knowledge state of a given player at any point in time. It processes and accumulates the results from the content component. It is also used to dynamically provide appropriate feedback and remediation to the players, to aid their learning process. The student model has three-fold benefits associated with it. Firstly, it provides a dynamic assessment of the student knowledge state. Secondly, it can be used for dynamic feedback, remediation, and as a deterrent to behaviours that are not favourable to learning. Thirdly, it provides dynamic game adaptation capabilities to adjust the game or content difficulty on the basis of the skill level of the players and thus keep them in the zone of proximal development.

16.8 Conclusions

The growing volume of literature on game-based assessment suggest a bright future ahead. Games are intrinsically motivating and have the potential to promote sustained learning during the game-play session. The learning can be scaffolded into the gaming environment such that the mastery of learning is attained during the process of mastering the game environment. As opposed to traditional forms of assessment which allows measurement of state variables only, game-based assessment enables quantification of trait variables, state variables as well as situation specific-variables. It enables measuring skills such as persistence and systems thinking that are hard to measure using pen-and-paper tests while keeping the test anxiety at bay. It can be used for all sorts of assessment, diagnostic, formative, as well as summative. There is a wide range of assessment techniques available at our disposal. Emerging practices for game-based assessment involve tackling multiple content assessments using a single game without making the assessment obvious to the learner while building and adapting the learning strategy as the learner progresses through the game. Dynamically personalizing the game in accordance with the skill level of a player not only helps in keeping the player in flow but also helps in improving their learning.

References

Baron, T. (2017). *An Architecture for Designing Content Agnostic Game Mechanics for Educational Burst Games* (Doctoral dissertation, Arizona State University).

Baron, T., & Amresh, A. (2015). Word towers: Assessing domain knowledge with non-traditional genres. In *European Conference on Games Based Learning* (p. 638). Academic Conferences International Limited.

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*, 1.

Birk, M. V., Mandryk, R. L., & Atkins, C. (2016, October). The motivational push of games: The interplay of intrinsic motivation and external rewards in games for training. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play* (pp. 291–303). New York, NY: ACM.

Brawl Stars. (2017). *Supercell*.

Chatzisarantis, N. L., Biddle, S. J., & Meek, G. A. (1997). A self-determination theory approach to the study of intentions and the intention–behaviour relationship in children's physical activity. *British Journal of Health Psychology, 2*(4), 343–360.

Cheng, M.-T., Rosenheck, L., Lin, C.-Y., & Klopfer, E. (2017). Analyzing gameplay data to inform feedback loops in the radix endeavor. *Computers & Education, 111*, 60–73.

Chin, J., Dukes, R., & Gamson, W. (2009). Assessment in simulation and gaming: A review of the last 40 years. *Simulation & Gaming, 40*(4), 553–568.

Conrad, S., Clarke-Midura, J., & Klopfer, E. (2014). A framework for structuring learning assessment in a massively multiplayer online educational game: Experiment centered design. *International Journal of Game-Based Learning (IJGBL), 4*(1), 37–59.

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*(4), 253–278.

Csikszentmihalyi, M. (1975). Play and intrinsic rewards. *Journal of Humanistic Psychology, 75*, 41.

Baker, R. S., Gowda, S. M., Wixon, M., Kalka, J., Wagner, A. Z., Salvi, A., ... & Rossi, L. (2012). Towards sensor-free affect detection in cognitive tutor algebra. *International Educational Data Mining Society*.

D'Mello, S., & Graesser, A. (2010, June). Mining bodily patterns of affective experience during learning. In *Educational data mining 2010*.

Deci, E. L., & Ryan, R. M. (2008). Self-determination theory: A macrotheory of human motivation, development, and health. *Canadian Psychology/Psychologie Canadienne, 49*(3), 182.

Deci, E. L., & Vansteenkiste, M. (2004). Self-determination theory and basic need satisfaction: Understanding human development in positive psychology. *Ricerche di psicologia*.

Dragon Age: Origins. (2009). *BioWare*.

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system: Investigator's guide*. Palo Alto, CA: Consulting Psychologists Press.

Faulkenberry, T. J. (2016). Testing a direct mapping versus competition account of response dynamics in number comparison. *Journal of Cognitive Psychology, 28*(7), 825–842.

Fletcher, J., & Tobias, S. (2011). *Computer games and instruction*. Charlotte, NC: Information Age Publishing.

Freeman, B., & Higgins, K. (2016). A randomised controlled trial of a digital learning game in the context of a design-based research project. *International Journal of Technology Enhanced Learning, 8*(3–4), 297–317.

Freeman, J. B., & Ambady, N. (2009). Motions of the hand expose the partial and parallel activation of stereotypes. *Psychological Science, 20*(10), 1183–1188.

Freire, M., Serrano-Laguna, Á., Iglesias, B. M., Martínez-Ortiz, I., Moreno-Ger, P., & Fernández-Manjón, B. (2016). Game learning analytics: Learning analytics for serious games. In *Learning, design, and technology* (pp. 1–29). Cham, Switzerland: Springer.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning, 29*(2–3), 131–163.

García, P., Amandi, A., Schiaffino, S., & Campo, M. (2007). Evaluating Bayesian networks' precision for detecting students' learning styles. *Computers & Education, 49*(3), 794–808.

Gee, J. P. (2003). What video games have to teach us about learning and literacy. *Computers in Entertainment (CIE), 1*(1), 20–20.

Järvinen, A. (2008). *Games without frontiers: Theories and methods for game studies and design*. Tampere, Finland: Tampere University Press.

Kibler, J. (2011). Cognitive disequilibrium. *Encyclopedia of child behavior and development*, 380–380.

League of Legends. (2009). *Riot Games*.
Lepora, N. F., & Pezzulo, G. (2015). Embodied choice: How action influences perceptual decision making. *PLoS Computational Biology, 11*(4), e1004110.
Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science, 5*(4), 333–369.
Metrics. (2019). Retrieved from https://developer.affectiva.com/metrics/
Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 2003*(1), i–29.
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.
Moreno-Ger, P., Martinez-Ortiz, I., Freire, M., Manero, B., & Fernandez-Manjon, B. (2014). Serious games: A journey from research to application. In *2014 IEEE Frontiers in Education Conference (FIE) Proceedings* (pp. 1–4). IEEE.
Papesh, M. H., & Goldinger, S. D. (2012). Memory in motion: Movement dynamics reveal memory strength. *Psychonomic Bulletin & Review, 19*(5), 906–913.
Pardos, Z. A., & Heffernan, N. T. (2010). Modeling individualization in a Bayesian networks implementation of knowledge tracing. In *International Conference on User Modeling, Adaptation, and Personalization* (pp. 255–266). Berlin, Germany: Springer.
Plass, J. L., Homer, B. D., Kinzer, C. K., Chang, Y. K., Frye, J., Kaczetow, W., … Perlin, K. (2013). Metrics in simulations and games for learning. In *Game analytics* (pp. 697–729). London, UK: Springer.
Renard, L. (2016a). *The ABC of a school without grades. An interview with Christel moors*. Retrieved from https://www.bookwidgets.com/blog/2016/11/the-abc-of-a-school-without-grades-an-interview-with-christel-moors
Renard, L. (2016b). *What is formative assessment? The 6 building blocks.* Retrieved from https://www.bookwidgets.com/blog/2016/11/what-is-formative-assessment-the-6-building-blocks
Rheem, H., Verma, V., & Becker, V. (2017). *Give and take in a moustracking choice paradigm*. Poster presented at Proceedings of the 58th Annual Meeting of Psychonomic Society, Vancouver, BC.
Rheem, H., Verma, V., & Becker, V. (2018). Use of mouse-tracking method to measure cognitive load. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 62, pp. 1982–1986). Los Angeles, CA: SAGE Publications.
Ritterfeld, U., Cody, M., & Vorderer, P. (2009). *Serious games: Mechanisms and effects*. New York, NY: Routledge.
Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25*(1), 54–67.
Ryan, R. M., LaGuardia, J. G., & Rawsthorne, L. J. (2005). Self-complexity and the authenticity of self-aspects: Effects on well being and resilience to stressful events. *North American Journal of Psychology, 7*(3), 431.
Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion, 30*(4), 344–360.
Scheuer, O., & McLaren, B. M. (2012). Educational data mining. In *Encyclopedia of the sciences of learning* (pp. 1075–1079). Berlin, Germany: Springer.
Second Life. (2003). *Linden Labs*.
Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction, 55*(2), 503–524.
Shute, V. J., & Spector, J. M. (2008). Scorm 2.0 white paper: Stealth assessment in virtual worlds. *Unpublished manuscript*. Tallahassee, FL: Florida State University
Shute, V. J., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. Cambridge, MA: MIT Press.
Sicart, M. (2008). Defining game mechanics. *Game Studies, 8*(2). Retrieved from: http://gamestudies.org/0802/articles/sicart.

Snow, E., Jacovina, M., Varner, L., Dai, J., & McNamara, D. (2014, July). Entropy: A stealth measure of agency in learning environments. In *Educational Data Mining 2014*.

Sørebø, Ø., & Hæhre, R. (2012). Investigating students' perceived discipline relevance subsequent to playing educational computer games: A personal interest and self-determination theory approach. *Scandinavian Journal of Educational Research, 56*(4), 345–362.

Team Ico. (2005). *Shadow of the Colossus* (PlayStation 2).

Van Eck, R. (2006). Digital game-based learning: It's not just the digital natives who are restless. *Educause Review, 41*(2), 16.

Vygotsky, L. (1978). Interaction between learning and development. *Readings on the Development of Children, 23*(3), 34–41.

Wang, L., Shute, V. J., & Moore, G. (2015). *Best practices and lessons learned of stealth assessment*. Retrieved December, 9, 2015.

West, D. M., & Bleiberg, J. (2013). Education technology success stories. *Issues in Governance Studies.*

Yamauchi, T., & Xiao, K. (2018). Reading emotion from mouse cursor motions: Affective computing approach. *Cognitive Science, 42*(3), 771–819.

Index

D. Ifenthaler, Y. J. Kim (eds.), *Game-Based Assessment Revisited*, Advances in Game-Based Learning, https://doi.org/10.1007/978-3-030-15569-8

H

I