Chapter 2
Measurement Challenges of Interactive Educational Assessment

David C. Gibson, Mary E. Webb, and Dirk Ifenthaler

Abstract This chapter discusses four measurement challenges of data science in educational assessments that are enabled by technology: (1) Dealing with change over time. (2) How a digital performance space's relationships interact with learner actions, communications and products. (3) How layers of interpretation are formed from translations of atomistic data into meaningful larger units suitable for making inferences about what someone knows and can do. (4) How to represent the dynamics of interactions between and among learners who are being assessed by their interactions with each other as well as with digital resources and agents in digital performance spaces. Because of the movement from paper-based tests to online learning, and in order to make progress on these challenges, the authors advocate the restructuring of training of the next generation of researchers and psychometricians in technology-enabled assessments.

D. C. Gibson
Curtin University, Perth, WA, Australia
e-mail: david.c.gibson@curtin.edu.au; https://staffportal.curtin.edu.au/staff/profile/view/David.C.Gibson

M. E. Webb
School of Education, Communication and Society, King's College London, London, UK
e-mail: mary.webb@kcl.ac.uk; https://www.kcl.ac.uk/sspp/departments/education/people/academic/webbm.aspx

D. Ifenthaler (✉)
Curtin University, Perth, WA, Australia

Economic and Business Education, Learning, Design and Technology, University of Mannheim, Mannheim, Germany
e-mail: dirk@ifenthaler.info; http://www.ifenthaler.info

D. Sampson et al. (eds.), *Learning Technologies for Transforming Large-Scale Teaching, Learning, and Assessment*, https://doi.org/10.1007/978-3-030-15130-0_2

1 Introduction

Assessment and learning analytics challenges have dramatically increased since new digital performance affordances, interactive user interfaces and the targets of technology-enabled assessments have become more complex. The increased complexity is due in part to technology's capabilities and roles in presenting interactive learning experiences and collecting rich data (de Freitas, 2014; Quellmalz et al., 2012) which is leading to the infusion of data science methods and techniques into learning and behavioural science research (Gibson & Knezek, 2011; Kozleski, Gibson, & Hynds, 2012). These changes require new quantitative methods as well as a reconceptualization of mixed methods (Tashakkori & Teddlie, 2003) that include domain experts as well as stakeholders in the construction of knowledge of such complex systems (Gibson & Ifenthaler, 2017).

In technology-enhanced assessments, the emergence of 'big data'—which at a minimum are defined as data with a large numbers of records, of widely differing data types, that are rapidly collected for immediate action (IBM, 2015; Margetts & Sutcliffe, 2013)—underscores the need to develop assessment literacy (Stiggins, 1995) in teachers, learners and other audiences of assessment. Assessment literacy has become more important than ever for understanding how technology influences and impacts assessment types and processes and especially for developing confidence in creating and analysing arguments from evidence, based on a user's current understanding of validation (Black, Harrison, Hodgen, Marshall, & Serret, 2010).

This chapter discusses four key challenges associated with applying data science methods to address aspects of an interactive digital media assessment's psychometric properties; time sensitivity; digital performance and the problem space for analysis; the hierarchy of tasks, turns and translations between different levels and the dynamics of interrelationships in assessment systems. First, in traditional assessments such as quizzes, tests and on-demand performances, change over time (e.g. whether a learner has newly acquired knowledge or skill, or has learned) is a matter of comparing a series of summative results. But in interactive digital media learning, acquisition of knowledge or skill may be evident during the learner's dynamic engagement with the media (Quellmalz et al., 2012). Second, the problem space in a traditional assessment is designed around the concept of a valid construct (Cronbach & Meehl, 1955); whereas in interactive digital media, it is perhaps more relevant to speak of an authentic performance (Wiggins, 1989). Third, in traditional assessments, a relatively static model pre-exists as the backdrop for the relationship of the test taker's task to the construct being measured (Fischer & Molenaar, 2012); while in interactive assessments, unexpected performances, resource utilizations and therefore interpretation models can dynamically emerge. Finally, in traditional assessments, the relationship of a learner's performance to the construct (including any error) is treated as invariant, but in interactive assessments, the challenges of longitudinal data analysis are evident (Hedeker & Gibbons, 2006).

In this chapter, the OECD (Organisation for Economic Co-operation and Development) PISA (Programme for International Student Assessment) plan for assessing complex problem-solving (CPS) is used as an example to explain these challenges in relation to a complex problem space. The chapter then illustrates a learning analytics case that shows how the identified challenges have been addressed in the development of assessments.

2 Background

There is uncertainty as to whether and how different perspectives on assessment—providing feedback information, supporting improvement decisions, identifying the degree of engagement and understanding, and making value judgments—can coexist to the benefit of learners (Webb, Gibson, & Forkosh-Baruch, 2013). Even with the increased possibilities that information technology provides there is not yet a way to say confidently that the multiple purposes for which some assessments have been used (Mansell, James, & The Assessment Reform Group, 2009) can or should be supported through the same assessment systems. This is because *the impacts of some purposes interact with the validation processes of others* (Messick, 1994). For example, the validity of an assessment for a learner may be related to its relevance to knowledge needed immediately to improve performance, but the purpose of the assessment might have been designed to provide information about program effectiveness to a school authority. The test taker, not seeing the relevance, might not perform as well as possible. Therefore, in considering assessment design for multiple purposes for example for formative as well as summative purposes, assessment designers need to examine potential impacts carefully in order to minimise negative consequences on learning and learners.

Developing theory for the application of data science methods in educational research is important for two primary reasons. First, assessment of virtual performance presents new challenges for psychometrics (Clarke-Midura & Dede, 2010; Ifenthaler, Eseryel, & Ge, 2012; Quellmalz et al., 2012). Secondly, new tools are needed for discovery of patterns and drivers in complex systems for working with 'big data' in educational research preparation and practice (Gibson, 2012; Ifenthaler, 2015; Patton, 2011). Indicators of progress in theory development would be an increase in research exploring and articulating the use of data science methods in learning analytics to improve learning and achievement; and the expansion of methods beyond traditional statistics and qualitative approaches in educational research, to include data mining, machine learning and, in general, the methods of data science.

3 Four Psychometric Challenges

Psychometrics is the branch of psychology that deals with the design, administration and interpretation of quantitative tests for the measurement of psychological variables such as intelligence, aptitude and personality traits ("Psychometrics", 2014). A good psychometric test is “internally consistent, reliable over time, discriminating and of demonstrated validity in respect of its correlations with other tests, its predictive power and the performance of various criterion groups. It also has good norms” (Kline, 1998, p. 92).

Until recently, the field dealt almost exclusively with the construction and validation of measurement instruments such as questionnaires, tests and personality assessments. However, there is now a need to expand to include highly interactive digital learning and adaptive test experiences, such as the OECD PISA assessment of CPS. In brief, PISA is a triennial international survey that aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students in order to determine the extent to which they can apply their knowledge to real-life situations and hence are prepared for full participation in society. To constrain the quite complex variables that would be involved if the collaboration was among a set of real people, the OECD PISA CPS assessment utilizes the computer to play roles as collaborators in a *virtual performance assessment* (Clarke-Midura, Code, Dede, Mayrath, & Zap, 2012; Zervas & Sampson, 2018). The PISA assessment plan incorporates a complex behaviour space that illustrates some of the new demands on psychometrics.

The challenge with technology-enabled assessments that produce big data is to evolve the procedural foundations of psychometrics, which until recently have been primarily based on population statistics and static snapshots of data. Elements of a new foundation outlined here highlight the need to include time sensitivity, digital performance space relationships, multiple layers of aggregations at different scales and representations of the dynamics of a complex behaviour space (Gibson & Jakl, 2013; Quellmalz et al., 2012).

3.1 Time Sensitivity

In the OECD PISA CPS assessment, time is controlled as a boundary variable of the test and the computer is used to prompt the test taker to ‘move on’ when the evidence rule system detects that the student needs to be rescued from an unproductive problem-solving path. The decision to redirect appears natural to the situation because the computer is playing the role of one or more collaborators, so the suggestion to move on comes from a simulated peer. This situation illustrates that a technology-enabled assessment might well give the student perceived or actual control over time, compared to an assessment that only displays test item prompts in a timed test. In some virtual performance assessments, time is open-ended, and the use of item resources (e.g. in what order, with or without returning to the resources multiple times, time spent with each resource, timing of the appropriate use of a resource and total time to utilize the appropriate resources to accomplish the task) may be critical to the classification of the learner’s response (Gibson & Jakl, 2013; Stevens & Palacio-Cayetano, 2003).

Table 2.1 Domain model for assessing collaborative problem-solving

	(1) Establishing and maintaining shared understanding	(2) Taking appropriate action to solve the problem	(3) Establishing and maintaining team organisation
(A) Exploring and understanding	(A1) Discovering perspectives and abilities of team members	(A2) Discovering the type of collaborative interaction to solve the problem, along with goals	(A3) Understanding roles to solve problem
(B) Representing and formulating	(B1) Building a shared representation and negotiating the meaning of the problem (common ground)	(B2) Identifying and describing tasks to be completed	(B3) Describe roles and team organisation (communication protocol/ rules of engagement)
(C) Planning and executing	(C1) Communicating with team members about the actions to be/being performed	(C2) Enacting plans	(C3) Following rules of engagement (e.g., prompting other team members to perform their tasks)
(D) Monitoring and reflecting	(D1) Monitoring and repairing the shared understanding	(D2) Monitoring results of actions and evaluating success in solving the problem	(D3) Monitoring, providing feedback and adapting the team organisation and roles

The OECD PISA CPS assessment solves the time sensitivity problem by parsing time into critical events and then monitoring the event patterns to detect the level of evidence of the competencies in the domain model (see Table 2.1). This is a form of *time segmentation*, because some events cannot happen until other events have occurred (e.g. establishing and maintaining team organisation must occur after establishing a shared vision, and while maintaining that vision and taking appropriate action to solve the problem). A planned sequence of activities and timed release of testing resources, known in game-based learning as a 'branching storyline' (Aldrich, 2005) is a method for controlling the evolution of a process.

Other problem-solving contexts, such as coordination of group actions needed for scientific inquiry and experimentation, require simultaneous actions mixed with sequences of actions. The classification system of the assessment has to handle *patterns of simultaneous and sequential interactions* in order to make valid links to time-sensitive evidence rules within the conceptual assessment framework (CAF), which is a key component of evidence-centred design (Mislevy, Steinberg, & Almond, 1999), an approach that is becoming increasingly prominent in assessment design and on which this analysis is based. The CAF has three core components: the student model, task model and evidence model (Mislevy et al., 1999; Mislevy, Steinberg, & Almond, 2003) within and among which the time-sensitive relationships adhere.

3.2 Digital Performance Space Relationships

A learning experience entails a designed structure of knowledge and action (Jonassen, 1997) and when that experience is interactive and digital there are many measurement challenges (Quellmalz et al., 2012). The emerging varieties of network analysis (e.g. social networks, visualization, artificial neural networks, decision trees) have arisen as new analytical tools and methods for understanding the structural relationships in technology-enhanced learning (Choi, Rupp, Gushta, & Sweet, 2010; Shaffer et al., 2009). In addition, the traces of knowledge and action (i.e., the actions, communications and products) created by a learner during the course of interacting with a digital learning application bear a relationship to that person's mental representations of the problem (Newell & Simon, 1972) and the knowledge and capability they acquired, accessed and utilized during the interaction (Ifenthaler, 2014; Pirnay-Dummer, Ifenthaler, & Spector, 2010; Thagard, 2010). This set of ideas are referred to here as '*digital performance space relationships*' which will be shown to be similar to 'items' and 'constructs' in classical test theory.

An interactive digital performance space can support several scenarios, each with one or more classification challenges for inferring what the test taker knows and can do. In the OECD PISA CPS assessment, for example, the scenarios presented to the student are designed to sample the digital performance space construct of 'collaborative problem-solving.' Each scenario allows the classification of the test taker into one or more cells of a matrix created by the intersection of three stages of 'collaboration' with four stages of 'problem-solving' (see Table 2.1). In classical test theory, the 'construct' plays a similar role to the digital performance space; several test items are used to make multiple measures of the construct. A review of the historical idea of a valid construct is helpful for making the bridge from classical testing to the digital age.

A valid construct was thought of as an *inductive summary* and as *part of a series* of validity investigations that included concurrent, predictive and content considerations. In addition, the *construct can change* and *become more elaborated over time*, as Cronbach noted (Cronbach & Meehl, 1955): When a construct is fairly new, there may be few specifiable associations by which to pin down the concept. As research proceeds, the construct sends out roots in many directions, which attach it to more and more facts or other constructs. Finally, the construct acquired validity through the idea of a *nomological network* which is a collection of overlapping *mappings* from (a) observable properties or quantities to one another; (b) different theoretical ideas to one another, or (c) theoretical constructs to observables (ibid). A single mapping might include examples of all these relations, as a construct might be a complex set of factors that interact with one another. The idea of a network of ideas and relationships was a fairly abstract philosophical idea in the 1950s but today has a renewed and concrete meaning that has become known as network theory in social science (Borgatti & Halgin, 2011) and network analysis in computational sciences, both of which are applied graph theory from mathematics (Brandes & Erlebach, 2005). This history outlines a bridge of ideas that carries forward into today when digital media learning spaces can record a network of traces of the actions of a learner.

Digital media learning presents problems as well as prompts for learner performance (e.g. problem-solving, collaboration) in a space that is characterized by hyperlinked resources that can be represented as nodes and relations in a network (Clarke-Midura et al., 2012; Quellmalz et al., 2012; Stevens, 2006). As a learner uses such a space to learn and perform (e.g. interacting with the resources to solve a problem, adding new information, re-arranging resources into new relationships) a new network can be created that represents the learner's response, a time-specific performance path through the digital performance space (Ifenthaler et al., 2012). The learner's performance network is a constructed knowledge structure that needs to be taken into account in assessment (Gijbels, Dochy, Van den Bossche, & Segers, 2005). The digital performance space and the constructed knowledge structure of the learner hold the same kind of relationship as the nomological network does to a demonstrated construct; the digital performance space holds the learning designer's view of the construct (e.g. what it means to act like a scientist in a given situation) and the constructed knowledge structure (e.g. what the learner did in this instance) holds evidence of the processes and products of knowing and doing.

The terms of the nomological network inference, which underpins a claim of construct validity, bear a similarity to the rules of a chain of a reasoned argument, which can lead to a claim concerning what a learner knows and can do as used in Evidence-Centered Design (ECD). In ECD, an argument has constituent claims, data, warrants and backing and must take account of alternative explanations. In a nomological network by comparison, there are observations, ideas and relationships, and a chain of inference must be used in order to establish a claim that a particular test is a measure of the construct.

The relationships and nodes of a network representation of the traces of learner interactions can be compared to the digital performance space resources and relationships to enable inferences about what the learner knows and can do (Al-Diban & Ifenthaler, 2011; Ifenthaler, 2010; Quellmalz, Timms, & Schneider, 2009). Network measures such as similarity, centrality, clusters and pattern matching are used in such inferences, where the patterns of the network imply functional and structural connectivity (Sporns, 2011). Digital performance space relationships examined with time-sensitive network analysis has increased the ability of research to characterise and make comment on processes, products, knowledge and know-how, and their complex entanglements in authentic performance settings.

3.3 Layers of Aggregations and Translations

In the OECD PISA CPS assessment, aggregations of events into tasks takes place in a hierarchy that begins at the top with a scenario and ends within each task of the scenario at the level of a 'turn'—a game-based learning concept that updates the state of the scenario based on the learner's input.

Each problem scenario (unit) contains multiple tasks. A task, for example, consensus building, is a particular phase within the scenario, with a beginning and an end. A task consists of a number of turns (exchanges, chats, actions, etc.) between the participants in the team. A finite number of options leading onto different paths are available to the participants after each turn, some of which constitute a step towards solving the problem. The end of a task forms an appropriate point to start the next task. Whenever the participants fail to reach this point a 'rescue' is programmed to ensure that the next task can be started (PISA, 2013).

With this hierarchy in mind (e.g. scenarios containing tasks that contain turns) the challenge of aggregating with time sensitivity and translating from one level of analysis to another can be addressed with moving averages, sliding time windows and event recognition. The OECD PISA CPS assessment uses event recognition, in which an action, communication or product of the test taker triggers a reaction by the test engine to update the scenario, which might include rescuing the test taker. In a moving average, some window of time is selected (e.g. every second, or after every three turns) and an average is performed to form an abstracted data layer that preserves some of the shape of the data movement over time. In the sliding time window (Choi et al., 2010; Han, Cheng, Xin, & Yan, 2007), a combination of event recognitions and moving averages, or some configuration of either, might be performed and then used as an abstracted data layer. In the example case summarized below, for example, the time stamps of every action were subtracted from each other to compute duration, which was then applied to each action, to nearby action-pairs and to action-n-grams (motifs) for further analysis.

Within any slice of time, or when comparing two or a few slices of time, standard statistical procedures and aggregations apply (e.g., means testing, correlations, regressions), but when high-resolution data is involved (e.g. many data points per record per unit of time) and where there are complex aggregations (e.g., widely varying sources of data and different units of measure) then data mining techniques are more applicable. Of note, regression techniques in data mining are not equivalent to the same methods in statistics, even though the terms sound and look the same. In data mining, regression represents a search within a complex nonlinear space for patterns and representations of structure and causal dynamic relationships, rather than the reduction of error of a linear model (Schmidt & Lipson, 2009). Thus, aggregations in the two approaches are also of different lineage and need to be considered as separate entities with separate representational functions, meaning and purposes (Bates & Watts, 1988).

3.4 Representations of Dynamics

Systems dynamics (Bar-Yam, 1997; Sterman, 1994) involves a mathematical modeling technique for framing, understanding and discussing the issues of time, digital performance space relationships and aggregation-translation in highly interactive technology-enhanced assessments. Field experiments with systems dynamics methods have, for example, focused on mid-level model-based theory building in an assessment context (Kopainsky, Pirnay-Dummer, & Alessi, 2010). The process of building a model from snapshots of a dynamic system is called a 'nonlinear state space reconstruction' (Sugihara et al., 2012). In such a state space equivalent to a network all the data fall within a finite band or manifold of behaviour. That is, every state of the system will be in one of the spaces created by the finite possibilities for each variable at some point in time. Such reconstructions of the underlying manifold governing the dynamics of a system can map to and uncover the causal relationships in a complex system (Schmidt & Lipson, 2009) including those that support inferences concerning what a user knows and can do.

Visualizing the current status of a learner's progress on an assessment is an example of representing a state of a dynamic system, as is visualizing the progress of the learner in relation to a domain model driving the assessment's evidence collection processes. The Khan Academy (Khan, 2011), for example, charts progress in learning mathematics or science content against a visualization of the content hierarchy. If the learner has mastered division, a visual tree shows how mastery fits with addition and subtraction and allows access to the next higher level of math skill. More dynamic and fine-grained visualizations are also possible, for example, that would trace the process steps of a solution, or document the details of a constructive process. Visualizations can aide pattern discovery involving both nonverbal and verbal expressions; for example, from bodies of text, from online student discussion forums, and from cognitive and mental model representations (Pirnay-Dummer et al., 2010).

To date the developments in learning analytics that provide visualisations of learning traces for learners and teachers have been represented by learning analytics dashboards. Such dashboards have been developed that keep track of time, social interactions for insights into collaboration, the use of documents and tools and the artefacts produced by students (Verbert, Duval, Klerkx, Govaerts, & Santos, 2013). While these dashboards currently fall far short of the detailed traces of assessment data that are possible to create, even these more limited opportunities for analysing their learning have been found to support learners' reflection and improve self-assessment as well as increasing course satisfaction (Verbert et al., 2013).

Examples of the more highly detailed traces are readily found in serious games, as well as casual games that are designed to be immersive and emotionally engaging rather than a simple pastime (Aldrich, 2005). In these game-based examples, the high-resolution feedback is always on, giving the player an up-to-date view of progress, hints about upcoming challenges, and a view to the longer-term goal (Prensky, 2001). Clearly educators and researchers might want to promote to policymakers the importance of researching the methods and impacts of presenting visualisations of data to teachers and learners along with developments in data processing that will better enable judgements of student performances.

Perhaps the biggest unresolved issue of representation of collaborative learning (and perhaps any learning progress during a complex process) is how to represent the moving and evolving quality of change over time. 'Movies' of dynamic educational processes have not yet been documented in many cases, and if existing, have not been widely disseminated into common practice. This lack of a practice base and experience hampers theory as well as practice in technology-enhanced assessments, and points to the need illustrated by the case in the next section, for future research and practice to create a shared understanding of the methods of data science in educational research.

4 Case Story: Virtual Performance Assessment

A case story illustrates how technology-enabled educational assessment can produce a large number of records, how time and process can be an included mediating factor in analysis and how machine learning and data mining methods are needed to support the rapid simultaneous testing of multiple hypotheses.

A game-based assessment of scientific thinking was created at Harvard (Clarke-Midura et al., 2012) and analysed by one of the authors (Gibson & Clarke-Midura, 2013) to ascertain the abilities of middle school students to design a scientific investigation and construct a causal explanation. A summary of the data science findings and issues included the observation of two of the three aspects of big data: volume (~821,000 records for 4000 subjects, or 205 records per subject); and variety of data (user actions, decisions and artefacts provided evidence of learning and thought processes). The third element of big data, velocity, was less important in this case; because the flow of data was not used in near-real time to give hints, correct mistakes, or inform the learner during the experience, so the data was streamed off to storage for later analysis.

This case illustrates several of the features of big data in educational assessment. First, the context was captured along with the learner action, decision and product, but that context needed to be effectively constructed from the smallest items of data into larger clusters of information. For example, a data element named ‘opened door’ by itself was relatively meaningless compared to knowing that it was a particular door, opened after another significant event such as talking to a scientist. Thus, patterns of action were transformed into *n-grams* (Scheffel, Niemann, & Leony, 2012) or *motifs*, which then became the transformed units of analysis. This concept of the unit of analysis containing the semantic, time and space contexts for lower levels of aggregation may be a new methodological requirement of digital assessments, and needs further study.

Second, as a large number of users traverse through the network of possibilities in a digital performance space, key movements of the population within the network can be counted and then used as the basis for *empirical prior probabilities* which assist in creating Bayesian inferences about the scientific problem-solving path-maps of learners (Stevens, Johnson, & Soller, 2005). In particular, each pathway in such a network can be further characterized or specified with a predictive nonlinear mathematical relationship (Gibson & Clarke-Midura, 2013), for example, found through *symbolic regression* an evolutionary machine learning technique (Schmidt & Lipson, 2009). Or, alternatively an *association rule network* can be created that distinguishes user action patterns and motifs according to the prevalence of utilizing one resource compared to another. For example, if 100% of the population goes to resource 3 after resource 1 (skipping over and not utilising resource 2), then with a very high probability, if the sample is a good sample of the greater population, the next user entering the system will follow that path and the inference system can make a highly probable educated guess about what the person now using resource 1 will do next.

The third feature is that the complex set of relationships in various analyses such as those just mentioned, bear a structural relationship to something meaningful about the digital performance space as outlined above. For example, a *cluster analysis* can reveal that some resources are critical to success and others are ignored and not important to the most successful learners (Quellmalz et al., 2012) or a *network visualization* can highlight how people relate to each other or to a task such as quoting and using scientific resources (Bollen et al., 2009).

5 Conclusion and Implications

This chapter has introduced four challenges of big data in educational assessments that are enabled by technology: how to deal with change over time and time-based data; how a digital performance space's relationships interact with learner actions, communications and products; how layers of interpretation are formed from translations of atomistic data into meaningful larger units; and how to represent the dynamics of interactions between and among learners who are being assessed by their interactions in digital performance spaces. The chapter linked the big data challenges to solutions offered in the OECD PISA CPS assessment of collaborative problem-solving, and then reviewed some of the same issues by briefly summarizing a particular case.

The challenges and issues discussed in this chapter reveal the requirements for developments in theory as well as some of the practical challenges that will need to be overcome if educators are to achieve the vision of providing learners and teachers with a 'quiet assessment' system in which the impact can be turned up at the request of learners and teachers as they seek to understand the progress of learning. This joint approach which emphasises assessment AS, FOR and OF learning (Bennett, 2010) is discussed further in the following publications (Gibson & Webb, 2015; Webb & Gibson, 2015; Webb & Ifenthaler, 2018). In moving forward to embrace the opportunities that could be provided by technology-enhanced assessments the challenges that remain to be addressed must not be underestimated before educators can use automated assessments of complex skills and understanding with confidence.

References

Al-Diban, S., & Ifenthaler, D. (2011). Comparison of two analysis approaches for measuring externalized mental models. *Educational Technology & Society, 14*(2), 16–30.

Aldrich, C. (2005). *Learning by doing: The essential guide to simulations, computer games, and pedagogy in e-learning and other educational experiences*. San Francisco, CA: Jossey-Bass.

Black, P., Harrison, C., Hodgen, J., Marshall, M., & Serrett, N. (2010). Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice, 17*(2), 215–232.

Bar-Yam, Y. (1997). *Dynamics of complex systems*. Reading, MA: Addison-Wesley.

Bates, D. M., & Watts, D. G. (1988). *Nonlinear regression analysis and its applications* (Vol. 32). New York, NY: John Wiley & Sons. https://doi.org/10.2307/2289810

Bennett, R. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research & Perspective, 8*(2–3), 70–91. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/15366367.2010.508686

Bollen, J., Van de Sompel, H., Hagberg, A., Bettencourt, L., Chute, R., Rodriguez, M. A., & Balakireva, L. (2009). Clickstream data yields high-resolution maps of science. *PLoS One, 4*, e4803. https://doi.org/10.1371/journal.pone.0004803

Borgatti, S. P., & Halgin, D. S. (2011). On network theory. *Organization Science, 22*, 1168–1181. https://doi.org/10.1287/orsc.1100.0641

Brandes, U., & Erlebach, T. (2005). *Network analysis: Methodological foundations* (Lecture notes in computer science) (Vol. 3418). Berlin, Germany: Springer. https://doi.org/10.1007/b106453

Choi, Y., Rupp, A., Gushta, M., & Sweet, S. (2010). Modeling learning trajectories with epistemic network analysis: An investigation of a novel analytic method for learning progressions in epistemic games. In *National Council on Measurement in Education* (pp. 1–39).

Clarke-Midura, J., Code, J., Dede, C., Mayrath, M., & Zap, N. (2012). Thinking outside the bubble: Virtual performance assessments for measuring complex learning. In *Technology-based assessments for 21st Century skills: Theoretical and practical implications from modern research* (pp. 125–148). Charlotte, NC: Information Age Publishers.

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

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

de Freitas, S. (2014). *Education in computer generated environments*. London, UK: Routledge.

Fischer, G. H. E., & Molenaar, I. W. E. (2012). *Rasch models: Foundations, recent developments, and applications*. New York, NY: Springer. https://doi.org/10.1007/978-1-4612-4230-7

Gibson, D. (2012). Game changers for transforming learning environments. In F. Miller (Ed.), *Transforming learning environments: Strategies to shape the next generation* (Advances in educational administration) (Vol. 16, pp. 215–235). Bingley, UK: Emerald Group Publishing Ltd. https://doi.org/10.1108/S1479-3660(2012)0000016014

Gibson, D., & Clarke-Midura, J. (2013). Some psychometric and design implications of game-based learning analytics. In D. Ifenthaler, J. Spector, P. Isaias, & D. Sampson (Eds.), *E-Learning systems, environments and approaches: Theory and implementation*. London, UK: Springer.

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

Gibson, D., & Jakl, P. (2013). *Data challenges of leveraging a simulation to assess learning*. West Lake Village, CA: Pragmatic Solutions. Retrieved from http://www.curveshift.com/images/Gibson_Jakl_data_challenges.pdf

Gibson, D., & Knezek, G. (2011). Game changers for teacher education. In P. Mishra & M. Koehler (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference 2011* (pp. 929–942). Chesapeake, VA: AACE.

Gibson, D., & Webb, M. E. (2015). Data science in educational assessment. *Education and Information Technologies*. https://doi.org/10.1007/s10639-015-9411-7

Gijbels, D., Dochy, F., Van den Bossche, P., & Segers, M. (2005). Effects of problem-based learning: A meta-analysis from the angle of assessment. *Review of Educational Research, 75*, 27–61.

Han, J., Cheng, H., Xin, D., & Yan, X. (2007). Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery, 15*(1), 55–86. https://doi.org/10.1007/s10618-006-0059-1

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/0470036486

IBM. (2015). *Big data*. Retrieved from http://www-01.ibm.com/software/au/data/bigdata/

Ifenthaler, D. (2010). Scope of graphical indices in educational diagnostics. In D. Ifenthaler, P. Pirnay-Dummer, & N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 213–234). 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.

Jonassen, D. H. (1997). Instructional design models for well-structured and ill-structured problem-solving learning outcomes. *Educational Technology Research and Development, 45*, 65–90. https://doi.org/10.1007/BF02299613

Khan, S. (2011). *Khan Academy*. Retrieved from http://www.khanacademy.org/

Kline, P. (1998). *The new psychometrics: Science, psychology, and measurement*. London, UK: Routledge.

Kopainsky, B., Pirnay-Dummer, P., & Alessi, S. M. (2010). Automated assessment of learner's understanding in complex dynamic systems. In *28th International Conference of the System Dynamics Society, Seoul* (pp. 1–39).

Kozleski, E., Gibson, D., & Hynds, A. (2012). Changing complex educational systems: Frameworks for collaborative social justice leadership. In C. Gersti-Pepin & J. Aiken (Eds.), *Defining social justice leadership in a global context* (pp. 263–286). Charlotte, NC: Information Age Publishing.

Mansell, W., James, M., & The Assessment Reform Group. (2009). *Assessment in schools. Fit for purpose? A Commentary by the Teaching and Learning Research Programme*. London. Retrieved from http://www.tlrp.org/pub/documents/assessment.pdf

Margetts, H., & Sutcliffe, D. (2013). Addressing the policy challenges and opportunities of "Big data". *Policy & Internet, 5*(2), 139–146. https://doi.org/10.1002/1944-2866.POI326

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

Mislevy, R., Steinberg, L., & Almond, R. (1999). *Evidence-centered assessment design*. Educational Testing Service. Retrieved from http://www.education.umd.edu/EDMS/mislevy/papers/ECD_overview.html

Mislevy, R., Steinberg, L., & Almond, R. (2003). On the structure of educational assessments. *Russell: The Journal of the Bertrand Russell Archives, 1*(1), 3–62.

Newell, A., & Simon, H. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Patton, M. (2011). *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. New York, NY: The Guilford Press.

Pirnay-Dummer, P., Ifenthaler, D., & Spector, M. (2010). Highly integrated model assessment technology and tools. *Educational Technology Research and Development, 58*(1), 3–18.

PISA. (2013). *Draft collaborative problem solving framework*. Paris, France: PISA.

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

Psychometrics. (2014). In *American heritage dictionary*. Boston, MA: Houghton-Mifflin Company.

Quellmalz, E., Timms, M., Buckley, B., Davenport, J., Loveland, M., & Silberglitt, M. (2012). 21st century dynamic assessment. In M. Mayrath, J. Clarke-Midura, & D. Robinson (Eds.), *Technology-based assessments for 21st Century skills: Theoretical and practical implications from modern research* (pp. 55–90). Charlotte, NC: Information Age Publishers.

Quellmalz, E., Timms, M., & Schneider, S. (2009). *Assessment of student learning in science simulations and games*. Washington, DC: National Research Council.

Scheffel, M., Niemann, K., & Leony, D. (2012). Key action extraction for learning analytics. *21st century learning …* (pp. 320–333). Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-33263-0_25

Schmidt, M., & Lipson, H. (2009). Symbolic regression of implicit equations. In *Genetic programming theory and practice* (Vol. 7, pp. 73–85). New York, NY: Springer.

Shaffer, D., Hatfield, D., Svarovsky, G., 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. Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/ijlm.2009.0013

Sporns, O. (2011). *Networks of the brain*. Cambridge, MA: MIT Press.

Sterman, J. D. (1994). Learning in and about complex systems. *System Dynamics Review, 10*(2/3), 291.

Stevens, R. (2006). Machine learning assessment systems for modeling patterns of student learning. In D. Gibson, C. Aldrich, & M. Prensky (Eds.), *Games and simulations in online learning: Research & development frameworks* (pp. 349–365). Hershey, PA: Idea Group.

Stevens, R., Johnson, D., & Soller, A. (2005). Probabilities and predictions: Modeling the development of scientific problem-solving skills. *Cell Biology Education, 4*(1), 42–57. https://doi.org/10.1187/cbe.04-03-0036

Stevens, R., & Palacio-Cayetano, J. (2003). Design and performance frameworks for constructing problem-solving simulations. *Cell Biology Education, 2*(Fall), 162–179.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan, 77*(3), 238–245.

Sugihara, G., May, R., Ye, H., Hsieh, C., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. *Science (New York, N.Y.), 338*(6106), 496–500. https://doi.org/10.1126/science.1227079

Tashakkori, A., & Teddlie, C. (2003). *Handbook of mixed methods in social & behavioral research*. Thousand Oaks, CA: SAGE. Retrieved from http://books.google.com/books?id=F8BFOM8DCKoC&pgis=1

Thagard, P. (2010). *How brains make mental models*. Retrieved from cogsci.uwaterloo.ca/Articles/Thagard.brains-models.2010.pdf

Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. L. (2013). Learning analytics dashboard applications. *American Behavioral Scientist*. https://doi.org/10.1177/0002764213479363

Webb, M. E., Gibson, D. C., & Forkosh-Baruch, A. (2013). Challenges for information technology supporting educational assessment. *Journal of Computer Assisted Learning, 29*(5), 451–462. https://doi.org/10.1111/jcal.12033

Webb, M., & Gibson, D. (2015). Technology enhanced assessment in complex collaborative settings. *Education and Information Technologies*. https://doi.org/10.1007/s10639-015-9413-5

Webb, M., & Ifenthaler, D. (2018). Assessment as, for and of 21st Century learning using information technology: An overview. In J. Voogt, G. Knezek, R. Christensen, & K.-W. Lai (Eds.), *International handbook of IT in primary and secondary education* (2nd ed., pp. 1–20). Cham, Switzerland: Springer.

Wiggins, G. (1989). Teaching to the (authentic) test. *Educational Leadership, 46*, 41–46.

Zervas, P., & Sampson, D. G. (2018). Supporting reflective lesson planning based on inquiry learning analytics for facilitating students' problem solving competence: The inspiring science education tools. In T.-W. Chang, R. Huang, & Kinshuk (Eds.), *Authentic learning through advances in technologies* (pp. 91–114). New York, NY: Springer.

David C. Gibson Director of Learning Futures at Curtin University in Australia and UNESCO Chair of Data Science in Higher Education Learning and Teaching, received his doctorate (Ed.D. Leadership and Policy Studies) from the University of Vermont in 1999 based on a study of complex systems modeling of educational change. His foundational research demonstrated the feasibility of bridging from qualitative information to quantifiable dynamic relationships in complex models that verify trajectories of organizational change. He provides thought leadership as a researcher, professor, learning scientist and innovator. He is the creator of simSchool, a classroom flight simulator for preparing educators, and eFolio an online performance-based assessment system, and provides vision and sponsorship for Curtin University's Challenge, a mobile, game-based learning platform for individuals and teams. He consults with project and system leaders, formulates strategies, and helps people articulate their vision for innovation; then helps connect people with the resources needed to fulfil their aspirations. His research has extended from learning analytics, complex systems analysis and modeling of education to application of complexity via games and simulations in teacher education, web applications and the future of learning. Dr. Gibson has also advanced the use of technology to personalize education via cognitive modeling, design and implementation.

Mary E. Webb is Reader in Information Technology in Education at King's College London. Mary is on the International Federation of Information processing (IFIP) Education Executive Committee and leads the Task Force on Computer Science in the Curriculum. Mary is internationally recognised for her research on pedagogy and formative assessment across both Computer Science as a subject and the uses of new technologies for learning. Mary's research has resulted in over 150 publications. Mary is joint editor of the Section: Using Information Technology for Assessment: Issues and Opportunities in the International Handbook entitled the use of IT in primary and Secondary Education which formed the foundation of the EDUsummITs. Mary is joint editor of: Tomorrow's learning: involving everyone, selected papers from World Conference on Computers in Education 2017.

Dirk Ifenthaler is Professor and Chair of Learning, Design and Technology at University of Mannheim, Germany 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 for 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. Dirk's research focuses on the intersection of cognitive psychology, educational technology, data analytics and organisational learning. His research outcomes include numerous co-authored books, book series, book chapters, journal articles and international conference papers, as well as successful grant funding in Australia, Germany and USA—see Dirk's website for a full list of scholarly outcomes at www.ifenthaler.info. He is the Editor-in-Chief of the Springer journal *Technology, Knowledge and Learning* (www.springer.com/10758).