Educ Inf Technol (2015) 20:697–713
DOI 10.1007/s10639-015-9411-7

Data science in educational assessment

David C. Gibson[1] **· Mary E. Webb**[2]

Abstract This article is the second of two articles in this special issue that were developed following discussions of the Assessment Working Group at EDUsummIT 2013. The article extends the analysis of assessments of collaborative problem solving (CPS) to examine the significance of the data concerning this complex assessment problem and then for educational assessment more broadly. The article discusses four measurement challenges of data science or 'big data' in educational assessments that are enabled by technology: 1. Dealing with change over time via time-based data. 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 call for the restructuring of training of the next generation of researchers and psychometricians to specialize in data science in technology enabled assessments.

Keywords Digital assessment · Educational data science · Assessment theory and practice · Virtual performance assessment · Collaborative problem-solving · Tests and measurements · Psychometrics

---

✉ David C. Gibson
david.c.gibson@curtin.edu.au

Mary E. Webb
mary.webb@kcl.ac.uk

[1] Curtin University, Perth, WA, Australia

[2] Kings College, London, UK

1 Introduction

Our previous analysis (Webb et al. 2013) following discussions at EDUsummIT 2011, identified student and teacher involvement in assessment, including digitally-enhanced assessment, as critical for 21st century learning. EDUsummit is a biennial meeting of the original authors, section editors and globally networked colleagues involved in the *International Handbook of Information Technology in Primary and Secondary Education* (Voogt and Knezek 2008), who assemble to create research updates, position papers, and calls to action to improve education. Digitally-enhanced assessments were defined in the 2011 summit as those that integrate: 1) an authentic learning experience involving digital media with 2) embedded continuous unobtrusive measures of performance, learning and knowledge, which 3) creates a highly detailed, high resolution data record which can be computationally analysed and displayed so that 4) learners and teachers can immediately utilize the information to improve learning. This unobtrusive measuring approach is a vision of 'quiet assessment' whose volume can be turned up by learners and teachers whenever they wish in order to check their progress.

This article is the second of two in this special issue that were developed following further discussions of the Assessment Working Group at EDUsummIT 2013. The paper aims to extend our analysis of assessments or evaluations of collaborative problem solving (CPS) as implemented by the OECD PISA started in Webb and Gibson (2015) to examine the significance of big data for this assessment approach and then for educational assessment more broadly. In the companion article (ibid.), the authors argue that in technology enhanced assessment, integration can occur to meet multiple purposes, because *the affordances of technology can redefine the nature of an assessment task*, and provide a high level outline of the processes for engaging in those considerations in the design of assessments of collaboration, particularly CPS. In this article, some of the major data challenges concerning the amount, type and velocity of information potentially available in technology-enhanced assessments are outlined and discussed.

Assessment and learning analytics challenges have dramatically increased since new digital performance affordances, 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 and Knezek 2011; Kozleski et al. 2012). These changes require new quantitative methods as well as a reconceptualization of mixed methods (Tashakkori and Teddlie 2003) that include domain experts as well as stakeholders in the construction of knowledge of such complex systems.

In technology-enhanced assessments, the emergence of "big data" - which are defined as data that has a large numbers of records, of widely differing data types, that is rapidly collected for immediate action (IBM 2015; Margetts and 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 article first outlines the background to our consideration of technology-enhanced assessments and the key issues that are crucial for policymakers, practitioners and learners in the near future. Then it discusses the main challenges associated with applying data science methods in educational assessment to address a digital 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. The OECD PISA plan for assessing CPS is used as an example to explain these challenges in relation to a complex problem space. The article then illustrates with three learning analytics cases that show how the identified challenges have been addressed in the development of assessments.

2 Background

When the EDUsummIT Assessment Working Group met again in 2013 some of the challenges identified in 2011 remained, including uncertainty as to whether and how the following four perspectives on assessment - feedback information, improvement decisions, degree of engagement and understanding, and value judgments - can co-exist to the benefit of learners (Webb et al. 2013). Even with the increased possibilities that IT provides there is not yet a way to say confidently that the multiple purposes for which some assessments have been used (Mansell et al. 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). Therefore in considering assessment design for multiple purposes for example for formative as well as summative purposes, users need to examine those impact factors carefully in order to minimise negative impacts on learning and learners.

Discussions in 2013 led to three main recommendations. First, researchers, policy-makers and practitioners agreed to examine and promote assessment of collaborative learning in problem solving environments as an important and complex digital performance space both for learning and for assessment. For example, significant challenges remain for developing validation approaches that can take account of the complexity of learning experiences for collaborative group tasks. Second, the group saw a need to develop theory for big data in educational research. Third, the group underscored the primacy of the need to engage teachers in the design of learning analytic tools for instructional practices and in interpreting and using results.

The working group concluded that 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 and Dede 2010; Ifenthaler et al. 2012; Quellmalz et al. 2012). Secondly, working with 'big data' needs to be included in educational research preparation and practice, because new tools are needed for discovery of patterns and drivers in complex systems (Gibson 2012; Patton 2011). Indicators of progress on this action item would be first, an increase in articles explaining the use of data science methods in learning analytics to improve learning and the achievement; second, the expansion beyond traditional statistics in educational research, to include data mining, machine learning, and in general, the methods of data science.

3 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). 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 discussed in the companion article (Webb and Gibson 2015). 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. The plan for the OECD assessment of collaborative problem-solving provides an example of an expert conception of how someone solves a problem, conjoined with how they do so in a collaborative environment (PISA 2013). To constrain the quite complex variables that would be involved if the collaboration was among a set of real people, the OECD assessment utilizes the computer to play roles as collaborators in a *virtual performance assessment* (Clarke-Midura et al. 2012). Even with these constraints, the PISA plan incorporates a complex behaviour space that illustrates some of the new demands on psychometrics.

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). The challenge with technology enabled assessments is to preserve these features while evolving the procedural foundations of psychometrics, which until recently have been primarily based on population statistics and static snapshots of data. The new foundation outlined here highlights 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 and Jakl 2013; Quellmalz et al. 2012).

3.1 Time sensitivity

In the OECD assessment of CPS, time is controlled as a boundary variable of the test as in traditional tests 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 is made somewhat more 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 an 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. However, in other cases of virtual performance assessment, time is truly 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) all may be critical to the classification of the learner's response (Gibson and Jakl 2013; Stevens and Palacio-Cayetano 2003).

A metaphor that helps illustrate the time sensitive aspects is to think of the problem of assessing the performance of a Beethoven symphony, a kind of collaborative problem-solving challenge for the orchestra (interpreting and performing) as well as the audience (listening and re-interpreting). It is not helpful to think of averaging all the notes into one event for all four movements (34 min), or for one movement (8 min), or even for one moment of one movement. It is the richness and complexity of the separate notes and *how they change over time* that is the appropriate context for an assessment of the performance; likewise with learning processes and performances such as collaborative problem solving and other learning situations.

The OECD 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 (Table 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 group-based 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 et al. 1999), an approach that is becoming increasingly prominent in assessment design, and on which our analysis is based. The CAF has three core components: the student model, task model and evidence model (Mislevy et al. 1999, 2003a) within and among which the time sensitive relationships adhere.

3.2 Digital performance space relationships

This section reviews how representations of knowledge and know-how have been discussed in test theory, in mental representations and model-based assessment in order to trace a path to the some of the challenges of big data in educational assessment and to make three points. First, 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). Second, the emerging varieties of network analysis (e.g., social networks, visualization, artificial neural networks, decision trees) are critical new analytical tools and methods for understanding technology-enhanced learning (Choi et al. 2010; Shaffer et al. 2009). Third, 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 and Simon 1972) and the knowledge and capability they acquired, accessed and utilized during the interaction (Pirnay-Dummer et al. 2010; Thagard 2010). This set of ideas, which have components in the real world as well as in the learner's mind, will

Table 1 Matrix of Collaborative problem solving skills for PISA 2015

	(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 he 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

be referred to as ‘*digital performance space relationships*’ which are taken 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 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. For example, in the OECD assessment discussed in the companion article, 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’ (Table 1). A review of the historical idea of a valid construct will be helpful for making the bridge from classical testing to the digital age.

In the mid-1950s the problem of validating a test was discussed by psychologists in order to address a concern that a variety of ideas about ‘construct validity’ had arisen in the preceding years, which opened the door to nonconfirmable test claims (Cronbach and Meehl 1955). A proposal was therefore put forward to conceptualize the problem of construct validity using the idea of a *nomological network* from the philosophy of science. A nomological network is a collection of overlapping *mappings* (i.e., statistical or deterministic rules that relate one thing to another) 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 *construct can change* and *become more elaborated over time*, as Cronbach noted:

> 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.

The construct was thought of as an *inductive summary* and as *part of a series* of validity investigations that included concurrent, predictive and content considerations. The idea of a network of ideas and relationships was a fairly abstract philosophical idea in the 1950’s but today has a concrete meaning that has become known as network theory in social science (Borgatti and Halgin 2011) and network analysis in computational sciences, both of which are applied graph theory from mathematics (Brandes and Erlebach 2005).

Bridging from the historical roots of construct validity in measurement theory into the present day of digital media learning experiences, a complex performance (a complex construct) can be recorded in high detail in terms of the actions, communications and products created by the learner. A bridge is possible because the structure and affordances of a digital performance space (e.g., the resources and affordances for action, communication and creation of artifacts) can be represented as a network by mapping each digital resource to a node and each action, communication or product creation relationship to an edge connecting nodes to one another.

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 et al. 2005). This section asserts that 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 (Cronbach and Meehl 1955); 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 (Table 2). 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

Table 2 Evidence Centered Design claims vs Nomological Network construct validation

Evidence-Centered Design Claims	Nomological Network Claims
Claims about what someone knows and can do based on an assessment are supported by a chain of reasoning or argument leading from data to the claim, which is supported by warrants (e.g., hypotheses or truth statements) and backing (e.g., historical data).	A claim that a test is a measure of a construct requires a chain of inference from a network of propositions of observables, ideas, and their relationships (the nomological network) to the construct.
A claim can face a counter-argument supported by alternative hypotheses and rebuttal data.	An individual measure only addresses and utilizes part of the nomological network.
Claims may be remote from data	Claims concerning a construct may be remote from observation
Evidence is interpretation of data to make a claim	Nomologicals may be derived from other parts of the network or new observations
Establishing validity entails making the warrant explicit, examining the network of beliefs and theories on which it relies, and testing its strength and credibility through various sources of backing. It requires determining conditions that weaken the warrant, exploring alternative explanations for good or poor performance, and feeding them back into the system to reduce inferential errors Mislevy et al. (2003b).	Establishing construct validity entails making contact with observations, and exhibiting explicit, public steps of inference (Cronbach and Meehl 1955)

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 and Ifenthaler 2011; Quellmalz et al. 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 assessment of CPS, 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, e.g., 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 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 et al. 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 1 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-ngrams (*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 and 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 and 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 preceding sections' 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 et al. 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 all the data falls 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 and 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, use of documents and tools, and artefacts produced by students (Verbert et al. 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 (other than those perhaps in the minds of expert teachers) have not yet been documented in some 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 cases in the next section, for future research and practice to create a shared understanding of the methods of data science in educational research.

4 Big data lessons from cases

Three cases illustrate 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.

4.1 Case 1: Virtual performance assessment

A game-based assessment of scientific thinking was created at Harvard (Clarke-Midura et al. 2012) and analysed by one of the authors (Gibson and 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 artifacts 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 et al. 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 et al. 2005). In particular, each pathway in such a network can be further characterized or specified with a predictive nonlinear mathematical relationship (Gibson and Clarke-Midura 2013), found through *symbolic regression*, an evolutionary machine learning technique (Schmidt and 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 analyses* 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).

4.2 Case 2: Student performance and completion in a MOOC

In the second case, a massively open online 4-week course (MOOC) in astronomy at a large university in Western Australia was the setting for an analytics study of the relationship of activity to completion (DeFreitas et al. 2015). This case again illustrates that volume of data, compared to traditional norms in educational research, is one of the new characteristic features; the size and resolution of the MOOC data comprised 31,000 records for 177 subjects or 175 records per subject. The data set did not have a high degree of variety so simple *descriptive statistics* were used to establish regimes of learner behaviour and completion status, which were then explored for predictive relationships. *Symbolic regression* equations were found to indicate plausible structural relationships between the digital performance space and learner performance characteristics, and the findings tended to confirm literature-based findings of a positive relationship between consistency of effort (Duckworth et al. 2007) and age. It is lamentable perhaps that early versions of massive open online courses are not able to utilize rapid analysis to provide feedback to learners until enough performance history has accumulated, so the users are not often provided with feedback based on such analyses. In this case, the analysis worked with traditional data structures and confirmed previous research while adding new detail about the specifics of the predictive relationship. For example, instead of only knowing that there is some positive linear correlation between age and consistency of effort in the MOOC, a specific nonlinear equation was produced specifying how much consistency of effort and in what relationship to age as a factor. This illustrates that a goal of data science in technology-enabled learning might integrate descriptive with inferential purposes.

4.3 Case 3: A study of retention

In the third case, an analytics methodology followed *a staged process of stakeholder involvement in data acquisition, preparation, discovery and analysis*, for the creation of a *self-organised map* that facilitated the simultaneous testing of over 50 hypotheses – a perhaps novel data science method in educational research (Gibson and de Freitas 2015). The three aspects of big data (volume, variety and velocity) are again noticeable in this case. The volume of data comprised 13 million records for 52,000 subjects, or about 250 records per subject. A wide variety of data was collected from ten digital sources that included study patterns, performance in units of study, attendance, survey question answers, demographic profiles, library records, and other diverse sources. The velocity element was represented by the short timeline for data acquisition, preparation, discovery and analysis managed in stages and engaging over 200 people in focus groups and feedback sessions. The *machine learning* method applied during the exploratory phase was an unsupervised *self-organizing map,* which proceeds based on a multidimensional similarity metric. Here, as in case 2, *nonlinear methods* were combined with *linear statistics* as appropriate.

The list of key elements and methods from these cases (Table 3) are relevant to technology-enhanced assessment of collaborative learning for several reasons: the potential scale of the data records, the inclusion of time as a mediating factor, the need for machine learning and data mining methods and the possibility of simultaneous testing of multiple hypotheses. The OECD plan to assess problem solving in a collaborative context, for example, transparently utilizes the notion of *motifs* and *dynamic data reduction* of action, communication and products because although the data record of the assessment might contain a large number of clicks, products and communications, the aggregation of elements into only 5 to 30 classifications (i.e., mini-assessments) per scenario, with each student input considered a 'test item' indicates that a layer of classification must take place on top of the click track, guided by the domain model (Table 1) into collections of evidence with constrained options (e.g., an action might be a binary 'low or high' on some scale, and the total number of scenario options is limited to 16 configurations of the digital performance space).

A rationale for a new foundation for research methods has been provided in Gibson (2012) based on the rise of complexity and data science and the observation that global communications and business interactions today have outpaced the tool sets for research in the learning and behavioural sciences. The evidence from these cases and from what the OECD assessment of collaborative problem solving is planning to do also supports the need for an expansion of the tools of educational research to include data science methods.

5 Conclusion and implications for teaching and learning

This article 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

Table 3 List of cases with key elements and methods

Case	Data Scale	Time & Processes	Analysis Foci
Virtual Performance Assessment	821,000 records for 4000 subjects 205 records per subject	*patterns of action* were transformed into *motifs* or *n-grams* user *actions* provided evidence of learning and thought processes	*empirical probabilities* were used to make inferences *symbolic regression* revealed predictive nonlinear relationships *association rule networks* distinguished user action patterns and motifs *cluster analyses* revealed structure of the digital performance space
MOOC Performance & Completion	31,000 records for 177 subjects 175 records per subject	*descriptive statistics* established regimes of learner behavior and completion status	*symbolic regression* equations found plausible structural relationships between the digital performance space and learner performance characteristics
University Retention	13 million records for 52,000 subjects 250 records per subject	*a staged process of stakeholder involvement in data acquisition, preparation, discovery and analysis* *lifetime retention defined to cover impacts over time*	*machine learning* using a unsupervised mapping based on multidimensional similarity *self-organizing map* created with a multidimensional similarity metric *nonlinear methods combined with linear statistics*

interactions between and among learners who are being assessed by their interactions in digital performance spaces. The article linked the big data challenges to solutions offered in the OECD PISA assessment of collaborative problem solving, and then reviewed some of the same issues by briefly summarizing how the newer methods were used in three additional cases.

The challenges and issues discussed in this article 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 our sister article (Webb and Gibson 2015).

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. In the sister article (Webb and Gibson 2015) some of the potential risks associated with technology-based assessments are examined, especially if they: 1) focus only on assessing what is possible to be assessed by technology; 2) fail to enable learners and teachers to understand the basis of judgements made and 3) do not involve teachers in the design of assessment.

References

Al-diban, S., & Ifenthaler, D. (2011). Comparison of two analysis approaches for measuring externalized mental models. *Educational Technology and 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: Jossey-Bass.

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

Bates, D., & Watts, D. (1988). Nonlinear regression analysis and its applications. John Wiley & Sons (Vol. 32). doi: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 and Perspective, 8*(2–3), 70–91. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/15366367.2010.508686..

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

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. doi:10.1371/journal.pone.0004803.

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

Brandes, U., & Erlebach, T. (2005). *Network Analysis: Methodological Foundations. Lecture Notes in Computer Science* (Vol. 3418). doi: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., & Dede, C. (2010). Assessment, technology, and change. *Journal of Research in Teacher Education, 42*(3), 309–328.

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: Information Age Publishers.

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

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

DeFreitas, S., Gibson, D., & Morgan, J. (2015). Will MOOCS transform teaching and learning in higher education? *British Journal of Educational Technology, in press.*

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*(6), 1087–1101. doi:10.1037/0022-3514.92.6.1087.

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, Volume 16)* (pp. 215 – 235). Emerald Group Publishing Ltd. doi: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: Springer.

Gibson, D., & de Freitas, S. (2015). Exploratory analysis in learning analytics. *Technology, Knowledge and Learning*. doi:10.1007/s10758-015-9249-5.

Gibson, D., & Jakl, P. (2013). *Data challenges of leveraging a simulation to assess learning*. CA: West Lake Village. 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.: AACE.

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. doi:10.1007/s10618-006-0059-1.

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

Ifenthaler, D., Eseryel, D., Ge, X., & Ge, X. (2012). In D. Ifenthaler & D. Eseryel (Eds.), *Assessment in game-based learning*. Lobdon: 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. doi:10.1007/BF02299613.

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

Kline, P. (1998). *The new psychometrics: Science, psychology, and measurement*. London: 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* (pp. 1–39). Seoul.

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: 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 and Internet, 5*(2), 139–146. doi: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. (2003a). On the structure of educational assessments. *Russell the Journal of the Bertrand Russell Archives, 1*(1), 3–62.

Mislevy, R., Wilson, M., Ercikan, K., & Chudowsky, N. (2003). Psychometric Principles in Student Assessment. In T. Kellaghan & D. Stufflebeam (Eds.), *International Handbook of Educational Evaluation* (pp. 489–531). Kluwer Int.

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

Patton, M. (2011). *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. New York: 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*. OECD: Organization for Economic and Community Development.

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

Psychometrics. (2014). In *American Heritage Dictionary*. Houghton-Mifflin Company

Quellmalz, E., Timms, M., & Schneider, S. (2009). *Assessment of Student Learning in Science Simulations and Games*. DC: National Research Council.

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: Information Age Publishers.

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. *Genetic Programming Theory and Practice, 7*, 73–85.

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: 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: Idea Group.

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

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. doi:10.1187/cbe.04-03-0036.

Stiggins, R. J. (1995). Assessment Literacy for the 21st Century. *Phi Delta Kappan*, *77*.

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. doi:10.1126/science.1227079.

Tashakkori, A., & Teddlie, . (2003). *Handbook of Mixed Methods in Social & Behavioral Research*. 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*. doi:10.1177/0002764213479363.

Voogt, J., & Knezek, G. (2008). *International Handbook of Information Technology in Primary and Secondary Education*. New York: Springer.

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

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