Chapter 13
Standards for Developing Assessments of Learning Using Process Data

Sandra Milligan

Abstract Digital technology is changing assessment of learning. Digitised assessment can be more administratively efficient, more easily scaled, more effectively targeted to individual levels of performance, more integrated into the learning environment, more interactive, and it can support more imaginative, colourful, interactive and timely feedback. However, in this chapter, it is argued that 'more, faster and prettier' is only part of the assessment story of the first quarter of the twenty-first century. Education institutions are also being pressed to make distinctive shifts in what is learned and thus in what is assessed. Students now need to establish mastery of complex learning outcomes that extend beyond the cognitive domain, and beyond mastery of content knowledge, to mastery of competence and skill, including soft skills, or general capabilities. This chapter explores this assessment frontier, examining whether and how large quantities of digital, process-oriented data generated from learning management systems and other digital learning tools can be used to make reliable and valid judgments about the degree to which student have mastered complex general capabilities. It is argued that 'metrolytic' standards for development of assessment tools can be applied to ensure the requisite validity and reliability.

13.1 Introduction

Shifts in thinking about assessment are not new. Assessment practice has always been responsive to the educational and social concerns of the day (Pellegrino 1999). For instance, the contemporary methods of psychometrics can be traced back to the interest in individual differences of anthropologists and eugenicists in the nineteenth century. The statisticians of the day, including Fisher, Spearman, and Pearson developed methods to manage evidence of individual attributes, many of which are still routinely applied. The multiple choice question was introduced in the early

S. Milligan (✉)
Assessment Research Centre, University of Melbourne, Parkville, VIC, Australia
e-mail: s.milligan@unimelb.edu.au

M. Bearman et al. (eds.), *Re-imagining University Assessment in a Digital World*, The Enabling Power of Assessment 7,
https://doi.org/10.1007/978-3-030-41956-1_13

twentieth century, in the attempt to improve objectivity and fairness into wide-scale assessments of applicants, to rank them reliably for positions in the US military. During the 1980s and 1990s, community demand for fairer selection methods, and greater accountability for student learning resulted in refinement of standardized testing methods. Statistically sophisticated methodologies were used with automated administration in support of large-scale ranking and monitoring, usually focused broadly on assessing 'scholastic aptitude', or mastery of basic learning like literacy, or numeracy.

Today, digital tools used in university teaching are visibly changing assessment practice, and the character of the relationship between assessor and assessed. Teachers have embraced digitally mediated tools to set assessments, scaffold student response to them, monitor cheating, harvest responses, mark responses, provide feedback, grade, compile and report. Digitisation of assessment procedures is administratively efficient, allows for scaling to accommodate large classes, allows more effective targeting to individual levels of student performance, is more interactive, and it can support more imaginative, colorful, interactive and timely feedback and faster more direct reporting. Digital scaffolding of assessment makes feasible assessment methods that require complex administration, including for example, peer and self-assessment (see Tai and Adachi, Chap. 15, in this volume). Embedding assessment into the learning environment of digital learning management systems allows better integration of teaching and learning and supports formative assessment practices. Digital mediation of teaching and assessment via ubiquitous learning platforms is now common, even in small on-campus classes. These forms of technological enhancement have arguably enabled assessment processes to become more efficient, faster and prettier, more responsive, more formative, more timely.

However, a further frontier for assessment is emerging that goes beyond technological enhancement of teacher assessment practice, to exploring the use of powerful digital technologies and digital data, particularly process data, to better assess and report on the growth in learning, particularly learning of complex competencies and general capabilities. This chapter focuses on this frontier, describing the difficulties associated with generating valid and reliable assessments. It argues that new standards for development of assessment tools—metrolytic standards—are required. These standards draw on methods used in the learning analytics community, combined with those commonly used in educational measurement, to provide a framework designed to ensure that any assessment can be trusted to have the required level of reliability and validity to warrant use for educational purposes.

13.2 Contemporary Pressures on Assessment Practice

The frontier arises directly from contemporary pressures on universities to change what students learn and how they learn it, which consequently changes what teachers assess and report (Griffin and Care 2015). The core idea is that the '4th industrial revolution' is in train, requiring educators to produce learners with different skills and capabilities than required by generations past (Tremblay et al. 2012; OECD 2018; Milligan et al. 2018). No single, simple driver can be identified. A range of factors are involved, including the ubiquity of digital communications and computing technology, the rapid expansion of knowledge, the impact of globalisation, and the increasing commitment to ensuring sustainable, equitable development for human wellbeing. All of this is redefining how we live and work in the twenty-first century. The net effect on education institutions is that they are being pressed to redefine what learners need to know, and to produce graduates who can demonstrate attributes other than just mastery of knowledge in a content domain. Students are now required to demonstrate *knowhow* in a domain of study, as well as content mastery. Curricula are being altered to supplement the cognitive outcomes of traditional disciplines with requirements that learners develop the constellations of knowledge, values, attitudes, skills and beliefs required for competent performance in any field (Dreyfus and Dreyfus 1980). This is sometimes characterised as a shift from *content* to *competence* in curricula (Griffin 2007). In this context, curriculum outcomes are extended beyond mastery in discipline or professional domains, to encompass attributes variously referred to as ‘soft’ or ‘21st Century skills’ or ‘transversal skills’, or (as referred to in this chapter) ‘general capabilities’, that might be applied in any field or domain (Asia-Pacific Education Research Institutes Network 2015; Griffin and Care 2015). For instance, the World Economic Forum (2015) described a range of required general learning capabilities, including critical thinking skills, communication, creativity, collaboration, scientific and ICT literacy, persistence and curiosity, amongst others.

A related shift requires that students develop skills as lifelong learners (Bransford et al. 2003). The argument is that it is not enough that students can learn when directed by teachers, in formal educational settings. They must also be able to learn by their own initiative. In modern times, this ability to learn is not considered to be a matter of IQ or any innate ability – it is more to do with the mastery of a set of knowledge, skills, understandings and beliefs about learning that equip individuals to a greater or lesser degree with the capability they need to learn (Milligan and Griffin 2016).

Formal inclusion of these general capabilities in curricula represent a professional challenge for teachers, especially in courses which use traditional approaches to higher education assessment, such as summative assessments made at the end of course, based on the evidence presented to a teacher via student essays. Now, the challenge is to assess the degree to which a learner has also mastered general capabilities.

Assessment of this sort is several degrees more difficult than assessment of mastery of content knowledge in the cognitive domain. It is a new field for teachers, and is especially challenging in large classes where teachers do not necessarily know their students. Assessment tasks that enable individuals to demonstrate complex capabilities are usually themselves complex, and often need to be conducted in non-standardized environments, involving performances or building of artifacts, or working with peers, sometimes in teams. Mastery of these capabilities usually takes time and practice, often in ‘authentic’ learning environments unlike traditional lecture halls. Feedback on performance is required at various stages of development, enabling learners and other stakeholders to chart the learner's gradual increase in mastery. Thus, in parallel with changes in curriculum, there are requisite changes in the purposes and methods of assessment. Assessments need to assist learners and teachers to determine the degree to which a learner has mastered complex general capabilities in a domain of learning. Important methodological frontiers for assessment are opening in this area.

13.3 The Promise of Big Data to Assess General Capabilities

Scardamalia et al. (2013) reported on the findings of a large, international research project established to explore how best to assess these general capabilities. They concluded that the best assessments are "embedded in the technologies used in the learning environment, concurrent and transformative" (p. 34). They pointed out that embedded technologies can generate automatic feedback, provide on-demand assessments, and prevent or reduce the separation of assessment from the learning experience. Response rates are unproblematic, as participant activities are always reflected in the log stream, and data collection requires no additional effort by participants. Use of digital traces of learner activity for assessment purposes has the potential to provide real-time calculation of scores, and could provide greater timeliness in calculation and feedback to participants during a course.

On the face of it, this approach seems feasible. There are copious quantities of learning-related digital information now available, including click stream data that captures every mouse click, swipe, or keyboard action of every learner as they use digital learning applications. Other information can be obtained from sophisticated digital data sensors in classrooms that capture anything from eye-gaze direction to heart-rate, from speech to physical movement. The capacity to 'see' what students say, do, make, or write in learning environments is vastly enhanced. Available data goes well beyond the inputs to traditional university assessments, such as observations by teachers in classrooms, or responses to assessment tasks, or results from standardized tests. Data now systematically captures information on the *process* of learning not just the outputs from it.

Such data can also be interrogated for meaning by a plethora of modern analytical methods, including, for example social network analysis, text analysis and various forms of data mining. These enable the construction of statistics that can, in theory, be used as indicators of performance, for assessment purposes. Teachers can 'see' the degree of connectedness to others in the class (from network analysis), the focus of a student's interests (from text analysis), the systematicity of study habits (from time-series analyses) and so on. Techniques of data mining and artificial intelligence can be applied to these indicators (He et al. 2016) potentially adding value. Data or analyses of this sort are now routinely presented in digital dashboards, or otherwise provided as feedback to teachers and learners (Corrin and DeBarba 2014).

Researchers early into the learning analytics field (Carmean and Mizzi 2010; Gasevic et al. 2015; Siemens and Long 2011; Greller and Draschler 2012) expected that such data would have many benefits for learners and teachers, making visible the process of teaching and learning, supporting reflection on practice by learners (on their learning) and by teachers (on their teaching); predicting and modeling learning, leading to better intervention; and making possible personalization of learning through real-time tracking and analysis of each individual. If backed with artificial intelligence tools, it is argued, digital responders might become more capable than their human predecessors in assessing learning.

This optimism was supported by belief that use of big data was not only possible, but it is the *preferred* option when it comes to assessing individuals' performances in the kind of complex general capabilities which are now the focus of curriculum (Scardamalia et al. 2013). The traditional techniques for assessing individuals' attributes or capabilities include use of self-report scales, direct observation by experts, employing think-aloud protocols, analysing artifacts like respondent diaries, and using micro-analytic approaches involving codification of behaviours such as eye-gaze or facial expression to infer an individual's level of an attribute (Cleary et al. 2012). But, such techniques are impractical for use in real learning environments. They are too costly and labour-intensive, leading teachers and assessors to look for better, more practical approaches. Naturally, they look to the possibility of using big, digital data, derived from sensors embedded in the learning environment, that provide systematic evidence about the process of learning used by learners.

13.4 The Need to Ensure Validity and Reliability

While the optimism is high, there are acknowledged difficulties. Scholars of learning analytics have always been quick to point out that that bigger digital data generated as a by-product of learning is not always better data (Greller and Draschler 2012; Siemens and Long 2011). A key question not yet convincingly answered is whether or not these digital traces can be used to construct indicators of learning, or whether they merely record processes that *may or may not* reflect the degree of learning attained. In addition, it is not clear whether the process data embodies sufficient information: perhaps missing elements may be exactly those required to explain learning. Platforms or digital sensors cannot capture all 'off-line' activity such as reflection or note taking, or what student are thinking, but these missing elements might be vital (Gunnarsson and Alterman 2013). Big natural databases tempt the search for interesting relationships based on correlations, or factor and cluster analysis. When interesting patterns are found, as they are bound to be, and if they attain statistical significance, the temptation is to impute explanatory value, and to infer meaning about learning. But it is not at all clear that these interesting, statistically significant patterns are a suitable basis on which to judge an individual's learning. The patterns may in reality be products of chance, or be inconsequential for learning, and may not evidence an underlying phenomenon with explanatory value. Statistical relationships demonstrate only that discovered relationships are unlikely to be random, which is an insufficient basis for interpreting any numbers as measures of learning for an individual.

The most robust of the contemporary work using process data to assess and report on the growth in complex competencies and general capabilities uses a combination of analytics methods and the methods of educational measurement (Buckingham Shum and Deakin Crick 2016; Griffin and Care 2015; He et al. 2016; Milligan and Griffin 2016; Shute and Ventura 2013; Wilson et al. 2016; Polyak et al. 2017). Measurement principles and techniques such as those that underpin Wilson's (2005) Constructing Measures approach, or the Evidence Centered Design approach (Mislevy and Haertel 2006) are designed to engender trust in assessments by ensuring that scores measure something of value, to a requisite standard. A careful, methodical, iterative, curriculum-focused process is used to develop scores for individuals, involving the development of constructs and evidence maps, sampling of evidence using rules and procedures, and so on. The measurement sciences establish standards that evidence claims that measures are appropriately used to assess learning – that is, they are valid and reliable, and can safely be used to judge an individual's learning progress.

It is notable that researchers who have adapted traditional measurement techniques to analytics-based process data tend to be cautious, reflecting a growing awareness in the assessment and analytics communities that such work is still in its infancy. The key to understanding the difficulties associated with this frontier is that, when constructing measures of complex constructs from digital data, assumptions crucial to quality of assessment need to be made explicit, and tested (Wright and Masters 1982). For instance, assessment of learning is based on an assumption that individuals can possess different amounts of an attribute being assessed, and that a description of that attribute should provide the basis of assessment. The underlying attribute must make sense and be plausible, and there should be practical benefit in assessing it. The attribute must have dimensionality and it must be possible to understand how it is that people can have more or less of it. 'More' or 'less' must be capable of consistent representation as a progression for all individuals, using units of equal value, and the units should be additive and repetitive. Although attributes cannot be directly observed, it must be possible to understand how different levels of performance is explained by differences in observable behaviour of individuals: what individuals do, or say, or make, or write (Glaser 1994a, 1994b). The behavioural differences must have explanatory value, and it should be possible to infer from these observed differences in behaviours the amount of the attribute that a person has. Such assumptions need to be tested and evidence and argument presented to satisfy stakeholders that the use of assessments for educational purposes are warranted. An important source of thinking about the standards that should apply can be found in discussion of validity in measurement science (Cronbach and Meehl 1955; Kane 2013; Messick 1995; Wolfe and Smith 2007) and in discussion of quality in learning analytics (Dringus 2012; Greller and Draschler 2012).

Table 13.1 outlines a set of indicative ‘standards’ that might be used to interrogate the quality of any score, before it is used to make decisions about an individual. These standards are derived from both the practice of measurement science and learning analytics. They are best described as metrolytics standards, derived from the Greek words *metron*, which is the root of the word measurement and means limited proportion, and *analutikós*, meaning to analyse. In an ideal world, the standards would provide the basis for presentation by an assessment designer of evidence and argument to support the validity and reliability of any assessment constructed from process data, in exactly the same way that high-stakes test developers are required to present arguments as to the validity and reliability of their tests. A set of indicative standards is presented in Table 13.1.

Recently, the learning analytics community expressed concern about the burgeoning array of analytics apps, and whether or not there is a sufficient evidence base about the use of analytics to warrant trust by stakeholders (Bergner et al. 2017; Ferguson and Clow 2017). The concerns intensify when the results are used to control or shape the treatment a person receives, as is often the case in assessment of learning. It would not be unreasonable for stakeholders (learners, teachers, employers, professional associations), to mistrust that an instrument can assess complex attributes, especially if it combines a variety of data types and utilizes complex data transformation or algorithms.

Adoption of metrolytic standards provides one way to address this issue, requiring methods that ensure that an assessment has utility for the intended purpose. The standards provide the framework for presenting the evidence and argument to engender trust in the assessments. In this, it is important that evidence presented should not merely defend the interpretation of an assessment. Rather, it should be capable of convincing a reader that the assessment design or method tested the assumptions on which it was based, and that no other plausible alternative interpretations of what is being assessing are supported. This involves imaginatively identifying and examining potentially disconfirmatory evidence as well as confirmatory evidence, to address what are the risks that could apply to a particular assessment.

13.5 Methodological Frontiers

Consideration of the metrolytics standards outlined in Table 13.1 highlights some of the practical difficulties facing analysts attempting to build reliable and valid assessments of complex general capabilities using process data. For instance, the standard requires as a prerequisite clarity about what is to be assessed. In traditional classrooms, what is to be assessed is usually operationally defined as ‘what has been taught’, often content-based. Assessments of the newer general capabilities require a similarly clear view of what is being assessed. In practice, this demands specification of the

Table 13.1 Indicative metrolytics standards for assessments derived from digital process data

Utility There is a clear purpose for making assessment(s) that are of value to stakeholders.
There is clarity about the nature of the attribute There exists a clear definition of the attribute being assessed, expressed as the constellation of knowledge, understandings, skills, beliefs, attitudes, and values required for mastery, differentiated at each level of mastery. This definition should be understandable and acceptable to stakeholders, including teachers and learners.
The attribute has dimensionality It is plausible to assume that individuals have more or less of the attribute of interest, and can be arrayed along an underlying continuous scale from more to less. Ideally, the scale reflects typical learning trajectories of learners' progress from level to level as their mastery grows. These trajectories, typically called 'progressions' or 'learning continua' must make sense to learners and teachers alike.
Data relates to learning behavior Selected data comprises representations of what learners say, do, make or write when learning. Data does not include characteristics or status of individuals that might be related to learning, (such as dispositions or social or demographic features) but which do not reflect learning gains in a particular environment.
Process data is 'clean' and understood Data husbandry methods are used, including, for instance: checking plausibility of the range and distribution of values; transforming raw log stream data into variables, counts or categories; identifying and mitigating corrupt, incomplete, misleading or miscoded data; ensuring that data definitions do not change over time; ensuring data definitions are stable and uniform; ensuring granularity matches the purposes to which data were being put (e.g., in time series studies, should analyses be precise to the second, to the hour, to the week or to the year); adopting analytical and sampling techniques to manage high volume; and so on.
Data is mapped to the levels of progression using statistical indicators The data elements selected to construct the assessments should be capable of generating stable behavioural indicators for each learner. For instance, when using network analysis, interactions between students might generate statistical indicators of connectedness. Differences in behavior of individuals on each behavioural indicator should plausibly explain different levels of attribute of an individual.
Interpretation of indicators is stable Direct comparability of scores over time is maintained. This is especially important when machine learning is used to construct scores or indictors. If the algorithms change indicators, the construct being measured might change. Any teaching policy changes are identified which may change inferences that can be made from data: for instance, voluntary participation in forums might normally reflect student engagement, but, if it is compulsory to post, it might reflect compliance.
Attribute is fully represented The behavioural indicators comprising the score together provide a balanced and full representation of the attribute along the full range of levels; indicators are not skewed by missing or irrelevant data; there are no important features missing. For instance if off-line work is vital to a process, it is difficult to see how an online automated assessment can be other than skewed.
Scoring and data transformation is transparent There is a transparent audit trail describing each stage of the transformations of data to indicators to scores. Metrics and algorithms at all stages of development are transparent.
Technical quality is adequate The psychometric qualities of accuracy, discrimination and reliability of indicators and scores are demonstrated: the indicators cohere with a simple developmental integrity, generating a scale with even intervals, and the criterion of conjoint measurement applies, there are no biases evident for sub groups. These characteristics can be examined by testing fit to a measurement model.

(continued)

Table 13.1 (continued)

Scores are interpretable There is no other plausible interpretation of the assessments other than that they reflect differences between individuals in the level of competence or capability they possess.
Alternative methods are canvassed There are no simpler, alternative methods for arriving at assessments.
Unintended consequences are identified There are likely to be no unintended negative consequences arising from use of the assessment that arise from shortcomings of the assessment.
Appeals are possible There is a process of appeals for review, important if assessments are the product of complex algorithms that are difficult for stakeholders to understand.

constellation of knowledge, understandings, skills, beliefs, attitudes and values that define different levels of mastery, based on understanding of typical trajectories or progressions described by learners as they learn. One difficulty for assessment designers seeking to use process-based data is that there are as yet not many examples of such progressions that describe the development trajectories for general capabilities. The first task of teachers or analysts seeking to design assessments is, therefore, to define a progression that plausibly describes the likely patterns of behaviours of individuals who have more or less of the attribute, so that they be arrayed along an underlying continuous scale from more to less. Defining a progression is not a simple task, and is often neglected by analysts or teachers keen to skip straight to the data. Unfortunately, without a theoretical progression on which to base empirical work, there is no means of determining the validity, utility or interpretability of assessment scores.

There are also difficulties in assessing individual performance when the individual is performing with others in a group. This is especially difficult for assessment of general capabilities like teamwork and collaboration, since such attributes are exhibited only in social environments. Teachers understand this, and work through it every time they assess group work. The relationship between performance of the group and the performance of the group member is complex. This complexity is endemic in data derived from digital forum participation, or collaborative work, or multi-user interactive activities. Current measurement and assessment methodologies struggle to untangle the complexities sufficiently to support individual assessments from confounded data of this sort. It is notable that, after reviewing the last 10 years of years of solid, large-scale psychometrically-based work on the topic of assessment of collaborative problem solving capacity, a recent report for the prestigious National Assessment of Educational Progress (NAEP) in the US acknowledged that the world is not yet in a position to a confidently make claims to be able to measure the collaborative problem solving ability of students (Fiore et al. 2017). Solutions to this problem are still developing (Wilson and Scalise 2016). A non-technical approach is to recognise that when the performance of a group in a domain is at issue, it is appropriate to focus assessment on the performance of the group rather that the performance of the individual. This approach would involve assessors taking seriously a socio-centric view of knowledge generation rather than a psycho-centric view of it. A more technically oriented team (von Davier and Halpin 2013) noted that the confounding effects of group membership are traditionally treated as ‘random error’ in assessments. They have proposed turning this around, psychometrically, so that indicators of the degree of error in individual measurement might be used to indicate presence of group capacity independent of individuals’ skill.

Further work is also required to determine whether or not an individual’s capacity to perform at particular levels on measures of general capabilities such as problem solving, or communication, or persistence, are generalisable, and thus worth assessing in one context or domain, because they are transferable to another. For example, will a learner who demonstrates good problem solving skills in a chemistry classroom demonstrate good problem solving skills in workplace, or in a physics lab? Will a student who collaborates well in online games be a good collaborator in face-to-face groups? Early research suggests that in general, transfer of these general capabilities, with some notable exceptions, is low (Perkins and Salomon 1992). The interpretations of scores relating to complex general skills in any given context or domain might need to be treated circumspectly, regarded as being context dependent, and used in a way that recognises their limits.

Teachers or assessment designers needing to collect evidence about complex competencies may decide to combine diverse evidence sources (such as peer, self and teacher assessments of portfolios, or combining patterns of participation in forums and in lectures). Unfortunately, when the relationship between diverse indicators is unexamined, it may be that the combined results are of poor quality. Unless each indictor taps into the same underlying attribute, the resultant assessment is likely to be characterized by poor validity, low reliability, inaccuracy and poor utility.

A legitimate response to technical difficulties is to question whether some general capabilities might best be left unmeasured, creating no-go zones for individual assessments. For instance, Masters (2018) gives the example of ‘creativity’, notoriously difficult to assess in any environment. He suggests that it may be appropriate to question whether there exists a generalisable attribute called ‘creativity’ that can be assessed, even in a specific domain. Unless it is possible for the community of stakeholders to reach agreement on a clear definition and progression, it may not be possible to assess.

In addition, the construction of measures that meets metrolytic standards is time consuming, expensive and demands technical skill. It becomes economically feasible only when used at scale. This in itself is likely to limit use of these methodologies in the short term.

This discussion highlights some of the practical difficulties associated with using process data to generate high quality assessments of general capabilities. Methodological challenges include that there is a paucity of clear, empirically derived progressions for these capabilities which need to underpin assessment, the scores generated may lack generalisability, there are questions about feasibility, and technical methodologies are currently limited when it comes to disentangling the effects of group membership on achievement. Further there are dangers in combining scores. Even the robust methodological tools of measurement science may fall short, for a range of technical and transparency reasons. Skepticism is warranted that any particular assessment tool will have basic qualities such as reliability, validity, accuracy, utility or interpretability.

13.6 Conclusion

The contemporary frontier for assessment focuses on whether and how to use powerful digital technologies to make use of digital data, particularly process data, to better assess and report on the attainment in learning, particularly of general capabilities. The underlying danger is always that 'data' will be mistaken for 'assessments'. This area of assessment is challenging, and it is still uncertain to what degree the new, big data sets, especially process data, will support valid and reliable assessment of learning. It is also uncertain whether current analytical methodologies are up to the task, or whether stakeholders (including learners, teachers and employers) will trust the results. Because of the technical complexity often associated with use of process data, including complex algorithms and data transformation, there are inevitable questions about whether these data have utility for assessment. The impact on learners of inaccurate, unreliable and invalid assessments of general capabilities should not be underestimated in the digital, automated, self-regulated world. Feedback and reporting generate powerful, real-life consequences for learners, which can be positive, but which can also be negative or destructive (Hattie and Timperley 2007). This concern is likely to be acute when assessments are automated to shape the treatment a person receives, and especially if there is no obvious appeal to an independent person who is familiar with the work of the assessed.

Metrolytic standards provide a framework for building confidence in assessments that use process data. It might be that the expected level of rigour inherent in the standards is too great to be feasible. The standards embody quality requirements for reliability or validity, accuracy, or interpretability of the sort normally only applied to high stakes, scaled assessments such as PISA, or SAT or GMAT. However, the application of metrolytic standards is one option available to those working at this assessment frontier, combining the methodological strengths of learning analytics and measurement science, and underpinned by a solid understanding of assessment and its role in learning.

References

Asia-Pacific Education Research Institutes Network. (2015). *Regional study on transversal competencies in education policy and practice*. Bangkok/Paris: UNESCO. Retrieved from: http://www.unesdoc.unesco.org/images/0023/002319/231907E.pdf.

Bergner, Y., Lang, C., & Gray, G. (2017). Measurement and its uses in learning analytics. In C. Lang, G. Siemens, A. Wise & D. Gasevic (Eds.), *Handbook of learning analytics* (pp. 35–48). New York: SOLAR. https://doi.org/10.18608/hla17. Retrieved from: https://www.solaresearch.org/hla-17/.

Bransford, J.D., Brown, J.D., & Cocking, R.R. (2003). *How people learn: Brain, mind, experience, and school* (Expanded ed.). Washington DC: National Academy Press.. Retrieved from http://www.nap.edu/read/9853/chapter/1

Buckingham Shum, S., & Deakin Crick, R. (2016). Learning analytics for 21st century competencies. *Journal of Learning Analytics, 3*(2), 6–21.

Carmean, C., & Mizzi, P. (2010). The case for nudge analytics. *Educause Quarterly, 33*(4), 4.

Cleary, T. J., Callan, G. L., & Zimmerman, B. J. (2012). Assessing self-regulation as a cyclical, context-specific phenomenon: Overview and analysis of SLR microanalytic protocols. *Educational Research International., 2012*, 1–19. https://doi.org/10.1155/2012/428639.

Corrin, L., & de Barba, P. (2014). Exploring students' interpretation of feedback delivered through learning analytics dashboards. In B. Hegarty, J. McDonald, & S.-K. Loke (Eds.), *Rhetoric and reality: Critical perspectives on educational technology. Proceedings ASCILITE Dunedin 2014* (pp. 629–633). Dunedin: ASCILITE.

Cronbach, L., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Dreyfus, S. E., & Dreyfus, H. L. (1980). *A five-stage model of the mental activities involved in directed skill acquisition*. Retrieved from http://www.dtic.mil/get-tr-doc/pdf?AD=ADA084551

Dringus, L. P. (2012). Learning analytics considered harmful. *Journal of Asynchronous Learning Networks, 16*(3), 87–100.

Ferguson, R., & Clow, D. (2017). Where is the evidence? A call to action for learning analytics. In *Proceedings of the 7th International learning analytics & knowledge conference*. Vancouver**,** Canada.

Fiore, S. M., Graesser, A. G., Greiff, S., Griffin, P. G., Gong, B., Kyllonen, P., et al. (2017). *Collaborative problem solving: Considerations for the national assessment of educational progress*. Alexandria: National Center for Education Statistics. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/researchcenter/collaborative_problem_solving.pdf.

Gasevic, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends, 59*(1), 64–71.

Glaser, R. (1994a). Criterion-referenced tests: Part I: Origins. *Educational Measurement: Issues and Practice, 13*(4), 9–11.

Glaser, R. (1994b). Criterion-referenced tests: Part II: Unfinished business. *Educational Measurement: Issues and Practice, 13*(4), 27–30.

Greller, W., & Draschler, H. (2012). Translating learning into numbers: A framework for learning analytics. *Educational Technology and Society, 15*(3), 42–47.

Griffin, P. (2007). The comfort of competence and the uncertainty of assessment. *Studies in Educational Evaluation, 33*(1), 87–99.

Griffin, P., & Care, E. (2015). *Assessment and teaching of 21st century skills: Methods and approach* (Vol. 2). Dordrecht: Springer.

Gunnarsson, B. L., & Alterman, R. (2013). *Understanding promotions in a case study of student blogging*. Paper presented at the 3rd international conference on learning analytics and knowledge, Leuven, Belgium.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487.

He, J., Rubinstein, B. I. P., Bailey, J., Zhang, R., Milligan, S., & Chan, J. (2016). *MOOCs meet measurement theory: A topic-modelling approach*. Paper presented at the 30th AAAI conference on artificial intelligence (AAAI-16), Phoenix, Arizona.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Masters, G. (2018, July 31). But can we measure it. *Teacher Magazine*. Retrieved from https://www.teachermagazine.com.au/columnists/geoff-masters/but-can-we-measure-it

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5–8.

Milligan, S. K., & Griffin, P. (2016). Understanding learning and learning design in MOOCs: A measurement-based interpretation. *Journal of Learning Analytics, 3*(2), 88–115. https://doi.org/10.18608/jla.2016.32.5.

Milligan, S. K., Kennedy, G., & Israel, D. (2018). *Assessment, credentialling and recognition in the digital era: Recent developments in a fertile field* (Seminar series 272). Melbourne: Centre for Strategic Education.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues & Practice, 25*(4), 6–20. https://doi.org/10.1111/j.1745-3992.2006.00075.x.

OECD. (2018). *The future of education and skills: Education 2030*. Retrieved from http://www.oecd.org/education/2030/E2030%20Position%20Paper%20(05.04.2018).pdf

Pellegrino, J.W. (1999). The evolution of educational assessment: Considering the past and imagining the future. William H. Angoff Memorial Lecture. Retrieved from https://www.ets.org/Media/Research/pdf/PICANG6.pdf

Perkins, D. N., & Salomon, G. (1992). Transfer of learning. In T. Husen & T. Neville Postlethwaite (Eds.), *International encyclopedia of education* (2nd ed.). Oxford: Pergamon Press.

Polyak, S. T., von Davier, A., & Peterschmidt, K. (2017). Analyzing game-based collaborative problem solving with computational psychometrics. In *Proceedings of 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Halifax, ACTNext.

Scardamalia, M., Bransford, J., Kozma, B., & Quellmalz, E. (2013). New assessments and environments for knowledge building. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (Vol. 1, pp. 231–300). New York: Springer.

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

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review, 46*(5), 30–32.

Tai, J., & Adachi, C. (this volume). The future of self and peer assessment: Are technology or people the key? In M. Bearman, P. Dawson, J. Tai, R. Ajjawi, & D. Boud (Eds.), *Reimagining assessment in a digital world (chapter 15)*. Dordrecht: Springer.

Tremblay, K., Lalancette, D., & Roseveare, D. (2012). Assessment of higher education learning outcomes. *Feasibility study report, volume 1 – Design and implementation*. OECD. Retrieved from http://www.oecd.org/edu/skills-beyond-school/AHELOFSReportVolume1.pdf

von Davier, A. A., & Halpin, P. F. (2013). *Collaborative problem-solving and the assessment of cognitive skills: Psychometric considerations*. ETS: Princeton.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. New York: Taylor & Francis Group.

Wilson, M., & Scalise, K. (2016). Learning analytics: Negotiating the intersection of measurement technology and information technology. In J. M. Spector, B. B. Lockee, & M. D. Childress (Eds.), *Learning, design, and technology*. New York: Springer.

Wilson, M., Scalise, K., & Gochyyev, P. (2016). Assessment of learning in digital interactive social networks: A learning analytics approach. *Online Learning, 20*(2), 97–119.

Wolfe, E. W., & Smith, E. V., Jr. (2007). Instrument development tools and activities for measure validation using Rasch models: Part 1 & part 2. In E. V. Smith Jr. & R. M. Smith (Eds.), *Rasch measurement: Advanced and specialized applications* (pp. 202–290). Minnesota: JAM Press.

World Economic Forum. (2015). *New vision for education: Unlocking the potential of technology*. Geneva: WEF & The Boston Consulting Group.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: Mesa Press.

Sandra Milligan is Director of the Assessment Research Centre at the University of Melbourne. Sandra's current research focuses on assessment, micro-credentialing and warranting of hard-to-assess curriculum or competency areas, including use of 'big data', learning analytics and developmental assessment. Sandra has an unusually wide engagement with the education industry, and in educational research. Originally a teacher of science and mathematics, she is also a former Director of Curriculum in an Australian state education department, and has held senior research, management and governance positions in a range of educational organisations, including government agencies, not-for-profits, start-up businesses and large, listed, international corporations.