Learning Analytics: Negotiating the Intersection of Measurement Technology and Information Technology

Mark Wilson and Kathleen Scalise

Contents

Abstract

In this chapter, we will review the current state of play in the area of overlap between learning analytics (LA), specifically data mining and exploratory analytics, and the field of measurement science. We will review the logic of measurement science, as instantiated through the BEAR Assessment System (BAS), and illustrate it in the context of a LA example. An example is presented showing how complex digital assessments can be designed through BAS with attention to measurement science, while LA approaches can help to score some of the complex digital artifacts embedded in the design. With that background, we

M. Wilson (✉)
University of California, Berkeley, CA, USA
e-mail: markw@berkeley.edu

K. Scalise (✉)
University of Oregon, Eugene, OR, USA
e-mail: kscalise@uoregon.edu

J.M. Spector et al. (eds.), *Learning, Design, and Technology*,
DOI 10.1007/978-3-319-17727-4_44-1

suggest ways that the two approaches can be seen to support and complement one another, leading to a larger perspective. This chapter concludes with a discussion of the implications of this emerging intersection and a survey of possible next steps.

Keywords
Learning analytics • Data mining • Measurement science • BEAR Assessment System • Twenty-first-century skills • ATC21S

Introduction: Why There Is a Problem

By some accounts, measurement is defined as the assignment of numbers to categories of observations. The properties of numbers then become the properties of a measure – such as nominal, ordinal, interval, and ratio (Stevens, 1946). But assigning numbers to categories is just one feature of measurement. Steps in measurement science before and after provide a key interpretive context upon which modern measures are based.

This chapter is about those steps, and how they are necessary when engaging in exploratory learning analytics, if the goal is to measure. Generating patterns from data sets through machine learning, for instance, can yield a set of results, i.e., those specific patterns. But what do these results mean and how can they be used to measure some underlying variable? This is where the measurement comes in.

In educational assessment, for instance, values on many variables for a student or a teacher are not manifest, or in other words, they cannot be directly measured such as one might measure height or eye color. Rather, in learning performances, a set of evidence is gathered on a "latent" construct. The property of latency means the element to be measured remains hidden to the observer, until circumstances are constructed suitable for the manifestation or elicitation of the evidence for the construct. This elicitation maps back to the construct and forward to interpretation. Together, these interpretive elements can make the numbers meaningful.

Our thesis is that to believe that the analytics themselves – the assignment of numbers – is the only or even the main goal of measurement is to miss the point. We argue that measures take on coherent evidentiary properties, such as validity, utility, and inferential characteristics, only when the numbers clearly map to a construct, to its indicators (observations), and to the interpretations around which the claims are to be made.

Applying learning analytics to educational assessment therefore requires negotiating a key intersection, which is at the interface of measurement technology and information technology. Here we first discuss both sides of that intersection, starting with principles of measurement science and moving to learning analytics. Then we share an example of opportunities that arise when the intersection is successfully navigated from both sides.

The Logic of Measurement

Some may say the IT sphere needs to expand its consciousness about measurement. It is also likely true that traditional measurement science may need to expand its consciousness about what can, these days, reasonably comprise a data set with sufficient construct-relevant variance for some measurement claims, albeit the data structure may be much more complex and incorporate a great deal of more noise and extraneous elements than one might have attempted to analyze in the past. Information technology has made great leaps forward in collecting such data, which will be discussed in the next section.

First, it is important to understand the logic of measurement. Here we use a framework that includes four principles of good assessment and measurement practice. The framework is part of the BEAR (BEAR Center = Berkeley Evaluation and Assessment Research Center) Assessment System (BAS: Wilson, 2005), which describes techniques used in the construction of high-quality assessments. A diagram of BAS is shown in Fig. 1. (These four principles also relate to the assessment triangle developed by the National Research Council Committee on the Foundations of Assessment and published in their report, “Knowing What Students Know” (2001).) We begin with a description of the four principles (Wilson, 2005), in the context of technology-enhanced assessments and learning (Scalise et al., 2007):

- Principle 1: Assessments should be based on a developmental perspective of student learning.
- Principle 2: Assessments in learning should be clearly aligned with the goals of instruction.

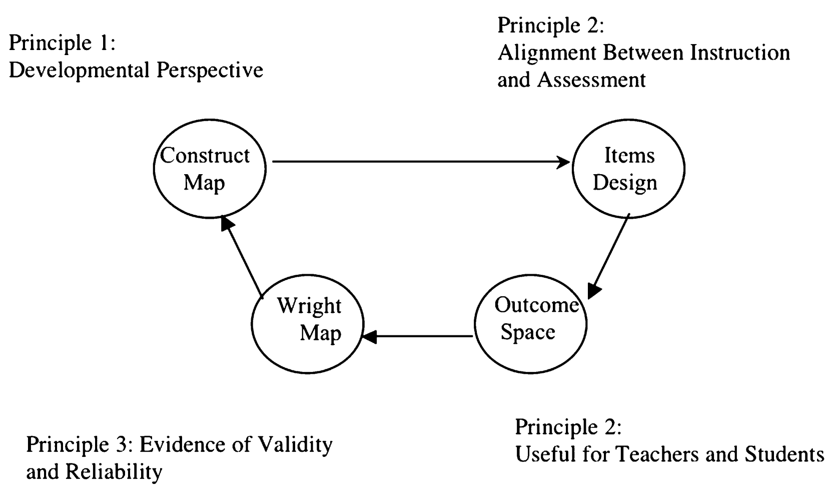


Fig. 1 A diagram of BAS, shown both the principles and four building blocks of measurement (Wilson & Sloane, 2000)

- Principle 3: Assessments must produce valid and reliable evidence of what students know and can do.
- Principle 4: Assessment data should provide information that is useful to teachers and students to improve learning outcomes.

Principle 1, a developmental perspective of student learning, means that we should be considering how student understanding of particular concepts and skills develops over time, rather than taking a one-shot view. A developmental perspective requires clear definitions of what students are expected to learn at particular points in their development, as well as a theoretical framework of how that learning is expected to unfold as the student progresses through the instructional material.

Traditional classroom assessment strongly supports a developmental perspective. Here, we affirm what is perhaps the obvious: For diagnostic information to be diagnostic, it must be collected in relationship to some set of goals about what is to be learned. Principle 2, establishing a good match between what is taught and what is assessed, means that the goals of learning and the measurements and inferences made regarding learning should be related. Reports abound of teachers interrupting their regular curricular materials to "teach the material" students will encounter on district- or statewide tests, and this is the antithesis of Principle 2.

Resnick and Resnick (1992) argued that "Assessments must be designed so that when teachers do the natural thing – that is, prepare their students to perform well – they will exercise the kinds of abilities and develop the kinds of skill and knowledge that are the real goals of educational reform" (pp. 37–76). Diagnostic assessment approaches that do not match the goals of instruction fail this test.

Principle 3, quality evidence, addresses issues of technical quality in assessments. By making inferences about students that are supported by evidence for their validity and reliability, numerous technology-enhanced learning assessment procedures are gaining "currency" in the educational community. Reliability concerns the reproducibility of results, whereas validity relates to whether an assessment measures what it is intended to measure. To ensure comparability of results across time and context, these issues must be addressed in any serious attempt at technology-based measures.

Principle 4, the value of assessment data to teachers and students, is perhaps the most critical: Learning assessment systems must provide information and approaches that are useful for improving learning outcomes. Teachers must have the tools to use systems efficiently and to explain resulting data and make inferences effectively and appropriately. Students also should be able to participate in the assessment process, and they should be encouraged to develop essential metacognitive skills that will further the learning process. If teachers and students are to be held accountable for performance, they need a good understanding of what students are expected to learn and of what counts as adequate evidence of student learning. Teachers are then in a better position, and a more central and responsible position, for presenting, explaining, analyzing, and defending their students' performances and outcomes of their instruction.

Students are better able to develop their own metacognitive skills and to bring them to bear in the learning process. In addition, learning assessment procedures should be accessible to teachers to avoid a climate of "black box" assessment, in which the logic of the assessments and personalization are known only to the software developers.

These four principles introduce a way to understand the advantages and disadvantages of measurement instruments, how to use such instruments, and how to apply these methods to develop new instruments or adapt old ones (Wilson, 2005). The four principles relate to four "building blocks" that make up an assessment – the construct map, the design plan for the items, the outcome space, and the statistical measurement model or algorithms to be used to compile and analyze patterns in the data, which can also be seen in Fig. 1.

They also focus our attention on quality control (QC) in the measures. With an interpretive context such as described by the BAS principles, QC of the measurement properties can rely heavily on the calibrated construct map and review how to check if scores are operating consistently and how to evaluate the reliability and validity evidence. This allows the assessment developer to employ a wide variety of item formats, including traditional questions in selected and constructed response formats but also behavioral observations, performance tasks, projects, portfolios, interview protocols, and active process data such as chat streams or click data in technology-enhanced assessments, when each of these forms of evidence is clearly designed to elicit observations mapped to be meaningful on the construct.

Relating This Logic to Data Mining and Exploratory Analytics

A commonly used definition of learning analytics that we will draw on here was proposed by the first International Conference on Learning Analytics and Knowledge (LAK 2011) and adopted by the Society for Learning Analytics Research (Society for Learning Analytics Research, 2011):

> Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs.

While this definition is helpful, two additional aspects are important to consider: the interpretation of results and the choice of appropriate data types and algorithms. We must underscore the point that, for learning analytics, it is critical to consider the meaningful interpretation of the data analysis, not simply reporting of the results (Wilson, 2005; Wilson et al., 2012; Wilson, Scalise, & Gochyyev, 2015).

Yet, interpretation is not directly included in the LAK/SoLAR definition of "collection, analysis, and reporting." This weakness in the definition can lead to the assumption that once results are composed and reported, their meaning for learners and learning outcomes is self-evident.

Meaningful interpretation means having an evidentiary framework, such as described in the four measurement principles above (Wilson, 2005). It must be designed to connect results clearly and on an empirical basis back to the goals and objectives of the analysis in order to make clear evidentiary claims about the learner (Mislevy, Almond, & Lukas, 2003; Wilson & Sloane, 2000). It also means being able to understand the uncertainty or range and degree of error likely to be present in the results.

Some groups have begun to establish standards of practice in learning analytics for twenty-first-century complex data analysis methodologies (Sclater, 2014; Wilson et al., 2012). In this chapter, we will present an example that helps establish the coherent evidentiary argument for the learning analytics involved through a framework called a “learning progression.” This framework connects the results to (a) the data and the learning analytic questions being asked and (b) to the techniques for the analytics employed.

Other researchers have begun to describe the need for such frameworks when learning analytics goes beyond data analysis alone and is to be used for predictive analytics, actionable intelligence, and decision-making (van Barneveld, Arnold, & Campbell, 2012).

In learning analytics, the need to establish a coherent evidentiary argument to support claims about learners can be approached either a priori (in advance of the analysis) or a posteriori (following the analysis). The a priori approach is essentially a theoretical approach, based on a strong theory or prior empirical information (or both), and thus might be considered a confirmatory learning analytic technique. It is also sometimes known as “supervised learning” (Russell & Norvig, 2009), in which factors, weights, network structures, or other characteristics of the LA learning algorithms are populated in advance with at least some prescribed characteristics, derived from prior work or from a theoretical basis.

The a posteriori approach can be considered generative or in other words an exploratory learning analytic approach and in many cases will need to be confirmed by a subsequent data collection and analysis. The exploratory approach is sometimes called by the name “data mining” (Papamitsiou & Economides, 2014) or machine learning. It can also be known sometimes as “unsupervised learning” (Russell & Norvig, 2009), in contrast to the supervised learning concept described above in which models are prepopulated to some extent with theoretical or prior empirical data. Exploratory approaches can be useful when the desire is to learn more about the patterns in the data sets in a context where little is yet understood or where new patterns may become evident that were not suspected before.

The entry point into the learning analytics paradigm then is an option to be considered when building the evidentiary argument to make claims about learners, the choice between an exploratory and confirmatory approach, depending on how much prior theory and/or empirics are available. Put together, these exploratory and confirmatory stages can be seen as a cycle in the evidence chain, as shown in Fig. 2. It depicts a simple example of a learning analytics interpretive cycle, where entry points can be either confirmatory, entering at the “theory or conceptualization” node, or exploratory, entering at “analysis and results” node for extant data or at the

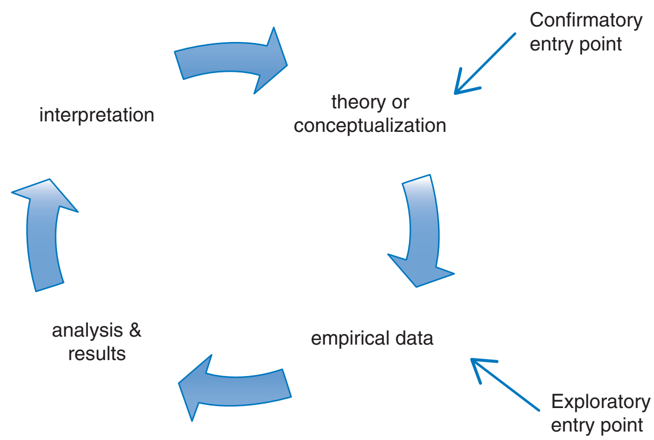


Fig. 2 Exploratory and confirmatory evidence chain cycle (Wilson, Scalise, & Gochyyev, pending)

"empirical data" node when observations will be designed and collected (see below for a discussion of extant and collected data).

No single entry point to the cycle is better in every situation. The choice can be determined by the intended *purposes* of interpretation and the current *state of claims* that can be made in a given context. In any particular situation, one relevant question to ask is does the analysis begin with an interpretive framework a priori, as in the theory component of the cycle below, or is interpretation intended to fall a posteriori, as when even the initial interpretive framework is derived from data because little is yet known?

In either case, the same cycle is present but with different points of entry and a different flow to the interacting elements.

Measurement science, then, really encompasses the "whole" thing or all of the cycle. But it doesn't truly become measurement until we have some of the confirmatory evidence that meets high-quality measurement standards – in other words exploring isn't enough to claim measurement. In this way, measurement can be seen in some contexts as a qualitative and quantitative cycle or an exploratory and confirmatory cycle. Just as for learning analytics, measurement science can be entered into at different points for a given construct depending on how generative or emerging of new theory the goal of the measures involves.

In terms of data types for which learning analytics by the LAK/SoLAR definition is likely to be more useful, in most cases, complex data should be involved. If not, other simpler techniques might be better employed (Ferguson, 2012). Complex data can take the form of large data sets (big data), multifaceted data sets, or other elements in the data that encode more complex patterns (Wilson et al., 2012) or, as described by Kathleen Scalise, hard-to-measure constructs not readily identifiable without complex analytic techniques (2012).

About the data sets, sometimes these can be preexisting or extant data sets, as described above. Examples of preexisting data include downloads from Twitter feeds, click streams in user data, or other online collections that often exist for another purpose originally (Baker & Siemens, 2014). At other times, data sets are collected at least in part directly for the purpose of applying learning analytics to the results. Data collection can include, for instance, an adaptive recommender where ratings on prior experiences are solicited for the purposes of prediction of respondent interest in future experiences (Chedrawy & Abidi, 2006; Dagger, Wade, & Conlan, 2005), or evidentiary data collection for educational or professional development, to address personalized or grouped components to support the learner in educational assessment (Brady, Conlan, Wade, & Dagger, 2006; Kennedy & Draney, 2006).

An extension to the LAK/SoLAR definition we propose here is specification that complex analytic techniques are needed to resolve the multifaceted or complex patterns. The same argument can be made as above for data sets. Complexity should be introduced in the analysis for a coherent evidentiary argument only when necessary. For instance, if a simple model that compares dyads of words in a narrative text stream produces results as good as comparing longer strings or more complicated relationships, the simpler model is supported (Russell & Norvig, 2009). So the usual parsimonious definition should be applied when models or other algorithms are used to fit learner data and resolve patterns.

Finally, it would be helpful if the LAK/SoLAR definition made reference to algorithms, or characteristics of algorithms, that might be to useful to apply for aggregating and parsing of patterns, since this is an important consideration in the use of learning analytics (Papamitsiou & Economides, 2014). While it is important to keep the definition general to be inclusive of many useful algorithms that might arise, as a general class, the approach typically needs to involve algorithms to automatically process the data, assuming the purposes of interpretation and the complexity of data require algorithmic approaches to the accumulation and parsing of patterns in the data. Algorithms can be statistical in nature, applied as inferential statistical tests or to yield inferential indices as part of the processing, which can help with assessing quality of results (Sclater, 2014).

Numerous algorithms in the form of measurement models have been created and applied that take a statistical form for learning outcomes. These are well established in the psychometrics research literature, and some of the advanced models as well as basic models can be appropriate to apply in learning analytics to complex twenty-first-century skill settings (Wilson et al., 2012).

Algorithms can also process patterns in more descriptive ways, yielding machine-readable results such as categorization or subsetting of respondents (Stanton, 2012). Note that since machine processing is required, however, the data sets at some point have to include machine-readable data. This may be text based or graphical in nature or in some other innovative format, depending on the processing requirements of the algorithm and platform, or the data sets may be numeric (Scalise & Gifford, 2006). The desired data characteristics may already be present for a given data set in any particular case or may require preprocessing. This could include types of scoring, ordering, subsetting, or other types of aggregation. For this, reliable data collection, warehousing, and prep can be a problem, so a variety of "cleanup" procedures may be needed. An important stage in learning analytics is reducing construct irrelevant variance including noise, user errors, or out-of-scope entry of data, which should be clarified and validated before conclusions can be drawn (Dringus, 2012).

In light of these sets of clarifications, we have suggested a revision to the LAK/SoLAR definition, which we propose as "Learning analytics definition, LAK/SoLAR.v2" (Wilson et al., pending):

> Learning analytics is the measurement, collection, analysis, *interpretation*, and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs, by means of a coherent evidentiary argument. Complexity should be introduced in the data and the analysis only when necessary to the development of the evidentiary argument.

Complex data will almost always be involved, which can take the form of large data sets (big data), multifaceted data sets, and/or other the data elements that encode patterns or hard-to-measure constructs not readily identifiable without advanced analytic techniques (Russell & Norvig, 2009; Scalise, 2012).

An Example: Synthesis of Measurement Technology and Learning Analytics

We began this chapter by describing a key intersection for learning analytics at the interface of measurement technology and information technology. But little has been said yet about information technology directly. As it turns out, much of the modern field of learning analytics has derived from the efforts of information technologists to successfully tackle complexity in data and analysis.

For instance, scalable machine learning for large data sets may take place using programming scripts in proprietary software or in more open-source solutions such as R and Python. A variety of distributed machine learning platforms are available designed for big data that now readily can run on a laptop or even tablet device. Core machine learning algorithms may be implemented in high-performance programming languages, with acceptable APIs (application programming interfaces) for interoperability through web interfaces.

Algorithm implementations may be distributed across virtual servers. This can permit software and analysis to scale for big data sets. When enough computational resources are available, a variety of algorithms may be employed, as discussed in other chapters of this handbook, from generalized linear models to gradient boosting and deep neural nets to dimensionality reduction methods (PCA, GLRM) and clustering algorithms (K-means). Anomaly detection, for instance, is becoming important to detect false positives and to improve the feedback within the state characteristics.

All of these are examples of advances not only in statistics in many cases but also in information technology. Scientists and engineers have many more solutions to offer due to the advances being made in IT.

That said, recall that the thesis of this chapter is that to believe the analytics themselves – the assignment of numbers, categories, or other sophisticated quantification or classification – is the only or even the main goal of measurement is to miss the point of measurement science. Measures take on coherent evidentiary properties, such as validity, utility, and inferential characteristics, only when the numbers clearly map to a construct, to its indicators (observations), and to the interpretations around which the claims are to be made.

So here we arrive at the intersection of measurement technology and information technology. In the context of applying learning analytics, for instance, to educational assessment, can the two perspectives work together to achieve a *gestalt* or more than the sum of the parts? Next we take up a brief example that attempted to incorporate both together, in the context of looking at the assessment of collaborative learning in digital interactive social networks.

The example here is taken from the Assessment and Teaching of Twenty-First Century Skills project (ATC21S), which was launched in 2009 by three information technology companies, Cisco, Intel, and Microsoft. An ATC21S project goal was to employ new analytical approaches in the assessment of learning.

For the ATC21S example, the BEAR Assessment System was applied to identify a set of distinctive information and communication technology (ICT) literacy goals for students (BAS: NRC, 2001; Wilson, 2005, 2009; Wilson & Sloane, 2001). The focus of ICT literacy was on collaborative digital activities, or *learning in networks*, which was seen as being made up of four strands of a learning progression:

- Functioning as a consumer in networks
- Functioning as a producer in networks
- Participating in the development of social capital through networks
- Participating in intellectual capital (i.e., collective intelligence) in networks

The four strands are seen as interacting together in the activity of learning in networks. They are conceptualized as parallel developments that are interconnected and make up that part of ICT literacy that is concerned with learning in networks.

First, functioning as a consumer in networks (CiN) involves obtaining, managing, and utilizing information and knowledge from shared digital resources and experts in order to benefit private and professional lives. It involves questions such as:

- Will a user be able to ascertain how to perform tasks (e.g., by exploration of the interface) without explicit instruction?
- How efficiently does an experienced user use a device, application, or other ICT strategy to find answers to a question?
- What arrangement of information on a display yields more effective visual search?
- How difficult will it be for a user to find information on a website?

Second, functioning as a producer in networks (PiN) involves creating, developing, organizing, and reorganizing information/knowledge in order to contribute to shared digital resources.

Third, developing and sustaining social capital through networks (SCN) involves using, developing, moderating, leading, and brokering the connectivities within and between individuals and social groups in order to marshal collaborative action, build communities, maintain an awareness of opportunities, and integrate diverse perspectives at community, societal, and global levels.

Fourth, developing and sustaining intellectual capital through networks (ICN) involves understanding how tools, media, and social networks operate and using appropriate techniques through these resources to build collective intelligence and integrate new insights into personal understandings.

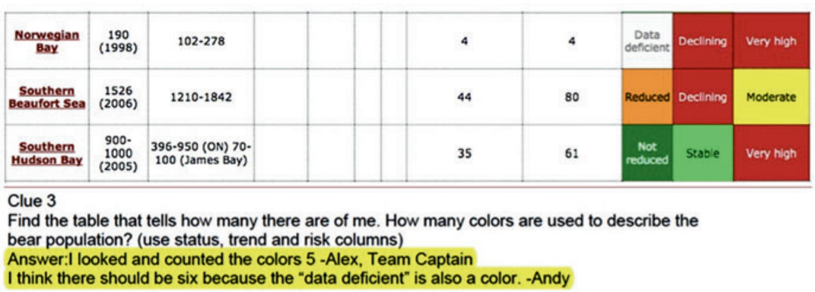
Using the four principles described above, assessments were designed to clearly align with learning goals through these constructs, to produce valid and reliable evidence of what students know and can do for the development perspective, and to generate evidence useful to teachers and students.

One potential mechanism to achieve these goals is to model assessment practice through a set of exemplary classroom materials. The example module here was developed based on using some of the "Go North!" expedition findings, originally posted as a K-12 virtual project by the University of Minnesota and partners. While no materials from the site were actually brought into the assessments shown here, students were allowed to use navigation links available through their browsers to view some of the scientific expedition materials. This was an example of using publicly available online resources to access a range of rich materials in the classroom. The website was developed at the University of Minnesota in collaboration with NOMADS Online Classroom Expeditions, GoNorth! This online adventure learning project was based around arctic environmental expeditions. The website was a learning hub with a broad range of information and different mechanisms to support networking with students, teachers, and experts.

ICT literacy resources developed relating to this module focus mainly on the functioning as a consumer in networks strand. The tour through the site for the ATC21S demonstration scenario is conceived as a "collaboration contest" or virtual treasure hunt (see Fig. 3 for a sample screen). The Arctic Trek scenario views social networks through ICT as an aggregation of different tools, resources, and people that together build community in areas of interest. In this task, students in small teams ponder tools and approaches to unravel clues through the Go North site, via touring scientific and mathematic expeditions of actual scientists.

The Arctic Trek task in which students work in teams is demonstrated in Figs. 4 and 5. In that task, students are expected to find the colors that are used to describe the bear population in the table, part of which is shown at the top. The highlighted chat log of students at the bottom of the figure, which actually takes the form of a collaborative laboratory notebook, indicates that students are communicating in order to identify signal versus noise in the supplied information. The colors in the text are the colors shown in the columns on the right of the table. Requiring both identifying signal versus noise in information and interrogating data for meaning,

Fig. 3 ATC21S Arctic Trek math and science task opening screen



Norwegian Bay	190 (1998)	102-278					4	4	Data deficient	Declining	Very high
Southern Beaufort Sea	1526 (2006)	1210-1842					44	80	Reduced	Declining	Moderate
Southern Hudson Bay	900-1000 (2005)	396-950 (ON) 70-100 (James Bay)					35	61	Not reduced	Stable	Very high

Clue 3
Find the table that tells how many there are of me. How many colors are used to describe the bear population? (use status, trend and risk columns)
Answer:I looked and counted the colors 5 -Alex, Team Captain
I think there should be six because the "data deficient" is also a color. -Andy

Fig. 4 Example of student collaborative chat in Arctic Trek task

this performance can be mapped into the ICN3 level ("Proficient builder") of the ICN strand (Wilson & Scalise, 2014). For further examples of activities and items from the Arctic Trek scenario, see Wilson and Scalise (2014).

For this example, the connection at the intersection of measurement science and learning analytics can be made in two ways. First, the statistical analytic technique used to compile scores in measurement science is called a "measurement model." It serves as an algorithm to gather the results together and make inferences about learners. Other fields such as computer science that come to learning analytics from

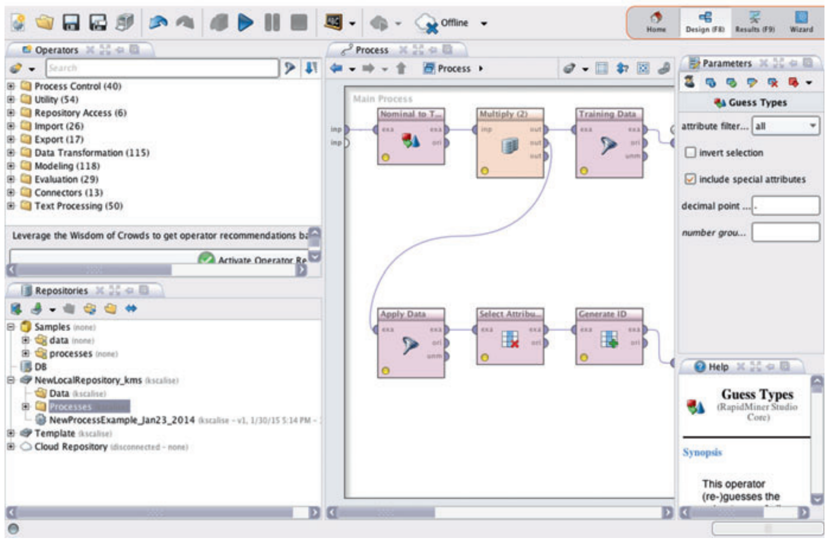


Fig. 5 Sentiment analysis design window for ATC21S example

a different historical basis often use a different vocabulary to describe such algorithms. For instance, the Rasch model often used in educational assessment from a computer science perspective would be considered an LA algorithm employing a multilayer feed-forward network (Russell & Norvig, 2009) with *g* as the Rasch function (a semi-linear or sigmoidal curve-fitting function), in which weights (item discrimination) are constrained to one for all inputs, and the item parameters estimated are only the thresholds on each item node (item difficulty). The 2PL IRT model, by contrast, is an algorithm employing a multilayer feed-forward network with *g* as the 2PL function (also a sigmoidal curve-fitting function), in which both weights (item discrimination) and thresholds on each item node (item difficulty) are estimated. In a further example of a commonly used measurement model, the 3PL model is an algorithm employing a multilayer feed-forward network with *g* as the 3PL function (sigmoidal), in which weights (item discrimination), thresholds on each item node (item difficulty), and a lower asymptote (guessing parameter) are estimated.

Secondly, the point we want to illustrate in this chapter is that additional specific learning analytics tools can be added or embedded within the traditional measurement model. Here we show an example of such embedding through an automated scoring engine. Scores produced by a scoring engine can be incorporated into a data set to be treated by a measurement model. To exemplify this, some of the complex student work products from the Arctic Trek module were treated under a learning analytics approach called "sentiment analysis." This involves predictions of team success in the collaborative notebooks.

For this example, some notebooks to be used for a training set of the LA engine were first scored by handscoring, using traditional tools such as rubrics and exemplars. A set of 28 handscored notebooks, which were work products for approximately 112 students, provided this training set. The training set was then made available to RapidMiner (Hoffman & Klinkenberg 2013) for the LA sentiment analysis approach.

Sentiment analysis in RapidMiner is an LA technique intended to extract information from large full-text data sources such as online reviews and social media discussions. It is often used to interpret and optimize what is being thought, said, or discussed about a company or its products – or in this case, what is being discussed in a collaborative learning situation for a science and mathematics learning activity.

The basic approach in sentiment analysis is to classify an expressed opinion in a document, a sentence, or an entity feature as positive or negative. In this case, "positive" means that the notebook shows some good evidence of learning in networks, based on the construct ideas described above. To calibrate the engine, first, both positive and negative "reviews" of the task results are considered – or in other words, a training set of scored collaborative notebooks are provided to the engine.

For the engine, first all of the words are stemmed into root words. Then, a vector word list and a model are created. Using the training set, the model compares each word in the given notebook being considered with that of words that come under different predictions stored earlier. The notebook prediction is estimated based on the majority of words that occur under a polarity or a trend direction toward a negative or positive prediction. In this way, sentiment analysis is a "bag of words" artificial intelligence technique (Russell & Norvig, 2009). More sophistication can be added to the sentiment analysis data mining engine to include a variety of relationships between words, if desired, and data adjustments such as spelling corrections, "blacklists," and "whitelists" that are addendums or eliminations from the data dictionary, and so forth. Here, an example of the sentiment analysis design window shows in Fig. 5.

The components of the full analysis for the Arctic Trek sentiment analysis engine used here are shown in Fig. 6.

For this project, following the establishment of the training set, four additional collaborative notebooks were added to the work product data set for the sentiment analysis. These additional notebooks did not associate a prediction for the sentiment analysis a priori. Rather the goal was for the LA engine to generate the prediction for each of the four notebooks. However, the four notebooks were handscored in advance using the same human scoring approaches as for the other notebooks. The point was to see if the LA engine could match and even potentially add to the results generated by the handscoring.

If so, this would provide some evidence that an LA sentiment analysis engine (in this case, via RapidMiner) might effectively be incorporated into the measurement science approach here. This could help to satisfy the measurement principle of usability by teachers and students, since an effective LA engine might eliminate some of the extensive handscoring. Then, use of the complex and interesting

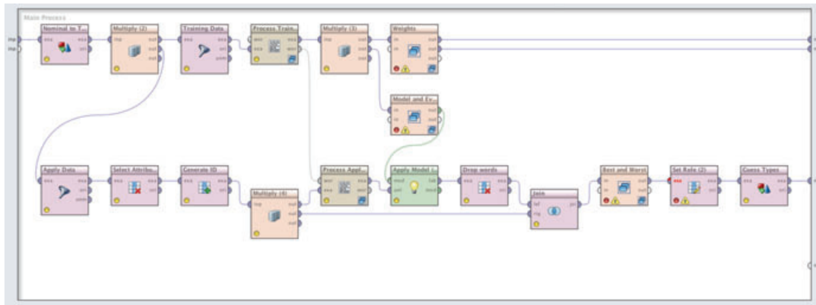


Fig. 6 Sentiment analysis component elements for LA engine in Arctic Trek

learning activities in the classroom could be much more possible and practical for teachers and students. Using such digital assessment tasks to generate measurement evidence as well as provide an effective classroom activity might require such tools for teachers.

The four notebooks selected represented a small but purposive sample for the engine to score. Only one notebook was high scoring according to the human rating. A second low-scoring notebook illustrated a similar text complexity but without nearly as much substantively correct information and with few patterns of collaboration incorporated. Two additional notebooks represented sparser incorrect versions, with little or no evidence of effective *learning in networks* practices, based on the construct ideas described above. All notebooks were supplied to the engine in their native formats, without editing or correction for any of the attributes of the student work.

One caveat for limitations that should be noted in advance of reporting the results is that this is a very small data set for most purposes but can serve for an illustrative example, and a larger set would be needed to provide a more formal example. Thus, this example should not be considered conclusive evidence of the sentiment engine here as being effective or ineffective for such purposes. Rather it should be considered illustrative of the larger topic, the potential intersection of measurement science and learning analytics. Collaborative data sets with teams of four yield fewer unique work products than in individual assessments. A larger data set of 150–175 notebooks, or therefore about 600–900 students if composed of collaborative teams of four students per notebook, would be more desirable for training an engine. Furthermore, the reader should note that if larger live action collaborative data sets were available, other algorithms might be more desirable (Chi et al., 2008; Pirolli, 2007, 2009; Pirolli, Preece, & Shneiderman, 2010; Pirolli & Wilson, 1998).

A brief example of the results of the sentiment analysis is shown in Table 1. Results show that the LA sentiment engine in this case was able to rank the four notebooks in the same order as the handscoring did. The high-scoring notebook was rated considerably higher than the next ranked notebook, even though text complexity between the two work products was similar. Furthermore, the LA engine

Table 1 Sentiment analysis results for Arctic Trek four notebooks

Notebook ID number	Sentiment ranking (pos/neg)	RapidMiner “score”	Handscore ranking
A (original case number 32)	Positive	78.0	1 (only notebook of the four judged as high-scoring, illustrated strong elements of collaboration)
B (original case number 11)	Negative	46.0	2 (low-scoring notebook but with some beginner elements of collaboration, text complexity similar to notebook A above)
C (original case number 14)	Negative	39.0	3 ties (low-scoring notebook, few if any relevant elements of collaboration visible)
D (original case number 13)	Negative	35.0	3 ties (low-scoring notebook, few if any relevant elements of collaboration visible)

seemed also able to do a reasonable job of awarding a type of “partial credit,” establishing a score substantially higher for the top notebook, but also ranking the next notebook somewhat higher than the other two, as had been the case for the human ratings. The notes in the handscore ranking column provide some interpretive context for teachers and students and could be applied to the LA results as well and mapped to the construct information described above.

How Findings Could Improve or Inform Teaching

For many teachers, the idea of teaching twenty-first-century standards such as digital collaboration is challenging (Partnership for 21st Century Skills & American Association of Colleges of Teacher Education, 2010; Scalise, 2016; Schrum & Levin, 2014). Teachers can help students be more successful in both their tools for working and ways of working digitally, but to do so, schools must have ways, means, and opportunities to help students master working in digital collaboration (Binkley et al., 2012; Griffin, McGaw, & Care, 2012). Digital literacy skills include social and intellectual capital, which are needed for virtual collaboration when the goal is learning in networks (Wilson, Scalise, & Gochyyev, in press). Yet these goals and objectives are not yet built into most educational systems, curricular materials, or approaches that teachers learn in professional development to support student learning.

Here, helping educators understand what a successful performance looks like in a collaborative digital space is important to improve teaching, if the improvement of twenty-first-century skills such as digital collaboration for learning in social networks is a goal. Furthermore, providing tools at the intersection of measurement science and LA, as described here, helps to inform teaching so that teachers know how such skills can be effectively assessed and whether and how students should be expected to improve over time.

Instructors have considerable experience recognizing more traditional work products in the classroom, but sometimes don't know if they can effectively recognize increasing student proficiency in an area such as digital collaboration. They haven't seen many examples, and they have few assessment tools to formally support the new learning environments. Together, LA and measurement science could make large contributions to teacher efforts in supporting complex twenty-first-century skills for students. Such approaches can allow teachers to have high-quality use of evidence without reducing or impoverishing the objectives or student experience in hard-to-measure constructs (K. Scalise, 2012). Furthermore, new types of feedback and enhanced feedback can be provided (Timms, 2016; Timms, DeVelle, & Schwanter, 2015).

One key topic that teachers specifically ponder in digital collaboration is how to effectively evaluate collaborative work in an online setting (McFarlane, 2003). They often feel they are good at evaluating work products in their subject matter areas, for instance, they can "grade" and provide feedback for language, math, or science competencies in a given assignment. But what factors might they tap as indicators of growing student proficiency (Wilson et al., 2012) in collaborative online digital literacy more generally? Without some indicators, it can be difficult for teachers to gauge how they are helping students improve in this type of educational practice. Working together at the intersection of LA and measurement science can provide new ways to help improve and inform teaching. This is true especially when the learning goal or learning products are not simple or traditional.

Conclusion and Next Steps

The proceeding descriptions and example review the current state of play in the area of overlap between learning analytics (LA), specifically data mining and exploratory analytics, and the field of measurement science. The logic of measurement science was reviewed briefly, definitions for LA introduced and extended slightly, and a brief example given showing how the two approaches can support and complement one another.

Next, we summarize some thoughts on what measurement can learn from LA, what LA can learn from measurement, and what the two fields must now do together, to realize the potential of the intersection.

What measurement can learn from LA. Learning analytics has shown a fearlessness in taking advantage of the new sources and large scope of data that have become available in the digital age. As well as hugely expanding the types and volume of data available to education, this has opened entirely new possibilities that simply did not exist before, from moment-to-moment data collection in educational settings, to detailed observations of interactive settings such as one-on-one conversations and classroom discussions, to the representation as complex data of objects that were previously not available, to quantitative analysis such as syntactic and content representations of document, student products, and so forth.

But it is not only the collection of data that is being revolutionized, it is the speed and possibility of feedback that opens up significant possibilities for education. No longer do educators have to wait for the “back-room experts” to spend weeks (or months) analyzing the data and preparing reports. They can obtain virtually instantaneous feedback once the student has responded. In our judgment, it is this that holds the greatest promise. The impact of classroom assessment on student success has been well documented in a conclusive meta-analysis (Black & Wiliam, 1998). But this impact has had little to do with measurement in the past as the classroom environment was too ephemeral for the “slow and serious” pace of traditional educational measurement. Measurement has, partly by virtue of its usual funding sources (policy-level decision-makers) and partly due to the lack of appropriate technology as noted above, been focused on large-scale samples of rather slim amounts of data for each sample student. This has proven useful for administrative and program evaluation purposes, but largely skipped over the most important site of educational change and improvement.

In addition, we agree with our colleague Bob Mislevy (2016) who has explained that while early measurement scientists often had a strong domain grounding in what they were trying to measure (e.g., psychologists trying to measure psychological traits concerning which they were pioneering experts), measurement science became its own specialty, and much of the domain expertise has been lost directly by the psychometricians. In contrast, LA researchers have built strong, distributed teams that bring that expertise back into play in ways that measurement science can learn from. They can tackle much more complex work products and data streams, but only because they pay a lot of attention to having actual educational professionals and domain analysts for the given area of interest working with them closely.

What LA can learn from measurement. The discussion above provides several aspects of the strength of the measurement approach as a framework for LA. First, every time that someone interprets LA results pertaining to student performance, they are making certain assumptions. Over many years, and across a wide range of contexts, the nature of these assumptions has been considered and contested with the domain of the science of measurement. Above, we have emphasized the importance of having a scientific theory that is the basis for the interpretation of the results – the construct map in the context of the BAS (although, of course, there could be many other such bases). Equally, there needs to be an understanding of how the actual data sources relate back to this scientific theory (this was embodied in the items design and the outcome space in the BAS). And, in order to have some means to appreciate the way that the accumulated evidence might relate to the hypothesized scientific construct, it is essential to have a statistical model for estimation and for uncertainty evaluation (which is one aspect of the measurement model in the BAS).

In addition, quality control considerations need to be invoked, and these are expressed in the measurement approach through concepts such as validity and reliability evidence (e.g., AERA/APA/NCME 2014), which summarize the grounds on which one can be assured that the interpretations one would like to make of the LA results are indeed valid. No amount of data, frequency of responses, nor novelty of data format will reduce the need for these issues to be considered and responded to. Ignoring this need may be possible at the initial stages of implementation, but long practice in many different domains has told us that such willful ignorance is fraught with risk, not just for the learning analysts but also for the students and teachers who rely on them.

What LA and measurement can do together. Perhaps even more important than that the two disciplines learn from one another is that they need to work together. Our example above has been intended to show some of the complementarities that exist between the two approaches, and our principal arguments above are not based on any necessary oppositions between the two, but rather on how they can be seen to offer ways that each can extend the other.

Looking back over our chapter, we see that new research directions at the intersection of LA and measurement science have been prompted by our discussions. First, in thinking about how interactions with LA can improve and expand measurement science, we noted the following possibilities. Measurement science needs to adapt to the important new directions and possibilities that LA affords with respect to the gathering of new types of data relating to student behaviors beyond the standard measurement science formats of the test and the questionnaire/survey to incorporate not just student "answers" but also their many steps and actions toward those answers. Measurement science also needs to welcome the invigoratingly broader horizon of being able to examine the entire time of student educational experiences, not just a single event in a single classroom in a single year, but by having access to the whole range of operational data that will be available regarding students. The very size of LA data sets is also a challenge to standard measurement science – the typical techniques of statistical analysis will have to give way to more flexible and fast algorithms and means of communicating results.

Second, thinking about how interactions with measurement science can improve and expand LA, we came up with the following possibilities. One possibility will, of course, include new LA algorithms and aggregation approaches. These are likely to be situated in data density – but they will also rely on more pattern finding and likely noisier patterns, with more construct irrelevant variance, included in less structured but larger data sets. A good direction for assessing efficacious algorithms and methods of classification and feedback, specifically for educational applications, will be to search for methods that add to the explained variance of models already employed in measurement science. As LA matures to focus not only on predictive validity but also to the establishment of well-accepted procedures for quality and measurement standards, new research directions will emerge in the science of LA assessment. These include technical studies and simulations to understand and address reliability and precision information for LA, assessment form creation, linking and equating, adaptive administrations, evaluating assumptions, and checking data-model fit. Furthermore, as LA opens up more opportunities for rich assessment of hard-to-measure constructs that are instructionally relevant, the interpretive focus of LA becomes more prominent. LA will need to add perspectives and practices regarding validity evidence for the interpretations of LA results: Measurement science has had 100 years of experience in this, and it will be much more efficient for LA to learn from that than to repeat those 100 years

Thinking from both sides, an important terrain of research directions emerges related to improving and informing instruction. Research questions to be asked include how and whether teaching and feedback opportunities can enrich student learning outcomes and whether they can address that need for all students, including disadvantaged students. Technology can help to level the playing and close achievement gaps – but it can also further marginalize some populations

Thus there is a need for new R&D projects that combine the two approaches together. This must provide wide dissemination of outcomes in order to reach the widely distributed fields of application, which often do not share the same source materials. Joint publication of books that combine the approaches and synthesize approaches would be helpful. Finally, training programs are needed that combine the two, both for graduate students and for working professionals and academics.

To sum up, as we enter a new age of digitally extended data collection, we need to match the fearlessness of LA with the strength and reassurance of measurement science.

References

American Educational Research Association, American Psychological Association, National Council for Measurement in Education (AERA, APA, NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Baker, R. S., & Siemens, G. (2014). Educational data mining and learning analytics. In K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences*. Cambridge, UK: Cambridge University Press, (pp. 253–274).

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (Vol. 1). Dordrecht, The Netherands/New York, NY: Springer.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. doi:10.1080/0969595980050102.

Brady, A., Conlan, O., Wade, V., & Dagger, D. (2006). *Supporting users in creating pedagogically sound personalised learning objects.* Paper presented at the Adaptive Hypermedia and Adaptive Web-based Systems, Dublin, Ireland.

Chedrawy, Z., & Abidi, S. S. R. (2006). *An adaptive personalized recommendation strategy featuring context sensitive content adaptation.* Paper presented at the Adaptive Hypermedia and Adaptive Web-Based Systems, 4th International Conference, AH 2006, Dublin, Ireland.

Chi, E. H., Pirolli, P., Suh, B., Kittur, A., Pendleton, B., & Mytkowicz, T. (2008). *Augmented social cognition*. Palo Alto, CA: Palo Alto Research Center.

Council, N. R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Dagger, D., Wade, V., & Conlan, O. (2005). Personalisation for all: Making adaptive course composition easy. *Educational Technology & Society, 8*(3), 9–25.

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

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning, 4*(5/6), 301–317. doi:10.1504/IJTEL.2012.051816.

Gasevic, G., Dawson, C., Ferguson, S. B., Duval, E., Verbert, K., & Baker, R. S. J. D. (2011). Open learning analytics: An integrated & modularized platform (Concept paper). Society for Learning Analytics Research. Retrieved from http://solaresearch.org/OpenLearningAnalytics.pdf

Griffin, P., McGaw, B., & Care, E. (Eds.). (2012). *Assessment and teaching of 21st century skills*. Dordrecht, The Netherlands/New York, NY: Springer.

Kennedy, C. A., & Draney, K. (2006). Interpreting and using multidimensional performance data to improve learning. In X. Liu (Ed.), *Applications of Rasch measurement to science education*. Chicago, IL: JAM Press.

McFarlane, A. (2003). Assessment for the digital age. *Assessment in Education: Principles, Policy & Practice, 10,* 261–266.

Mislevy, R. J. (2016). [Discussion of learning analytics].

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to Evidence-Centered Design. *CRESST Technical Paper Series.* Los Angeles, CA: CRESST.

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society, 17*(4), 49–64.

Partnership for 21st Century Skills & American Association of Colleges of Teacher Education. (2010). 21st Century Knowledge and Skills in Educator Preparation. http://www.p21.org/storage/documents/aacte_p21_whitepaper2010.pdf

Pirolli, P. (2007). Cognitive models of human-information interaction. In F. T. Durso (Ed.), *Handbook of applied cognition* (pp. 443–470). New York, NY: Wiley.

Pirolli, P. (2009, April 3–9). *An elementary social information foraging model.* Paper presented at the CHI 2009, ACM Conference on Human Factors in Computing Systems, Boston, MA.

Pirolli, P., Preece, J., & Shneiderman, B. (2010). Cyberinfrastructure for social action on national priorities. *IEEE Computer, 43*(11), 20–21.

Pirolli, P., & Wilson, M. (1998). A theory of the measurement of knowledge content, access, and learning. *Psychological Review, 105*(1), 58–82.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 37–76). Boston, MA: Kluwer.

Russell, S., & Norvig, P. (2009). *Artificial intelligence, a modern approach* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Scalise, K. (2012). Using technology to assess hard-to-measure constructs in the CCSS and to expand accessibility. *Invitational Research Symposium on Technology Enhanced Assessments*. http://www.k12center.org/events/research_meetings/tea.html

Scalise, K. (2016). Student collaboration and school educational technology: Technology integration practices in the classroom. *Journal on School Educational Technology, 11*(4), 39–49.

Scalise, K., & Gifford, B. R. (2006). Computer-based assessment in E-Learning: A framework for constructing "Intermediate Constraint" questions and tasks for technology platforms. *Journal of Teaching, Learning and Assessment, 4*(6), 7.

Scalise, K., Bernbaum, D. J., Timms, M. J., Veeragoudar Harrell, S., Burmester, K., Kennedy, C. A., & Wilson, M. (2007). Adaptive technology for e-Learning: Principles and case studies of an emerging field. *Journal of the American Society for Information Science and Technology, 58* (14), 001–015.

Schrum, L., & Levin, B. B. (2014). *Evidence-based strategies for leading 21st century schools*. Thousand Oaks, CA: Corwin.

Sclater, N. (2014). JISC: Code of practice for learning analytics: A literature review of the ethical and legal issues. http://repository.jisc.ac.uk/5661/1/Learning_Analytics_A-_Literature_Review.pdf

Stanton, J. M. (2012). *An introduction to data science.* Retrieved from http://surface.syr.edu/istpub/165/

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 221–263.

Timms, M. (2016). Towards a model of how learners process feedback: A deeper look at learning. *Australian Journal of Education*. doi:10.1177/0004944116652912.

Timms, M., DeVelle, S., & Schwanter, U. (2015). *Towards a model of how learners process feedback.* Paper presented at the Artificial Intelligence in Education Conference 2015, Switzerland.

van Barneveld, A., Arnold, K. E., & Campbell, J. P. (2012). Analytics in higher education: Establishing a common language. *EDUCAUSE Learning Initiative*. https://qa.itap.purdue.edu/learning/docs/research/ELI3026.pdf

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Assoc.

Wilson, M., Bejar, I., Scalise, K., Templin, J., Wiliam, D., & Torres-Irribarra, D. (2012). Perspectives on methodological issues. In P. Griffin, B. McGaw B., & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 67–142). Dordrecht: Springer.

Wilson, M., Scalise, K., & Gochyyev, P. (in press). ICT Literacy – Learning in digital networks. In R. W. Lissitz & H. Jiao (Eds.), Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective. Charlotte, NC: Information Age Publisher.

Wilson, M., Scalise, K., & Gochyyev, P. (2016). Assessment of learning in digital interactive social networks: A learning analytics approach. *Online Learning Journal, 20*(2). ISSN 2472–5730. http://olj.onlinelearningconsortium.org/index.php/olj/article/view/799/205

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal for Research in Science Teaching, 46*(6), 716–730.

Wilson, M., & Scalise, K. (2014). Assessment of learning in digital networks. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills: Vol. 2. Methods & approaches*. Dordrecht: Springer.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181–208.

Wilson, M., Scalise, K., & Gochyyev, P. (2015). Rethinking ICT literacy: From computer skills to social network settings. *Thinking Skills & Creativity, 18*, 65–80.

Mark Wilson's interests focus on measurement and applied statistics. His work spans a range of issues in measurement and assessment from the development of new statistical models for analyzing measurement data, to the development of new assessments in subject matter areas such as science education, patient-reported outcomes, and child development, to policy issues in the use of assessment data in accountability systems. He has recently published three books: the first, Constructing measures: An item response modeling approach (Erlbaum), is an introduction to modern measurement; the second (with Paul De Boeck of the University of Leuven in Belgium), Explanatory item response models: A generalized linear and nonlinear approach (Springer-Verlag), introduces an overarching framework for the statistical modeling of measurements; the third, Towards coherence between classroom assessment and accountability (University of Chicago Press: National Society for the Study of Education), is an edited volume that explores the issues relating to the relationships between large-scale assessment and classroom-level assessment. He has chaired National Research Council committees on science achievement. He is founding editor of the new journal Measurement: Interdisciplinary Research and Perspectives. Dr. Wilson holds the Ph.D., University of Chicago, Educational Measurement & Educational Statistics, 1984.

Kathleen Scalise is an associate professor at the University of Oregon, in the Department of Educational Methodology, Policy and Leadership. She is also director of the U.S. National Assessment of Educational Progress (NAEP) Science for ETS. Her main research areas are technology-enhanced assessments in science and mathematics education, item response models with innovative item types, dynamically delivered content in e-learning, computer adaptive testing, and applications to equity studies. Previously, she was co-director of the UC Berkeley Evaluation and Assessment Research Center (BEAR) and she has served with the U.S. National Academies and other organizations. She also has served as a core member of the methodological group for the Assessment and Teaching of twenty-first-century Skills project created by Cisco, Intel and Microsoft; on the Oregon state task force for legislation on virtual public schools; and with the Curriculum Frameworks and Instructional Resources Division of the California Department of Education for the California Science Framework for K-12 Public Schools. She teaches the psychometric series of doctoral quantitative methods courses at U. Oregon and earned the Ph.D. in quantitative measurement at the University of California, Berkeley (2004). She holds teaching credentials for K-12 physical sciences and life sciences.