Chapter 3: Measurement and its Uses in Learning Analytics

Yoav Bergner

Learning Analytics Research Network, New York University, USA
DOI: 10.18608/hla17.003

ABSTRACT

Psychological measurement is a process for making warranted claims about states of mind. As such, it typically comprises the following: defining a construct; specifying a measurement model and (developing) a reliable instrument; analyzing and accounting for various sources of error (including operator error); and framing a valid argument for particular uses of the outcome. Measurement of latent variables is, after all, a noisy endeavor that can nevertheless have high-stakes consequences for individuals and groups. This chapter is intended to serve as an introduction to educational and psychological measurement for practitioners in learning analytics and educational data mining. It is organized thematically rather than historically, from more conceptual material about constructs, instruments, and sources of measurement error toward increasing technical detail about particular measurement models and their uses. Some of the philosophical differences between explanatory and predictive modelling are explored toward the end.

Keywords: Measurement, latent variable models, model fit

Knowing what students know and – given the increased attention to affective measures – how they feel is the basis for many conversations about learning. Measuring a student's knowledge, skills, attitudes/aptitudes/abilities (KSAs), and/or emotions is, however, less straightforward than measuring his or her height or weight. Psychological measurement is a noisy endeavor that can have high-stakes consequences, such as assignment to a special program (advanced or remedial), admission to a university, employment, hospitalization, or incarceration. Even small errors of measurement at the individual level can have large consequences when results are aggregated for groups (Kane, 2010). Sensitivity to these consequences has emerged over a century of methodology research enshrined in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). Insofar as measurement may be used in learning analytics and educational data mining for the purposes of understanding and optimizing learning and learning environments (Siemens & Baker, 2012), what are the tolerances for errors of measurement? After all, it has been argued that "harnessing the digital ocean" of data could ultimately replace the need for separate assessments (Behrens & DiCerbo, 2014). In the meantime, at minimum, one would like to avoid *misunderstanding* learning or *diminishing* learner experiences.

WHAT IS MEASUREMENT? PHILOSOPHY AND BASIC IDEAS

Discussions of psychological measurement often begin by drawing contrasts with physical measurement (for example, Armstrong, 1967; Borsboom, 2008; DeVellis, 2003; Lord & Novick, 1968; Maul, Irribarra, & Wilson, 2016; Michell, 1999; Sijtsma, 2011). A number of important facets of psychological measurement are raised in the process, namely its instrumentation or operationalization, the repeatability or precision of measurements, sources of error, and the interpretation of the measure itself. It can be said that psychological measurement comprises the following: defining a construct; specifying a measurement model and (developing) a reliable instrument; analyzing and accounting for various sources of error (including operator error); and framing a valid argument for particular uses of the outcome.

Constructs

Do psychological constructs really exist? In what sense can we really know a student's state of mind? We say that variables like physical length of an object are directly observed, or manifest, whereas a person's mental states or psychological traits are only indirectly observed, or latent. The term construct is used interchangeably with *latent variable*, while *trait* is used to imply a construct that is stable over time (Lord & Novick, 1968). In fact, even physical measurement is indirectly instrumented. Although we can perceive length directly through our senses, the *measurement* of length involves a process of comparison with a reference object or instrument, such as a tape measure. The tape measure provides a scale, such as inches or centimeters, which formalizes comparisons of length. For example, we can quantify the difference in two lengths by subtracting one measurement from the other.

In the first half of the twentieth century, efforts to settle philosophical issues of measurement led Bridgman (1927) and others to operationalism, wherein physical concepts like length, mass, and intensity are understood to be "synonymous with" the operations used to measure them. That is, length is understood as the outcome of a (possibly hypothetical) length measurement procedure. This idea can be carried over to psychological constructs, such as math ability and extraversion, by equating the constructs to scores on instruments used to measure them. Math ability is then equivalent to a score on a math test, and extraversion is a score on a Likert-item questionnaire. This positivist attitude is reflected in Stevens' definition of measurement as, "the assignment of numerals to objects or events according to rules" (1946, p. 677). The operationalist view of constructs was highly influential in the past, but it has been rejected for a host of reasons (Maul, Irribarra, & Wilson, 2016; Michell, 1999), notably that operationalism forces a redefinition of the construct for every instrument that exists to measure it.

If an operationalist interpretation is rejected, it appears to leave open epistemological and ontological questions about latent variables. Mislevy (2009, 2012) articulates a constructivist-realist position, namely that we can talk *as if* a construct exists without a commitment to strict realism by committing to model-based reasoning. Model-based reasoning means accepting a simplified representation of a system — for example, a construct-mediated relationship between persons and responses — that captures salient aspects (e.g., patterns) and allows us to explain or predict phenomena (Mislevy, 2009; we return to the explanatory/predictive distinction later in this chapter). As George Box famously said, "all models are wrong, but some are useful" (Box, 1979). The challenge remains to come up with useful models or, in terms of Stevens' definition, useful measurement rules.

Physical theories tend to be few in number and more comprehensive, whereas psychological theories are numerous and narrowly defined (DeVellis, 2003). Since constructs are invented things, there is no empirical limit to their number. It is possible to talk about a construct in the absence of a measurement *instrument*, but a measurement instrument is always designed to measure something. Therefore, we can infer an extremely partial list of constructs relevant to learning analytics from the instruments already developed to measure them. Examples include intelligence (e.g., the Stanford–Binet Intelligence Scale), scholastic aptitude (e.g., that SAT test), academic achievement (numerous examples include both large-scale tests and course exams), personality (e.g., the "big five" factor model; Digman, 1990), achievement-goal orientation (e.g., Midgley et al., 2000), achievement emotions (Pekrun, Goetz, Frenzel, Barchfeld, & Perry, 2011), grit (Duckworth, Peterson, Matthews, & Kelly, 2007), self-theories of intelligence and fixed/growth mindset (Dweck, 2000; Yeager & Dweck, 2012), intrinsic motivation (Deci & Ryan, 1985; Guay, Vallerand, & Blanchard, 2000), self-regulated learning and self-efficacy (e.g., Pintrich & De Groot, 1990), learning power (Buckingham Shum & Deakin Crick, 2012; Crick, Broadfoot, & Claxton, 2004), and crowd-sourced learning ability (Milligan & Griffin, 2016).

Several of the constructs listed above are multidimensional, that is they include multiple factors. The value of combining versus separating out related constructs is a subject of debate (Edwards, 2001; Schwartz, 2007).

Measurement Instruments

Psychological measurement instruments are typically called tests or questionnaires (also surveys and inventories) and are made up of items or indicators. The word test is more often used for constructs like intelligence, cognitive ability, and psychomotor skills, wherein the subject, or examinee, is instructed to try to maximize his or her performance (Sijtsma, 2011). Questionnaire respondents, by contrast, are asked to respond honestly about their thoughts, feelings, and behaviours. (Response bias can blur this distinction, as we shall describe when we come to validity). Note that this description of how subjects are expected to interact with instruments reveals the rudiments of a measurement *model*. We assume that the more able test taker will obtain a higher score on an ability test and that the more anxious subject will obtain a higher score on an anxiety questionnaire.

Sometimes the term measurement scale is used interchangeably with the instrument (DeVellis, 2003). Scale implies that the test or questionnaire has been scored. Binary items that have correct and incorrect answers

and yes/no questions are usually scored dichotomously with values in {0, 1}. Likert scale, rating scale, and visual-analogue scales (Luria, 1975) are other item types that can take discrete or continuous numerical values. Adding up the scores of individual items into a sum score (also, raw score) is one procedure for scoring an instrument, but it is not the only or necessarily the best procedure (Lord & Novick, 1968; Millsap, 2012). Weighted sum scores and item response theory (IRT; Baker & Kim, 2004) offer a range of alternatives.

The use of tests and questionnaires is a matter of both efficiency and standardization, compared with the alternative of observing people in real life and waiting for them to spontaneously express thoughts or exhibit the behaviours of interest (Sijtsma, 2011). In learning analytics, efficient collection of data is usually not the problem, but the lack of standardization can make it challenging to account for measurement error.

Source of Error in Measurements

We know from experience that psychological measurements are not as consistently repeatable as physical measurements. We also know that people's responses to an instrument may not faithfully reflect their abilities, attitudes, or other constructs of interest. Statistical models allow us to think of items, indicators, or tests as random samples of a latent variable. The latent variable can be a random variable, or it can be fixed, as in true score theory (Lord & Novick, 1968). Either way, the measurement samples will have error resulting from the inherent non-repeatability, which is sometimes called random error and is unbiased (in the sense of having an expectation value of zero over some distribution of repeated measures). There can also be systematic error, which is biased.

More precise or formal statements about error arise when we adopt a measurement framework or model. For example, in true score theory and factor analysis we can reason in terms of parallel tests or equivalent forms to derive estimates of an instrument's reliability. Measurement error can also be defined as any variance in the data not attributed to the construct, as explained by the model (AERA, APA, & NCME, 2014). We will revisit the sources of error after we flesh out our discussion of measurement models.

Reliability

Reliability is attributed to an instrument and is a measure of the consistency of scores (AERA, APA, & NCME, 2014), specifically the proportion of the total variance in scores attributed to the latent variable (DeVellis, 2003). It can be sample-dependent (in true score theory) and model-dependent (in more complicated models). The word is sometimes used to mean a particular reliability coefficient, most commonly Cronbach's (1951) alpha, α, which ranges from [0,1]. However, the term reliability is also used in the sense of test-retest reliability, which is actually a correlation, and inter-rater reliability (e.g., Cohen's kappa, κ; Cohen, 1968). Practitioners sometimes lean uncritically on guidelines for acceptable values of α, such as .70 as a lower bound (Cortina, 1993), to decide that scales are good enough to use. But it should be noted that statistical power improves with higher values of α (DeVellis, 2003). Thus, effort in improving the reliability of a scale can often outweigh the benefit of recruiting larger samples.

Validity

Validity is the foremost topic in the *Standards*, whose first chapter begins, "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests … It is incorrect to use the unqualified phrase 'the validity of the test'" (p. 11). Substituting the broader term "measure" for the narrower "test," it should be self-evident that validity is of paramount importance to learning analytics. There is a palpable focus in the *Standards* on shaping the language used in validation arguments, an approach also evident in Messick's (1995) influential reworking of Cronbach and Meehl (1955) (see also Kane, 2001). Types of evidence about validity (rather than "types of validity") include evidence about response processes, evidence about the internal structure of the instrument, convergent and discriminant evidence, criterion references (including predictive criteria), and evidence of generalizability.

We referred earlier in this chapter to the assumption that responses to questionnaires correspond to honest thoughts and feelings. However, there is extensive literature on types of response bias, from acquiescence bias (yea-saying; Messick & Jackson, 1961) to social desirability bias (also, faking good; Nederhof, 1985) to bias from extreme and moderate types of responders (i.e., people who tend to choose extreme ends of Likert-scales) (Bachman & O'Malley, 1984). Although more often documented for questionnaires and surveys about sensitive topics such as willingness to cheat, sexual fantasies, or attitudes about race, self-tuning or censoring of responses can also happen on educational tests, such as the force concept inventory (FCI; Hestenes, Wells, & Swackhamer, 1992) used to assess Newtonian thinking. Mazur (2007) reported a student specifically asking, "How should I answer these questions? According to what you taught us, or by the way I think about these things?" Finally, intentional rapid guessing behaviour can be thought of as a form of response bias (Wise & Kong, 2005). It should be clear that all of these sources of response bias challenge the uncritical interpretation of scale scores.

Measurement Models

The rubber meets the road in the technical details of measurement models. A measurement model is a formal mathematical relationship between a latent variable or set of variables and an observable variable or set of variables. A fully statistical measurement model may specify a distribution for the latent variable(s), a distribution for the observed variable(s), and a functional relationship between them. The latent variables are often understood as *causally explaining* the observations, which are subject to errors. Variances and covariances of random variables are described, explicitly or implicitly, in the model. Models make assumptions, for example the assumption of monotonicity (or, stricter, linearity) of the relationship between the construct and the measure or the assumption of zero covariance between error terms of unique items. If the assumptions of a model are violated, inferences made using the model may be wrong (Lord & Novick, 1968).

Since categorical and continuous variables involve different statistical methods, types of measurement models are sometimes classified into families according to the type of latent and observed variables, as shown in Table 3.1. This classification is not exhaustive, as hybrid models exist as well as generalized frameworks (Skrondal & Rabe-Hesketh, 2004) in which these model families become special cases. Growth models are extensions of measurement models to repeated measures and can apply to both continuous and categorical latent variables (e.g., Meredith & Tisak, 1990; Rabiner, 1989; Raudenbush & Bryk, 2002).

Table 3.1. Families of Latent Variable Models

Latent/Observed	Observed continuous	Observed categorical
Latent continuous	Factor models (Bollen, 1989; Mulaik, 2009)	Item response models (Lord & Novick, 1968; Baker & Kim, 2004)
Latent categorical	Latent mixture models (McLachlan & Peel, 2004)	Latent class models (Goodman, 2002)

SPECIFIC USES OF MEASUREMENT MODELS IN LEARNING ANALYTICS

We mentioned previously that psychological and educational measurement is applied for a variety of purposes including classification, diagnosis, ranking, placement, and certification of individuals as well as corresponding inferences about groups. Work in learning analytics and educational data mining also explores the complex web of relationships between psychological scales, behaviour, and performance in digital learning environments (Tempelaar, Rienties, & Giesbers, 2015). The purpose of this section is to provide a bit more depth about models and their uses in learning analytics and educational data mining. All topics are not treated equally, reflecting both space constraints and selection bias.

Factor Analysis

Factor analysis (Mulaik, 2009) models the correlations among observed variables through a linear relationship to a set of latent variables known as factors. The original one-factor model is Spearman's (1904) model of general intelligence *g*, used to explain correlations between scores on unrelated subject tests. True score theory, also known as classical test theory (Lord & Novick, 1968), can be derived as a special case of a single factor model in which all of the item factor loadings are the same. Thurstone (1947) developed the multiple (seven) factors model of intelligence.

Factor analysis is commonly divided into two enterprises. Exploratory factor analysis (EFA) is used to determine the number of latent factors from data without strong theoretical assumptions and is commonly part of scale development. However, EFA requires a number of important methodological decisions which, if made poorly, can lead to problematic results (Fabrigar, Wegener, MacCallum, & Strahan, 1999). In particular, Fabrigar et al. (1999) caution against confusing EFA with principal components analysis (PCA), a dimensionality reduction technique, which can result in erroneous conclusions about true factor structure. Confirmatory factor analysis (CFA) is a complementary set of techniques to test a theoretically proposed factor model by examining residuals between expected and observed correlations. Thus, CFA can be used to reject a model. CFA, along with path analysis and latent growth models, is subsumed by structural equation modelling (SEM; Bollen, 1989; Kline, 2010). Confirmatory factor analysis is not the same thing as running EFA multiple times with different population samples, although the case has been made for doing the latter (DeVellis, 2003).

Some learning analytics research is directly concerned with scale development and its integration with data gathered from learning management systems (e.g., Buckingham Shum & Deakin Crick, 2012; Milligan & Griffin, 2016). Other work focuses on associations between existing scales and outcome measures, such as the relationship between achievement emotions (Pekrun et al., 2011) and decisions regarding face-to-face and online instruction (Tempelaar, Niculescu, Rienties, Giesbers, & Gijselaers, 2012) or between motivational measures and completion of a massive open online course (Wang & Baker, 2015). When adapting an instrument or, especially, part of an instrument for new purposes, practitioners should be mindful of whether these new uses merit new validation arguments.

Latent Class and Latent Mixture Models

Dedic, Rosenfeld, and Lasry (2010) used latent class analysis to understand the distribution of physics misconceptions based on students' wrong answers on a physics concept test. Data came from administrations both before and at the end of a physics course (pre- and post-test). The authors identified an apparent progression from Aristotelian to Newtonian thinking through discrete classes of dominance fallacies. A widely used method for topic modelling of documents, latent Dirichlet allocation (LDA; Blei, Ng, & Jordan, 2003; see also several chapters in this volume) is a latent mixture model. Mixed membership models (Erosheva, Fienberg, & Lafferty, 2004) further generalize latent mixtures by allowing "fuzzy" or weighted assignments of an individual to multiple classes. The Gaussian mixture model forms the basis for model-based cluster analysis (Fraley & Raftery, 1998) applied to performance trajectories of MOOC learners (Bergner, Kerr, & Pritchard, 2015). It should be noted that not all clustering algorithms, however, are latent mixture models.

Item Response Theory (IRT)

Item response theory distinguished itself in the historical development of testing theory by modelling individual person-item interactions rather than total test scores, as in classical test theory. Conceptually, the purpose of IRT is "to describe the items by item parameters and the examinees by examinee parameters in such a way that we can predict probabilistically the response of any examinee to any item, even if similar examinees have never taken similar items before" (Lord, 1980, p. 11). A sample item characteristic curve (ICC) or, equivalently, item response function (IRF) for a binary item (e.g., correct/incorrect, agree/disagree, et cetera) is shown in Figure 3.1.

The salient characteristics of Figure 3.1 are as follows:

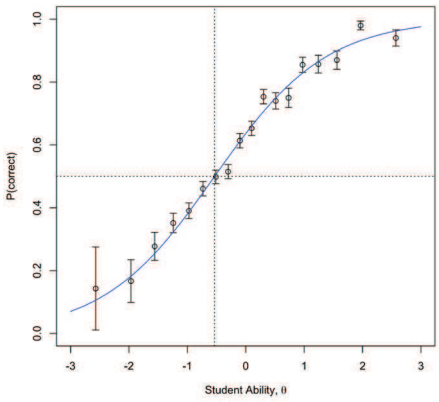

Figure 3.1. A sample item characteristic curve (ICC). Dotted lines indicate the P = 0.5 intercept.

1. The trait (e.g., ability) is quantified as a continuous random variable and is represented by θ on the horizontal axis. The variable is standardized to have a mean of zero and a variance of 1 in the population of interest. More of the trait, corresponding to a higher value of θ, is expected to increase the probability P of a positive (or correct) response. This is the *monotonicity assumption*. An observed violation of monotonicity means that that the fundamental person-item relationship is wrong, and including the item in a test would lead to bad fit and unreliable inferences.
2. Two ways of interpreting these curves were described by Holland (1990). In the stochastic subject interpretation, one literally imagines this curve as applying to an individual whose performance is inherently unpredictable. To paraphrase Holland, the stochastic subject explanation is intuitive, but not wholly satisfactory; we do not have a mechanistic explanation for the stochastic nature of the subject. In the random sampling interpretation, on the other hand, this curve makes sense as applied to a sample population of examinees. For example, among examinees within a certain ability range, some proportion will answer correctly. The points and error bars in the figure reflect this observation.[1]
3. The value of θ for which P = 0.5 is a reference intercept, which for a cognitive ability test item is called the difficulty. Note that difficulty is *ipso facto* on the same scale as ability, and so it makes sense to talk about the difference between a person's ability and the difficulty of an item.
4. The form of the probability link is commonly parametric with respect to the trait θ_i of individual i and a (set of) item parameters β_j, for item j,

$$P_{ij} = P(X_{ij} = 1|\theta_i, \beta_j) = f(\theta_i, \beta_j), \quad (1)$$

as in the case of the Rasch model (a single difficulty parameter) or of the two-parameter logistic (2PL) model. The 2PL model is shown in Figure 3.2; the fit to data is visibly good, and a G^2 goodness-of-fit test confirms as much. It should be noted that non-parametric IRT methods exist (Sijtsma, 1998).

When a person responds to several items in a measurement instrument, the idea is to combine the response information to make posterior estimates of the trait. For the likelihood of a response vector to factor into a product of individual item-level probabilities, the responses must be otherwise independent, conditional on the trait. This *conditional independence assumption*

[1] For the stochastic subject, these sample values would have to represent a set of identical trials by the same subject with no memory of the other trials. Although this seems odd in a cognitive test item, it is plausible in a psychomotor context. See Spray (1997).

may require the introduction of additional factors that explain inter-item dependence (e.g., Rijmen, 2010).

Evidence that IRT has some traction in education outside of high-stakes testing applications can be found in physics education research applications to the force concept inventory (FCI; Hestenes et al., 1992) and the mechanics baseline test (MBT; Hestenes & Wells, 1992). While these instruments have been in use for twenty-five years, item response model analyses started to appear more recently (Morris et al., 2006; Wang & Bao, 2010). Model-data fit for the FCI were generally acceptable. Cardamone et al. (2011), however, discovered two malfunctioning items in the MBT by inspecting the item response functions. An example is shown in Figure 3.2.

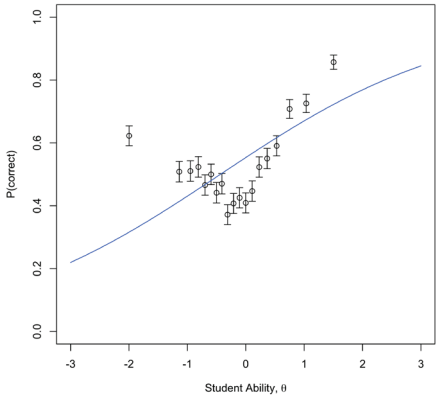

Figure 3.2. A poorly fitting item from the mechanics baseline test (MBT).

Something is fishy if low-ability students are more likely to answer an item correctly than average-ability students. Upon closer inspection, it was discovered that ambiguous wording of this test item allowed students holding a common misconception to misread the question and coincidentally choose the correct response for the wrong reason. In this case, two wrongs did make a right.

Following exploratory factor analyses of the FCI that identified multiple dimensions (Ding & Beichner, 2009; Scott, Schumayer, & Gray, 2012), a variation of multidimensional IRT was applied to the MBT (Bergner, Rayyan, Seaton, & Pritchard, 2013). Item response theory models have also been extended to the inherently sequential process behind multiple attempts to answer (answer-until-correct), an affordance which is common in online homework (Attali, 2011; Bergner, Colvin, & Pritchard, 2015; Culpepper, 2014).

Growth Models

Growth models apply any time a latent trait is changing systematically between measurements. They can be applied to changing attitudes (e.g., George, 2000), but we focus here on application to cognitive ability domains. There is an extensive literature in educational data mining on student models for intelligent problem-solving tutors, which are distinguished from curriculum sequencing tutors (Desmarais & Baker, 2011).

In cognitive tutors for mathematics (Anderson, Corbett, Koedinger, & Pelletier, 1995), sequences of practice items are designed to support mastery learning of fine-grained knowledge components (also, skills or productions), according to a cognitive model. Two approaches for modelling growth towards mastery in data from these systems are Bayesian knowledge tracing (BKT; Corbett & Anderson, 1995) and the additive factors models (AFM; Cen, Koedinger, & Junker, 2008; Draney, Pirolli, & Wilson, 1995). Learning curves analysis (Käser, Koedinger, & Gross, 2014; Martin, Mitrovic, Mathan, & Koedinger, 2010) has also been used to check for discrepancies between data and the cognitive model underlying the tutor.

According to the "law of practice" (Newell & Rosenbloom, 1981), the aggregate error rate T as a function of practice opportunity n should decay according to a power law $T=B_n^{-\alpha}$, where B and a are empirically determined. Bad fit between data and model, for example using r-squared measures, may motivate improvements to knowledge mapping. This may be seen as an analogue to the item analysis in Figure 3.2, where a faulty item is detected. In this case, however, the assignment of a sequence of items to a knowledge component is seen as faulty.

In BKT, the latent variable is mastery of a procedural knowledge component and is binary-valued, $M \in \{0, 1\}$. The probability link between mastery and correctness $X \in \{0, 1\}$ on any given opportunity is a 2x2 conditional probability table, but by analogy with Eq. (1), it can be written in terms of guess (g) and slip (s) parameters as,

$$P(X = 1|M) = (1 - s)^M \, g^{(1-M)} \qquad (2)$$

Importantly, the attempts are not viewed as independent. Rather, the key idea in BKT is that students begin with some prior probability of mastery and move towards mastery (they learn) on each practice opportunity according to the rule,

$$P(M_n) = P(M_{n-1}) + \tau(1 - P(M_{n-1})) \qquad (3)$$

Here τ is a growth parameter. Recently, van de Sande (2013) demonstrated that BKT implies an exponential rather than a power law relationship between practice attempts and error rates. This would make BKT a mis-specified model for data that satisfy a power law

of practice. The additive factors model, by contrast, is designed to fit the power law of practice paradigm. Käser et al. (2014) showed that prediction accuracy of BKT is often indistinguishable from AFM. Regarding fit of the latter, they noted systematic bias in aggregate residuals analyses.

AFM has been referred to as an extension of IRT (Koedinger, McLaughlin, & Stamper, 2012), and indeed the relation to the linear logistic test model (LLTM; Fischer, 1973) was clear in the progenitor of this model (Draney et al., 1995). However, in passing to its current form, the model was changed in a critical way. The LLTM is a Rasch-type IRT model in which the difficulty of an item is decomposed as a sum over potential properties of the item. Writing the Rasch model as,

$$logit(P_{ij}) = ln(P_{ij}/(1-P_{ij})) = \theta_i - \beta_j, \quad (4)$$

the difficulty β_j of item j is further decomposed,

$$\beta_j = c_j + \sum_k w_{jk}\alpha_k, \quad (5)$$

where α_k are difficulties of "basic" operations (Fischer's term) and the indicators w_{ik} are 0 or 1 depending on whether these operations are required in item j. If all items use the same operations, the model clearly reduces to the Rasch model with a simple offset,

$$\beta_j = c_j + \alpha. \quad (6)$$

Although the model of Draney et al. (1995) contained an item-level difficulty parameter, in AFM only the difficulties of the component skills are retained. In addition, a practice term is introduced,[2]

$$\beta_j^{AFM} = \sum_k w_{jk}\alpha_k - \sum_k w_{jk}\gamma_k T_{ik}, \quad (7)$$

where γ_k is a growth parameter and T_{ik} is a count of the previous practice attempts of learner i on skill k. If a sequence of practice problems all involve the same skills, which is common for tutor applications, then for each sequence, this parameter reduces to,

$$\beta_j^{AFM} = \alpha - \gamma T_i. \quad (8)$$

Importantly, this is not a property of the item at all, as is clear from the subscripts on the right hand side, which depend only on the learner. By dropping the c_j parameter in Equations (7)–(8), the AFM has actually become a fixed effect growth model.

From a modelling perspective, it is not surprising that the item-level difficulty parameter was removed, as keeping both difficulty and growth parameters creates a problem for identifiability. A model is identifiable if its parameters can be unambiguously learned given sufficient data. However, for students working on a fixed sequence of items, the increased success rate due to learning/growth can be attributed to decreasing item difficulty. The two effects cannot be distinguished unless item difficulties have been separately calibrated under conditions where there is no growth.

[2] One sign convention from Cen et al. (2008) has been changed to make the model consistent with the usual Rasch model, with a difficulty rather than an easiness parameter.

Cognitive Diagnostic Models

A seminal study of mixed-number subtraction using cognitive task analysis led Tatsuoka (1983) to develop the Q-matrix method and a model for diagnosing specific sub-skills (e.g., converting a whole number to a fraction) in an educational test. The Q-matrix is a discrete mapping of items to requisite sub-skills and is traditionally specified in the assessment model. Cognitive diagnostic models have since been considerably generalized (Rupp & Templin, 2008; von Davier, 2005), and efforts to learn the Q-matrix from data have appeared in educational data mining research (Barnes, 2005; Desmarais, 2012; Koedinger et al., 2012).

SOURCES OF ERROR, REVISITED

Having explored some of the measurement models involved in studying motivation, emotion, and cognition, it is worth revisiting the important subject of error. Practitioners should be mindful that additional sources of error could be introduced by using models with the wrong parameters, by using the wrong models, or by using the models wrongly.

The use of a model may depend on parameters whose estimation is itself subject to error. These uncertainties should be acknowledged, but they are not necessarily serious if the model is consistent as a *data-generating model* for the observed data. That is, we think of the statistical model as a stochastic process that can be used to generate (also, sample or simulate) data (Breiman, 2001). For example, we can simulate data from coin flips using a Bernoulli process, even if we are unsure about whether the real coin is fair. In principle, our parameter for the probability of heads in our model can be improved with more data from the real coin. This is different from the case when the model itself, either in terms of the latent variables or the link functions, is inconsistent with the true generating model. The second case is called model mis-specification (White, 1996). Goodness-of-fit tests evaluate the consistency between the observed data and the generating model to retain or reject the model (White, 1996; Haberman, 2009; Ames & Penfield, 2015).

EXPLANATION AND PREDICTION

Predictive modelling is one of the most prominent methodological approaches in educational data mining (Baker & Siemens, 2014; Baker & Yacef, 2009). Measurement theory, by contrast, is decidedly explanatory, as are most of the statistical methods traditionally used in the social sciences (Breiman, 2001; Shmueli,

2010). While an explanatory model can be used to make predictions – and an error-free explanatory model would make perfect predictions – a predictive model is not necessarily explanatory. Breiman (2001) expressed the distinction in terms of two cultures: the data modelling culture (98% of statistics, informally according to Breiman) and the algorithmic modelling culture (the 2%, in which Breiman included himself).[3] Shmueli (2010) contrasted the entire design process for statistical modelling when viewed from either a prediction or an explanation lens. The interpretability or non-interpretability of predictors in a complex prediction model is only one aspect of the distinction (see also Liu & Koedinger, this volume). The different viewpoints fundamentally inform how researchers handle error and uncertainty.

The predictive view is expressed, for example, in a recent best paper from the educational data mining conference. The authors assert that, "the only way to determine if model assumptions are correct is to construct an alternative model that makes different assumptions and to determine whether the alternative outperforms [out-predicts] BKT" (Khajah, Lindsey, & Mozer, 2016, p. 95, editorial note added). Strictly speaking, model prediction performance is not a way to determine if model assumptions are violated. By contrast, both informal checks and formal tests for goodness-of-fit have been discussed above. However, the quote is a reflection of the algorithmic modelling culture in which models are validated by predictive accuracy (Breiman, 2001). More problematically, it carries a presumption that predictive power points to the truer model. In fact, it is explanatory power that plays this role. Put in terms of variance components, "in explanatory modelling the focus is on minimizing bias to obtain the most accurate representation of the underlying theory. In contrast, predictive modelling seeks to minimize the combination of bias and variance, occasionally sacrificing theoretical accuracy for improved empirical precision" (Shmueli, 2010, p. 293). It should be emphasized that explanatory power and predictive power do not always point in the same direction. Indeed, Hagerty and Srinivasan (1991) proved that, in noisy circumstances, under-specified multiple regression models can have more predictive power than the correctly specified (true) model.

Suthers and Verbert (2013) have described learning analytics as a "middle space" between learning science and analytics. Perhaps it may also be thought of as occupying a methodological middle space between explanatory and predictive approaches. In that case, the field may benefit from understanding the nuances of both perspectives.

FURTHER READING

Psychological measurement is almost as old as psychology itself and as old as statistics. Authoritative, technical, and somewhat encyclopedic sources are the anthology of psychometrics in the *Handbook of Statistics* series (Rao & Sinharay, 2006) and the "bible" of *Educational Measurement*, now in its fourth edition (Brennan, 2006). Educational measurement volumes and the *Standards* (AERA, APA, & NCME, 2014) tend to emphasize testing, where specific issues are reliability, validity, generalizability, comparability, and fairness. DeVellis' (2003) concise volume on scale development is a non-technical introduction to psychological measurement and omits topics specific to large-scale testing, such as linking scores from parallel test forms.

[3] Breiman uses the term *information* in place of *explanation* and in contrast to *prediction*.

REFERENCES

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

Ames, A. J., & Penfield, R. D. (2015). An NCME instructional module on polytomous item response theory models. *Educational Measurement: Issues and Practice*, 34(3), 39–48. doi:10.1111/emip.12023

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4(2), 167–207.

Armstrong, J. S. (1967). Derivation of theory by means of factor analysis or Tom Swift and his electric factor analysis machine. *The American Statistician*, 21, 17–21.

Attali, Y. (2011). Immediate feedback and opportunity to revise answers: Application of a graded response IRT model. *Applied Psychological Measurement*, 35(6), 472–479.

Baker, F. B., & Kim, S.-H. (Eds.). (2004). *Item response theory: Parameter estimation techniques*. Boca Raton, FL: CRC Press.

Baker, R. S., & Siemens, G. (2014). Educational data mining and learning analytics. In R. Sawyer (Ed), *The Cambridge handbook of the learning sciences* (pp. 253–272). Cambridge University Press.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3–17.

Barnes, T. (2005). The Q-matrix method: Mining student response data for knowledge. In the Technical Report (WS-05-02) of the AAAI-05 Workshop on Educational Data Mining.

Behrens, J. T., & DiCerbo, K. E. (2014). Harnessing the currents of the digital ocean. In J. A. Larusson & B. White (Eds.), *Learning analytics: From research to practice* (pp. 39–60). New York: Springer.

Bachman, J. G., & O'Malley, P.M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly*, 48, 491–509.

Bergner, Y., Colvin, K., & Pritchard, D. E. (2015). Estimation of ability from homework items when there are missing and/or multiple attempts. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 118–125). New York: ACM.

Bergner, Y., Kerr, D., & Pritchard, D. E. (2015). Methodological challenges in the analysis of MOOC data for exploring the relationship between discussion forum views and learning outcomes. In O. C. Santos et al. (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining* (EDM2015), 26–29 June 2015, Madrid, Spain (pp. 234–241). International Educational Data Mining Society.

Bergner, Y., Rayyan, S., Seaton, D., & Pritchard, D. E. (2013). Multidimensional student skills with collaborative filtering. *AIP Conference Proceedings*, 1513(1), 74–77. doi:10.1063/1.4789655

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan.), 993–1022.

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.

Borsboom, D. (2008). Latent variable theory. *Measurement: Interdisciplinary Research & Perspective*, 6(1–2), 25–53. http://doi.org/10.1080/15366360802035497

Box, G. E. (1979). Robustness in the strategy of scientific model building. *Robustness in Statistics*, 1, 201–236.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199–215. http://doi.org/10.2307/2676681

Brennan, R. L. (Ed.). (2006). *Educational measurement*. Praeger Publishers.

Bridgman, P. W. (1927). *The logic of modern physics*. New York: Macmillan.

Buckingham Shum, S., & Deakin Crick, R. (2012). Learning dispositions and transferable competencies: Pedagogy, modeling and learning analytics. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (LAK '12), 29 April–2 May 2012, Vancouver, BC, Canada (pp. 92–101). New York: ACM.

Cardamone, C. N., Abbott, J. E., Rayyan, S., Seaton, D. T., Pawl, A., & Pritchard, D. E. (2011). Item response theory analysis of the mechanics baseline test. *Proceedings of the 2011 Physics Education Research Conference* (PERC 2011), 3–4 August 2011, Omaha, NE, USA (pp. 135–138). doi:10.1063/1.3680012

Cen, H., Koedinger, K. R., & Junker, B. (2008). Comparing two IRT models for conjunctive skills. In B. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie (Eds.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (ITS 2008), 23–27 June 2008, Montreal, PQ, Canada (pp. 796–798). Springer.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213–220.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4, 253–278.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78(1), 98.

Crick, R. D., Broadfoot, P., & Claxton, G. (2004). Developing an effective lifelong learning inventory: The ELLI project. *Assessment in Education: Principles, Policy & Practice*, 11(3), 247–272.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302.

Culpepper, S. A. (2014). If at first you don't succeed, try, try again: Applications of sequential IRT models to cognitive assessments. *Applied Psychological Measurement*, 38(8), 632–644. doi:10.1177/0146621614536464

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behaviour*. New York: Plenum.

Dedic, H., Rosenfield, S., & Lasry, N. (2010). Are all wrong FCI answers equivalent? *AIP Conference Proceedings*, 1289, 125–128. doi.org/10.1063/1.3515177

Desmarais, M. C. (2012). Mapping question items to skills with non-negative matrix factorization. *ACM SIGKDD Explorations Newsletter*, 13(2), 30–36.

Desmarais, M. C., & Baker, R. S. (2011). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, 22(1–2), 9–38. doi:10.1007/s11257-011-9106-8

DeVellis, R. F. (2003). *Scale development: Theory and applications*. Applied Social Research Methods Series (Vol. 26). Thousand Oaks, CA: Sage Publications.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41(1), 417–440.

Ding, L., & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics: Physics Education Research*, 5(2), 1–17. doi:10.1103/PhysRevSTPER.5.020103

Draney, K., Pirolli, P., & Wilson, M. R. (1995). A measurement model for a complex cognitive skill. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, 9, 1087–1101.

Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality and development*. Philadelphia, PA: Taylor & Francis.

Edwards, J. R. (2001). Multidimensional constructs in organizational behavior research: An integrative analytical framework. *Organizational Research Methods*, *4*(2), 144–192.

Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5220–5227.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359–374.

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, *41*(8), 578–588.

George, R. (2000). Measuring change in students' attitudes toward science over time: An application of latent variable growth modeling. *Journal of Science Education and Technology*, *9*(3), 213–225.

Goodman, L. (2002) Latent class analysis: The empirical study of latent types, latent variables, and latent structures. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 3–55). Cambridge, UK: Cambridge University Press.

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

Haberman, S. J. (2009). Use of generalized residuals to examine goodness of fit of item response models. *ETS Research Report RR-09-15*.

Hagerty, M. R., & Srinivasan, V. (1991). Comparing the predictive powers of alternative multiple regression models. *Psychometrika*, *56*(1), 77–85.

Hestenes, D., & Wells, M. (1992). A mechanics baseline test. *The Physics Teacher*, *30*(3), 159–166.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*(3), 141. doi:10.1119/1.2343497

Holland, P. W. (1990). On the sampling theory roundations of item response theory models. *Psychometrika*, *55*(4), 577–601. http://doi.org/10.1007/BF02294609

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342.

Kane, M. (2010). Errors of measurement, theory, and public policy. William H. Angoff Memorial Lecture Series. *Educational Testing Service*. https://www.ets.org/Media/Research/pdf/PICANG12.pdf

Käser, T., Koedinger, K. R., & Gross, M. (2014). Different parameters – same prediction: An analysis of learning curves. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (EDM2013), 6–9 July 2013, Memphis, TN, USA (pp. 52–59). International Educational Data Mining Society/Springer.

Khajah, M., Lindsey, R. V., & Mozer, M. C. (2016). How deep is knowledge tracing? In T. Barnes, M. Chi, & M. Feng (Eds.), *Proceedings of the 9th International Conference on Educational Data Mining* (EDM2016), 29 June–2 July 2016, Raleigh, NC, USA (pp. 94–101). International Educational Data Mining Society.

Kline, R. B. (2010). *Principles and practice of structural equation modeling*. New York: Guilford.

Koedinger, K. R., McLaughlin, E. A., & Stamper, J. (2012). Automated student model improvement. In K. Yacef et al. (Eds.), *Proceedings of the 5th International Conference on Educational Data Mining* (EDM2012), 19–21 June 2012, Chania, Greece. International Educational Data Mining Society. http://www.learnlab.org/research/wiki/images/e/e1/KoedingerMcLaughlinStamperEDM12.pdf

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Luria, R. E. (1975). The validity and reliability of the visual analogue mood scale. *Journal of Psychiatric Research*, 12(1), 51–57.

Martin, B., Mitrovic, T., Mathan, S., & Koedinger, K. R. (2010). Evaluating and improving adaptive educational systems with learning curves. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research*, *21*, 249–283.

Maul, A., Irribarra, D. T., & Wilson, M. (2016). On the philosophical foundations of psychological measurement. *Measurement*, 79, 311–320. http://doi.org/10.1016/j.measurement.2015.11.001

Mazur, E. (2007). Confessions of a converted lecturer. https://www.math.upenn.edu/~pemantle/active-papers/Mazurpubs_605.pdf

McLachlan, G., & Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, 55(1), 107–122.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749.

Messick, S., & Jackson, D. (1961). Acquiescence and the factorial interpretation of the MMPI. *Psychological Bulletin*, 58(4), 299–304

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept* (Vol. 53). Cambridge University Press.

Midgley, C., Maehr, M. L., Hruda, L., Anderman, E. M., Anderman, L., Freeman, K. E., et al. (2000). *Manual for the patterns of adaptive learning scales (PALS)*. Ann Arbor, MI: University of Michigan.

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.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R. L. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 83–108). Charlotte, NC: Information Age Publishing.

Mislevy, R. J. (2012). Four metaphors we need to understand assessment. Draft paper commissioned by the Gordon Commission. http://www.gordoncommission.com/rsc/pdfs/mislevy_four_metaphors_understand_assessment.pdf

Morris, G. A., Branum-Martin, L., Harshman, N., Baker, S. D., Mazur, E., Dutta, S., … McCauley, V. (2006). Testing the test: Item response curves and test quality. *American Journal of Physics*, 74(5), 449. doi:10.1119/1.2174053

Mulaik, S. A. (2009). *Foundations of factor analysis*. Boca Raton, FL: CRC Press.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, 15(3), 263–280. http://doi.org/10.1002/ejsp.2420150303

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. *Cognitive Skills and their Acquisition*, 6, 1–55.

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The achievement emotions questionnaire (AEQ). *Contemporary Educational Psychology*, 36(1), 36–48. http://doi.org/10.1016/j.cedpsych.2010.10.002

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, 82(1), 33.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257–286.

Rao, C. R., & Sinharay, S. (Eds.). (2006). *Handbook of statistics 26: Psychometrics*. Elsevier. doi:10.1016/S0169-7161(06)26037-1

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, 47(3), 361–372. doi:10.1111/j.1745-3984.2010.00118.x

Rupp, A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research & Perspective*, 6(4), 219–262. doi:10.1080/15366360802490866

Schwartz, S. (2007). The structure of identity consolidation: Multiple correlated constructs or one superordinate construct? *Identity*, 7(1), 27–49.

Scott, T. F., Schumayer, D., & Gray, A. R. (2012). Exploratory factor analysis of a force concept inventory data set. *Physical Review Special Topics: Physics Education Research*, 8(2). doi:10.1103/PhysRevSTPER.8.020105

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310. http://doi.org/10.1214/10-STS330

Siemens, G., & Baker, R. S. (2012). Learning analytics and educational data mining: Towards communication and collaboration. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (LAK '12), 29 April–2 May 2012, Vancouver, BC, Canada (pp. 252–254). New York: ACM.

Sijtsma, K. (2011). Introduction to the measurement of psychological attributes. *Measurement*, 44(7), 1209–1219. doi:10.1016/j.measurement.2011.03.019

Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement*, 22(1), 3–31. doi:10.1177/01466216980221001

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC Press.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2), 201–292.

Spray, J. A. (1997). Multiple-attempt, single-item response models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 209–220). New York: Springer.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677–680.

Suthers, D., & Verbert, K. (2013). Learning analytics as a middle space. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (LAK '13), 8–12 April 2013, Leuven, Belgium (pp. 1–4). New York: ACM.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20, 345–354.

Tempelaar, D. T., Niculescu, A., Rienties, B., Giesbers, B., & Gijselaers, W. H. (2012). How achievement emotions impact students' decisions for online learning, and what precedes those emotions. *Internet and Higher Education*, 15(3), 161–169. doi:10.1016/j.iheduc.2011.10.003

Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Computers in Human Behavior*, 47, 157–167. doi:10.1016/j.chb.2014.05.038

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago, IL: University of Chicago Press.

van de Sande, B. (2013). Properties of the Bayesian knowledge tracing model. *Journal of Educational Data Mining*, 5(2), 1–10.

von Davier, M. (2005). A general diagnostic model applied to language testing data. *The British Journal of Mathematical and Statistical Psychology*, *61*(Pt 2), 287–307. doi:10.1348/000711007X193957

Wang, Y., & Baker, R. S. (2015). Content or platform: Why do students complete MOOCs? *Journal of Online Learning and Teaching*, *11*(1), 17.

Wang, J., & Bao, L. (2010). Analyzing force concept inventory with item response theory. *American Journal of Physics*, *78*(10), 1064. doi:10.1119/1.3443565

White, H. (1996). *Estimation, inference and specification analysis* (No. 22). Cambridge University Press.

Wise, S., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183.

Yeager, D. S., & Dweck, C. S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational Psychologist*, *47*(4), 302–314.