Causal Rasch Models in Language Testing: An Application Rich Primer

A. Jackson Stenner, Mark Stone, William P. Fisher Jr. and Donald Burdick

A new paradigm for measurement in education and psychology, which mimics much more closely what goes on in the physical sciences was foreshadowed by Thurstone (1926) and Rasch (1961):

> It should be possible to omit several test questions at different levels of the scale without affecting the individual's score [measure].
>
> … a comparison between two individuals should be independent of which stimuli [test questions] within the class considered were instrumental for comparison; and it should also be independent of which other individuals were also compared, on the same or some other occasion.

Taken to the extreme, we can imagine a group of language test takers (reading, writing, speaking, or listening) being invariantly located on a scale without sharing a single item in common. i.e. no item is taken by more than one person. This context defines the limit case of omitting items and making comparisons independent of the particular questions answered by any test taker.

More formally we can contrast a fully crossed pxi design (persons crossed with items) in which all persons take the same set of items with a nested design i:p (all items are unique to a specific person). The more common design in language research is pxi simply because there is no method of data analysis that can extract

A.J. Stenner (✉) · D. Burdick
MetaMetrics, Durham, NC, USA
e-mail: jstenner@lexile.com

A.J. Stenner
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

M. Stone
Aurora University, Aurora, IL, USA

W.P. Fisher Jr.
University of California—Berkeley, Berkeley, CA, USA

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_1

invariant comparisons from an i:p design unless item calibrations are available from a previous calibration study or are theoretically specified.

But the i:p design is routinely encountered in physical science measurement contexts and in health care when, for example, parents report their child's temperature to a pediatrician. Children in different families do not share the same thermometers. Furthermore, the thermometers may not even share the same measurement mechanism (mercury in a tube vs. NexTemp technology, see Note 1). Yet, there is little doubt that the children can be invariantly ordered and spaced on any of several temperature scales.

The difference between the typical language testing and temperature scenarios is that the same construct theory, engineering specifications and manufacturing quality control procedures have been enforced for each and every thermometer, even though the measurement mechanism may vary. In addition, considerable resources have been expended in ensuring the measuring unit (°F or °C) has been consistently mapped to the measurement outcome (e.g. column height of mercury or cavity count turning black on a NexTemp thermometer) (Hunter 1980; Latour 1987). Substantive theory, engineering specifications, and functioning metrological networks—not data- render comparable measurement from these disparate thermometers. This contrast illustrates the dominant distinguishing feature between measurement in the physical and educational sciences including EFL, ESL and ENL language testing. Educational measurement does not, as a rule, make use of substantive theory in the ways the physical sciences do (Taagepera 2008). Nor does educational science embrace metric unification even when constructs (e.g. reading ability) repeatedly assert their separate independent existences (Fisher 1997, 1999, 2000a, b; Fisher et al. 1995).

Typical applications of Rasch models in language testing are thin on substantive theory. Rarely is there an a priori specification of the item calibrations (i.e. constrained model). Instead the researcher estimates both person parameters and item parameters from the same pxi data set. For Kuhn (1961) this practice is at odds with the scientific function of measurement in that substantive theory almost never will be revealed by measuring. Rather "the scientist often seems to be struggling with facts [measurement outcomes, raw scores], trying to force them to conformity with a theory s(he) does not doubt" (p. 163). Kuhn is speaking about substantive construct theory, not axiomatic measurement theory. Demonstrating data fit to a descriptive Rasch Model or sculpting a data set by eliminating misfitting items and persons and then rerunning the Rasch analysis to achieve satisfactory fit is, specifically not, the "struggling" Kuhn is referring to.

The gold standard demonstration that a construct is well specified is the capability to manufacture strictly parallel instruments. A strictly parallel instrument is one in which the correspondence table linking attribute measure to measurement outcome (count correct) is identical although items are different on each parallel instrument. So, imagine two 4000 word 1300L articles, one on 'atomic theory' and one on 'mythology'. Both articles are submitted to a machine that builds 45 four choice cloze items distributed about one item for every 80–100 words. These one-off items are assumed to have calibrations sampled from a normal distribution with a mean

equal to 1300L and a standard deviation equal to 132L. With this information, an ensemble Rasch model (Lattanzio et al. 2012) can produce a correspondence table linking count correct to Lexile measure. Since the specifications (test length, text measure, text length and item spread) are identical for the two articles, the correspondence tables will also be identical; on both forms 25 correct answers converts to 1151L and 40 correct answers converts to 1513L, and so on. The same basic structure plays out with NexTemp® thermometers. A NexTemp® thermometer has 45 cavities. Twenty-five cavities turning black converts to a temperature of 37.9 °C, whereas 40 cavities turning black converts to 39.4 °C. In both cases theory, engineering specifications and manufacturing guidelines combine to produce strictly parallel instruments for measuring reading ability and human temperature and in each case it is possible to manufacture large quantities of identical instruments. The capacity to manufacture "strictly" parallel instruments is a milestone in an evolving understanding of an attribute and its measurement. Richard Feynman wrote: "What I cannot create, I don't understand!" We demonstrate our understanding of how an instrument works by creating copies that function like the original.

Descriptive Rasch Models Versus Causal Rasch Models

Andrich (2004) makes the case that Rasch models are powerful tools precisely because they are prescriptive, not descriptive, and when model prescriptions meet data, anomalies arise. Rasch models invert the traditional statistical data-model relationship. Rasch models state a set of requirements that data must meet if those data are to be useful in making measurements. These model requirements are independent of the data. It does not matter if the data are bar presses, counts correct on a reading test, or wine taste preferences, if these data are to be useful in making measures of rat perseverance, reading ability, or vintage quality all three sets of data must conform to the same invariance requirements. When data fail to fit a model, Rasch measurement theory (Rasch 1960; Andrich 1988, 2010; Wright 1977, 1999) does not respond by relaxing the invariance requirements and adding, say, an item specific discrimination parameter to improve fit, as does Item Response Theory (Hambleton et al. 1991). Rather, the Rasch approach is to examine the items serving as the medium for making observations, and to change them in ways likely to produce new data conforming with theory and data model expectations.

A causal Rasch model (in which item calibrations come from theory, not data) is then doubly prescriptive (Stenner et al. 2009a, b). First, in accord with Rasch, it is prescriptive regarding the data structures that must be present:

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared. Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for comparison; and it should also be independent of which other individuals were also compared, on the same or on some other occasion (Rasch 1961, p. 321).

Second, causal Rasch Models (Burdick et al. 2006; Stenner et al. 2008) prescribe the values imposed by substantive theory on the item calibration estimates. Thus, the data, to be useful in making measures, must conform to both Rasch model invariance requirements *and* to substantive theory invariance requirements as specified by the theoretical item calibrations.

When data meet both sets of requirements then those data are useful not just for making measures of some vaguely defined construct but are useful for making measures of that precise construct specified by the equation that produced the theoretical item calibrations. We emphasize that these dual invariance requirements come into stark relief in the extreme case of no connectivity across stimuli or examinees (i:p). How, for example, are two readers to be measured on the same scale if they share no common text passages or items? If you read a Hunger Games novel and answer machine generated questions about it, and I read a Lord of the Rings novel and answer machine generated questions about it, how would it be possible to realize an invariant comparison of our reading abilities except by means of predictive theory? How else would it be possible to know that you read 250L better than I, and, furthermore, that you comprehended 95 % of what you read, whereas I comprehended 75 % of what I read? Most importantly, by what other means than theory would it ever be possible to reproduce this result to within a small range of error using another two completely different books as the basis of comparison?

Given that seemingly nothing is in common between the above two reading experiences, invariant comparisons might be thought impossible. Yet in the thermometer example, it is in fact a routine everyday experience for different instruments to be interpreted as informing comparable measures of temperature. Why are we so quick to accept that you have a 104 °F high grade fever and I have a 100 °F low grade fever (based on measurements from two different thermometers) and yet find the book reading example inexplicable? Is it because there are fundamental differences between physical science measurement and behavioral science measurement? No! The answer lies in well-developed construct theory, rigorously established instrument engineering principles, and uniform metrological conventions (Fisher 2009).

Clearly, each of us has had ample confirmation that weight denominated in pounds and kilograms can be well measured by any reputable manufacturer's bathroom scale. Experience with diverse bathroom scales has convinced us that, within a pound or two of error, these instruments will produce not just invariant relative differences between two persons but will also meet the more stringent expectation of invariant absolute magnitudes for each individual independent of instrument. Over centuries, instrument engineering has steadily improved to the point that for most purposes "uncertainty of measurement" (usually interpreted as the standard deviation of a distribution of imagined or actual replications taken on a single person) can be effectively ignored for most bathroom scale applications. And, quite importantly, by convention (i.e., the written or unwritten practice of a community) weight is denominated in standardized units (kilograms or pounds). The choice of any given unit is arbitrary, but what is decisive is that a unit is agreed to by

the community and is slavishly maintained through consistent implementation, instrument manufacture, and reporting. At present, language ability (reading, writing, speaking, and listening) does not enjoy a common construct definition, nor a widely promulgated set of instrument specifications, nor a conventionally accepted unit of measurement. The challenges that must be addressed in defining constructs, specifying instrument characteristics, and standardizing units include cultural assumptions about number and objectivity, political challenges in shaping legislation, resource allocation, and the expectations and procedures of social scientists (Fisher 2012, n.d.). In this context, the Lexile Framework for Reading (Stenner et al. 2006) stands as an exemplar of how psychosocial measurement can be unified in a manner precisely parallel to the way unification was achieved for length, temperature, weight and dozens of other useful attributes (Stenner and Stone 2010).

A causal (constrained) Rasch model (Stenner et al. 2009a, b) that fuses a substantive theory to a set of axioms for conjoint additive measurement affords a much richer context for the identification and interpretation of anomalies than does a descriptive i.e. unconstrained Rasch model. First, with the measurement model and the substantive theory fixed, anomalies are understood as problems with the data. Attending to the data ideally leads to improved observation models (e.g. new task types) that reduce unintended dependencies and variability. An example of this kind of improvement in measurement was realized when the Duke of Tuscany put a top on some of the early thermometers, thus reducing the contaminating influences of barometric pressure on the measurement of temperature. In contrast with the descriptive paradigm dominating much of education science, the Duke did not propose parameterizing barometric pressure in the model in the hope that the boiling point of water at sea level, as measured by open top thermoscopes, would then match the model expectations at 3000 ft above sea level (for more on the history of temperature see Chang 2004).

Second, with both model and construct theory fixed our task is to produce measurement outcomes that fit the invariance requirements of both measurement theory and construct theory. By analogy, not all fluids are ideal as thermometric fluids. Water, for example, is non-monotonic in its expansion with increasing temperature. Mercury, in contrast, has many useful properties as a thermometric fluid. Does the discovery that not all fluids are useful thermometric fluids invalidate the concept of temperature? No! In fact, a single fluid with the necessary properties would suffice to validate temperature as a useful construct. The existence of a persistent invariant framework makes it possible to identify anomalous behavior (water's strange behavior) and interpret it in an expanded theoretical framework (Chang 2004).

Analogously, finding that not all reading item types produce data that conform to the dual invariance requirements of a Rasch model and the Lexile theory does not invalidate either the axioms of conjoint measurement theory or the Lexile reading theory. Rather, the anomalous behaviors of some kinds of text (recipes, and, poems) are open invitations to expand the theory to account for these deviations from expectation. Notice here the subtle shift in perspective. We do not need to find 1000 unicorns; one will do to establish the reality of the class. The finding that reader

behavior on a minimum of two types of reading tasks can be regularized by the joint actions of the Lexile theory and a Rasch model is sufficient evidence for the existence of the reading construct (Markus and Borsboom 2013). Of course, actualizing this scientific reality to make the reading construct a universally uniform and available object in the world requires the investment of significant social, legal, and economic resources (Fisher 2005, 2009, 2000a, b, 2011, n.d.; Fisher and Stenner n.d.).

Equation (1) is a causal Rasch model for dichotomous data, which sets a measurement outcome (expected score) equal to a sum of modeled probabilities

$$\text{Expected score} =: \sum \frac{e^{(b-d_i)}}{1+e^{(b-d_i)}} \tag{1}$$

The measurement outcome is the dependent variable and the measure (e.g., person parameter, b) and instrument (e.g., the parameters d_i pertaining to the difficulty d of item i) are independent variables. The measurement outcome (e.g., count correct on a reading test) is observed, whereas the measure and instrument calibrations are not observed but can be estimated from the response data and substantive theory, respectively. When an interpretation invoking a predictive mechanism is imposed on the equation, the right-side variables are presumed to characterize the process that generates the measurement outcome on the left side. The symbol=: was proposed by Euler circa 1734 to distinguish an algebraic identity from a causal identity (right hand side causes the left hand side). This symbol (=:) was reintroduced by Judea Pearl and can be read as indicating that manipulation of the right hand side via experimental intervention will cause the prescribed change in the left hand side of the equation. Simple use of an equality (=) does not signal a causal interpretation of the equation.

A Rasch model combined with a substantive theory embodied in a specification equation provides a more or less complete explanation of how a measurement instrument works (Stenner et al. 2009a, b). A Rasch model in the absence of a specified measurement mechanism is merely a probability model. A probability model absent a theory may be useful for describing or summarizing a body of data, and for predicting the left side of the equation from the right side, but a Rasch model in which instrument calibrations come from a substantive theory that specifies how the instrument works is a causal model. That is, it enables prediction after intervention.

Below we summarize two key distinguishing features of causal Rasch models and highlight how these features can contribute to improved ENL, EFL and ESL measurement.

1. First, causal Rasch models are individually centered, meaning that a person's measure is estimated without recourse to any data on other individuals. The measurement mechanism that transmits variation in the language attribute (within person over time) to the measurement outcome (count correct on a reading test) is hypothesized to function the same way for every person. This hypothesis is testable at the individual level using Rasch Model fit statistics.

2. Figuring prominently in the measurement mechanism for language measurement is text complexity. The specification equation used to measure text complexity is hypothesized to function the same way for most text genres and for readers who are ENL, EFL and ESL. This hypothesis is, also, testable at the individual level but aggregations can be made to examine invariance over text types and reader characteristics.

EdSphere™ Reader App

The data for computing empirical text complexity measures came from the reader appliance in EdSphere™. Students access tens of millions of professionally authored digital text by opening EdSphere™ and clicking on the Reader App. Digital articles are drawn from hundreds of periodicals including Highlights for Children, Boys Life, Girls Life, Sports Illustrated, Newsweek, Discovery, Science, The Economist, Scientific American, etc. Such a large repository of high quality informational text is required to immerse students with widely varying reading abilities in daily deliberate practice across the K-16 education experience.

Students use three search strategies to locate articles targeted at their Lexile level: (1) click on suggested topics, (2) click the icon "Surprise Me", or the most frequently used method (3) type search terms into "Find a Book or Article" (see Fig. 1). In the example below, a 1069L reader typed "climate change" in the search box and found 13,304 articles close to her reading level. The first article is an 1100L 4-pager from *Scientific American* with a short abstract.

Readers browse the abstracts and refine the search terms until they find an appropriate length article about their interest topic (or a teacher assigned topic) at their reading level. Within one second of selecting an article, the machine builds a set of embedded semantic cloze items. Students choose from the four options that appear at the bottom of the page. The incorrect options have similar difficulty and part of speech to the correct answer. The answer is auto-scored and the correct answer is immediately restored in the text and color coded as to whether the student answered correctly or incorrectly.

Three instructional supports are built into the Reader App to facilitate comprehension. *First*, suggested strategies are presented to students during the reading process. *Second*, students have access to an in-line dictionary and thesaurus (one click access). Finally, a text-to-speech engine has been integrated into EdSphere, allowing words, phrases or sentences to be machine spoken to the reader.

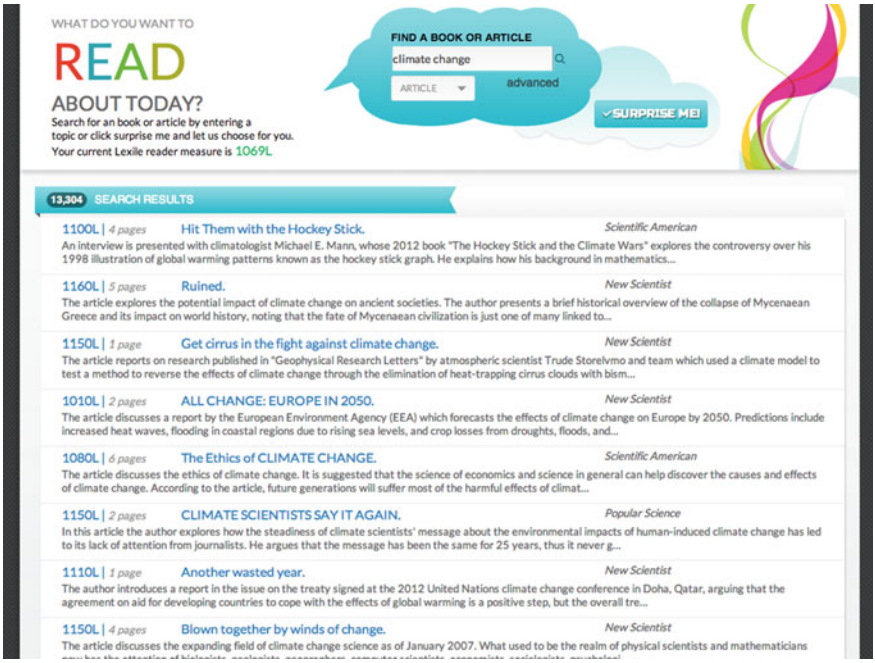

Fig. 1 Results from student's keyword search for articles about climate change; results include, publication titles, publication dates, and/or page length

Text Complexity 719's with Artifact Correction

Figure 2 presents the results of a multiyear study of the relationship between theoretical text complexity as measured by the Lexile Analyzer (freely available for non-commercial use at Lexile.com) and empirical text complexity as measured by the Edpshere™ platform. Each of the 719 articles included in this study was evaluated by the analyzer for semantic demand (log transformed frequency of each word's appearance in a multibillion word corpus) and syntactic demand (log transformed mean sentence length). The text preprocessing, what constitutes a *word*, involves thousands of lines of code. Modern computing enables the measurement of the Bible or Koran in a couple of seconds.

The Edpshere™ platform enables students to select articles of their choosing from a collection of over 100 million articles which have been published and measured over the past 20 years. As a student's reading ability grows a 200L window moves up the scale (100L below the student's ability to 100L above) and all articles relevant to a reader's search term that have text complexity measures in the window are returned to the reader. The machine generates a four choice cloze every 70–80 words and the count correct combined with the readers Lexile measure

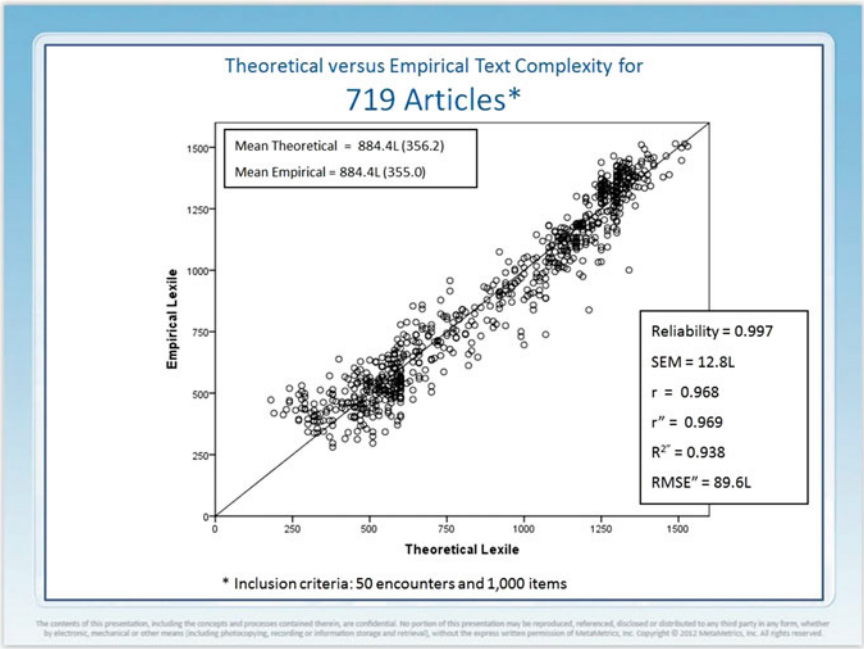

Fig. 2 Plot of Theoretical and Empirical text complexity measures

is used to compute an empirical text complexity for the article averaged over at least 50 readers and at least 1000 items.

The 719 articles chosen for this study were the first articles to meet the dual requirements of at least 50 readers and at least 1000 item responses. Well estimated reader measures were available prior to the encounter between an article and a reader. Thus, each of the articles has a theoretical text complexity measure from the Lexile Analyzer and an empirical text complexity from EdSphere. The correlation between theory and empirical text complexity is r = 0.968 (r^2 = 0.938).

Connecting Causal Rasch Models to theories of language development (Hanlon 2013; Swartz et al. 2015) has made extensive use of Ericsson's theory of deliberate practice in the acquisition of language expertise (Ericsson 1996, 2002, 2006). Deliberate practice is a core tenant of Ericsson's theory of expertise development. Hanlon (2013) distills five core principles of deliberate practice in the development of reading ability: targeted practice reading text that is not too easy and not too hard, (2) real time corrective feedback on embedded response requirements, (3) distributed practice over a long period of time (years, decades), (4) intensive practice that avoids burnout and (5) self-directed options when one on one coaching is not available. Each of these principles, when embedded into instructional technologies, benefits from individually centered psychometric models in which, for example, readers and text are measured in a common unit.

Swartz et al. (2015) provide a complete description of EdSphere, its history and components. The EdSphere Technology is designed to immerse students in deliberate practice in reading, writing, content vocabulary, and practice with conventions of standard English: "These principles of deliberate practice are strengthened by embedding psychometrically sound assessment approaches into learning activities. For example students respond to cloze items while reading, compose short and long constructed responses in response to prompts, correct different kinds of convention errors (i.e. spelling, grammar, punctuation, capitalization) in authentic text, and select words with common meanings from a Thesaurus-based activity. Each item encountered by students is auto-generated and auto-scored by software. The results of these learning embedded assessments are especially beneficial when assessment item types are linked to a developmental scale" (Swartz et al. 2015).

Figure 3 is an individual-centered reading growth trajectory denominated in Lexiles. All data comes from EdSphere. Student 1528 is an ESL seventh grade male (first language Spanish) who read 347 articles of his choosing (138,695 words) between May 2007 and April 2011. Each solid dot corresponds to a monthly average Lexile measure. The growth curve fits the monthly means quite well, and this young man is forecasted (big dot on the far right of the figure) to be a college-ready reader when he graduates from high school in 2016. The open dots distributed around 0 on the horizontal axis are the expected performance minus observed performance (in percents) for each month. Expected performance is computed using the Rasch model and inputs for each article's text complexity and

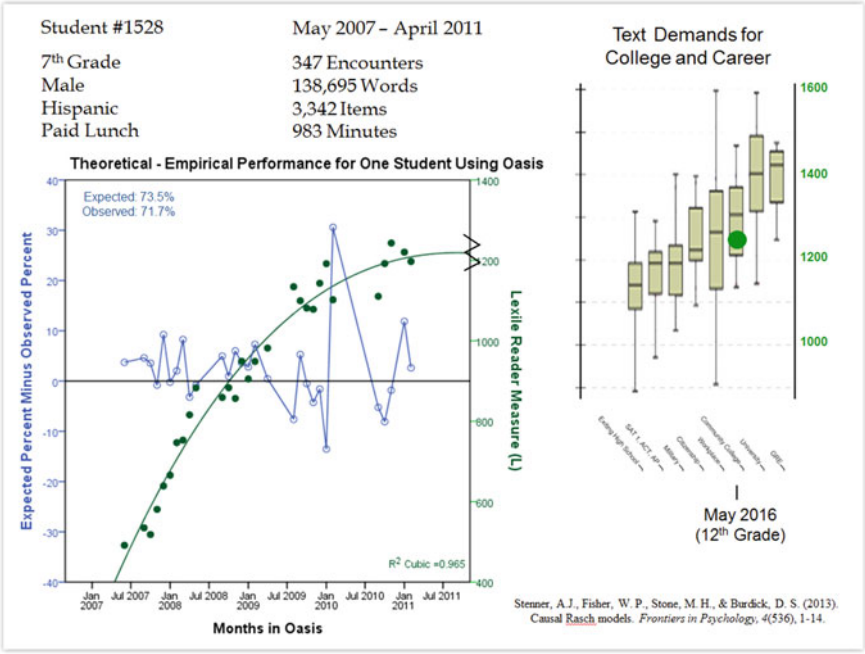

Fig. 3 An individual-centered reading growth trajectory denominated in Lexiles

the updated readers ability measure. Given these inputs, EdSphere forecasts a percent correct for each article encounter. The observed performance is the observed percentage correct for the month. The difference between what the substantive theory (Lexile Reading Framework) in cooperation with the Rasch model expects and what is actually observed is plotted by month. The upper left hand corner of the graphic summarizes the expected percentage correct over the four years (73.5 %) and observed percentage correct (71.7 %) across the 3342 items taken by this reader. Note that EdSphere is dynamically matching text complexity of the articles the reader can choose to the increasing reader ability over time. So, this graphic describes a within-person (intra-individual) test of the quantitative hypothesis: Can EdSphere trade-off a change in reader ability for a change in text complexity to hold constant the success rate (comprehension)? For this reader, the answer appears to be a resounding yes! This trade-off or cancellation affords an intra-personal test of the quantitative hypothesis (Michell 1999).

Figure 4 is a graphical depiction of the 99 % confidence interval for the artifact corrected correlation between theoretical and empirical text complexity. The artifacts included measurement error, double range restriction and construct invalidity. The artifact corrected correlation (coefficient of theoretical equivalence) is slightly higher than r = 1.0 suggesting that the Lexile Theory accounts for all of the true score variation in the empirical text complexity measures. The reader may be puzzled about how a correlation can be higher than r = 1.0, of course it can't be, but

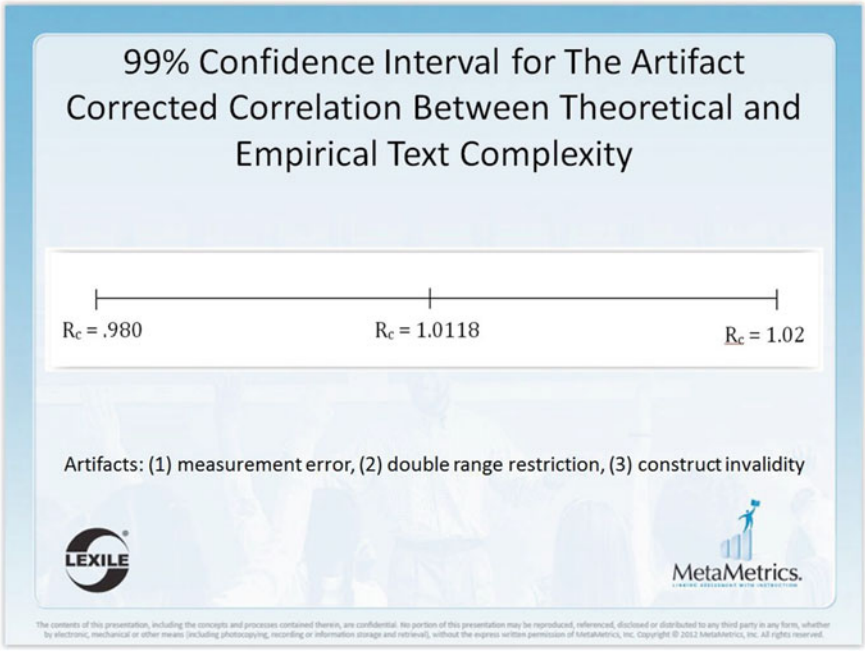

Fig. 4 Artifact corrected correlation between theory observed text complexity

an artifact corrected correlation can be if one or more artifactors used in the process are, perhaps due to a sampling error, lower than their population values.

In the temperature example, a uniform increase or decrease is the amount of soluble additive in each cavity, changes the correspondence table that links the number of cavities that turn black to degrees Fahrenheit or Celsius. Similarly, an increase or decrease in the text demand (Lexile) of the passages used to build reading tests, predictably alters the correspondence table that links count correct to Lexile reader measure. In the former case, a temperature theory that works in cooperation with a Guttman model produces temperature measures. In the latter case, a reading theory that works in cooperation with a Rasch model produces reader measures. In both cases, the measurement mechanism is well understood, and we exploit this understanding to address a vast array of counterfactuals (Woodward 2003). If things had been different (with the instrument or object of measurement), we could still answer the question as to what then would have happened to what we observe (i.e., the measurement outcome). It is this kind of relation that illustrates the meaning of the expression, "There is nothing so practical as a good theory" (Lewin 1951).

Notes

1. The NexTemp® thermometer is a small plastic strip pocked with multiple enclosed cavities. In the Fahrenheit version, 45 cavities arranged in a double matrix serve as the functioning end of the unit. Spaced at 0.2 °F intervals, the cavities cover a range from 96.0 °F to 104.8 °F. Each cavity contains three cholesteric liquid crystal compounds and a soluble additive. Together, this chemical composition provides discrete and repeatable change-of-state temperatures consistent with the device's numeric indicators. Change of state is displayed optically (cavities turn from green to black) and is easily read.
2. Text complexity is predicted from a construct specification equation incorporating sentence length and word frequency components. The squared correlation of observed and predicted item calibrations across hundreds of tests and millions of students over the last 15 years averages about 0.93. Recently available technology for measuring reading ability employs computer-generated items built "on-the-fly" for any continuous prose text in a manner similar to that described for mathematics items by Bejar et al. (2003). Counts correct are converted into Lexile measures via a Rasch model estimation algorithm employing theory-based calibrations. The Lexile measure of the target text and the expected spread of the cloze items are given by theory and associated equations. Differences between two readers' measures can be traded off for a difference in Lexile text measures to hold comprehension rate constant. When the item generation protocol is uniformly applied, the only active ingredient in the measurement mechanism is the choice of text complexity (choosing a 500L article on panda bears) and the cloze protocol implemented by the machine.

References

Andrich, D. (1988) *Rasch models for measurement* (Vols. 07-068). Sage University Paper Series on Quantitative Applications in the Social Sciences, Beverly Hills, California: Sage Publications.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*, 1–16.

Andrich, D. (2010). Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. *Psychometrika, 75*, 292–308.

Bejar, I., Lawless, R., Morley, M., Wagner, M., Bennett, R., & Revuelta, J. A. (2003). feasibility study of on-the-fly item generation in adaptive testing. *The Journal of Technology, Learning, and Assessment 2*(2003), 1–29. http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1663.

Burdick, D., Stone, M., & Stenner, A. J. (2006). The combined gas law and a Rasch reading law. *Rasch Measurement Transactions, 20*, 1059–1060.

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. New York: Oxford University Press.

Ericsson, K. A. (1996). The acquisition of expert performance: An introduction to some of the issues. In K. A. Ericsson (Ed.), *The road to excellence: The acquisition of expert performance in the arts and sciences, sports, and games* (pp. 1–50). Mahwah, NJ: Erlbaum.

Ericsson, K. A. (2002). Attaining excellence through deliberate practice: Insights from the study of expert performance. In M. Ferrari (Ed.), *The pursuit of excellence in education* (pp. 21–55). Hillsdale, NJ: Erlbaum.

Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, P. Feltovich, & R. R. Hoffman (Eds.), *Cambridge handbook of expertise and expert performance* (pp. 683–703). Cambridge, UK: Cambridge University Press.

Fisher, W., Jr. (1997). Physical disability construct convergence across instruments: Towards a universal metric. *Journal of Outcome Measurement, 1*, 87–113.

Fisher, W., Jr. (1999). Foundations for health status metrology: The stability of MOS SF-36 PF-10 calibrations across samples. *Journal of the Louisiana State Medical Society, 151*, 566–578.

Fisher, W., Jr. (2000a). Rasch measurement as the definition of scientific agency. *Rasch Measurement Transactions, 14*, 761.

Fisher, W., Jr. (2000b). Objectivity in psychosocial measurement: What, why, how. *Journal of Outcome Measurement, 4*, 527–563.

Fisher, W., Jr. (2005). Daredevil barnstorming to the tipping point: New aspirations for the human sciences. *Journal of Applied Measurement, 6*, 173–179.

Fisher, W., Jr. (2009). Invariance and traceability for measures of human, social, and natural capital: Theory and application. *Measurement, 42*, 1278–1287.

Fisher W., Jr. (2011). Bringing human, social, and natural capital to life: Practical consequences and opportunities. In N. Brown, B. Duckor, K. Draney, & M. Wilson (Eds.), *Advances in Rasch Measurement* (Vol. 2, pp. 1–27). Maple Grove, Minnesota: JAM Press.

Fisher W., Jr. (2012). What the world needs now: A bold plan for new standards, *Standards Engineering 64*, in press.

Fisher W., Jr. NIST critical national need idea White Paper: Metrological infrastructure for human, social, and natural capital, Retrieved 6 March 2012 from http://www.nist.gov/tip/wp/pswp/upload/202_metrological_infrastructure_for_human_social_natural.pdf. Washington, DC: National Institute for Standards and Technology.

Fisher W., Jr., & Stenner, A. J. Metrology for the social, behavioral, and economic sciences (Social, Behavioral, and Economic Sciences White Paper Series). Retrieved 6 March 2012, from http://www.nsf.gov/sbe/sbe_2020/submission_detail.cfm?upld_id=36, Washington, DC: National Science Foundation.

Fisher, W., Jr., Harvey, R., & Kilgore, K. (1995). New developments in functional assessment: Probabilistic models for gold standards. *NeuroRehabilitation, 5*, 3–25.

Hambleton, R., Swaminathan, H., & Rogers, L. (1991). *Fundamentals of item response theory*. Newbury Park, California: Sage Publications.

Hanlon, S. T. (2013). *The relationship between deliberate practice and reading ability* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses databases (AAT 3562741).

Hunter, J. (1980). The national system of scientific measurement. *Science, 210*, 869–874.

Kuhn, T. S. (1961). The Function of Measurement in Modern Physical Science. *Isis, 52*, 161–193.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. New York: Cambridge University Press.

Lattanzio, S., Burdick, D., & Stenner, A. J. (2012). *The Ensemble Rasch Model*. Durham, NC: MetaMetrics Paper Series.

Lewin, K. (1951). *Field theory in social science: Selected theoretical papers*. New York: Harper & Row.

Markus, K. A. & Borsboom, D. (2013). *Frontiers of Test Validity Theory*. Routledge.

Michell, J. (1999). *Measurement in Psychology*. Cambridge University Press.

Rasch, G. (1960) Probabilistic models for some intelligence and attainment tests (Reprint, with Foreword and Afterword by B. Wright, University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Rasch, G. (1961). *On general laws and the meaning of measurement in psychology, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, IV* (pp. 321–334). Berkeley, California: University of California Press.

Stenner, A. J., & Stone, M. (2010). Generally objective measurement of human temperature and reading ability: Some corollaries. *Journal of Applied Measurement, 11*, 244–252.

Stenner, A. J., Burdick, H., Sanford, E., & Burdick, D. (2006). How accurate are Lexile text measures? *Journal of Applied Measurement, 7*, 307–322.

Stenner, A. J., Burdick, D., & Stone, M. (2008). Formative and reflective models: Can a Rasch analysis tell the difference? *Rasch Measurement Transactions, 22*, 1152–1153.

Stenner, A. J., Stone, M., & Burdick, D. (2009a). The concept of a measurement mechanism. *Rasch Measurement Transactions, 23*, 1204–1206.

Stenner, A. J., Stone, M., & Burdick, D. (2009b). Indexing vs. measuring. *Rasch Measurement Transactions, 22*, 1176–1177.

Stenner, A. J., Fisher, W. P., Stone, M. H. & Burdick, D. S. (2013). Causal Rasch Models. *Frontiers in Psychology, 4*, 1–14.

Swartz, C. W., Hanlon, S. T., Stenner, A. J., & Childress, E. L. (2015). An approach to design-based implementation research to inform development of EdSphere®: A brief history about the evolution of one personalized learning platform. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on computational tools for real-world skill development*. IGI Global: Hersey, PA.

Taagepera, R. (2008). *Making social sciences more scientific: The need for predictive models*. New York: Oxford University Press.

Thurstone, L. L. (1926). The Scoring of Individual Performance. *Journal of Educational Psychology, 17*, 446–457.

Woodward, J. (2003). *Making things happen* (p. 410). Oxford: Oxford University Press. pp. vi.

Wright, B. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97–116.

Wright, B. (1999). Fundamental measurement for psychology. In S. Embretson & S. Hershberger (Eds.), *The new rules of measurement: What every educator and psychologist should know* (pp. 65–104). Hillsdale, New Jersey: Lawrence Erlbaum Associates.