urn:nbn:de:gbv:ilm1-2011imeko-027:0

Joint International IMEKO TC1+ TC7+ TC13 Symposium
August 31st – September 2nd, 2011, Jena, Germany
urn:nbn:de:gbv:ilm1-2011imeko:2

HOW TO MODEL AND TEST FOR THE MECHANISMS THAT MAKE MEASUREMENT SYSTEMS TICK

A. Jackson Stenner[1], *Mark Stone*[2], *and Donald Burdick*[3]

[1]Chairman, CEO, and Co-founder, MetaMetrics, Durham, North Carolina, USA
[2]Distinguished Service Professor, Adler School of Professional Psychology, Chicago, Illinois, USA
[3]Distinguished Research Scientist, MetaMetrics, Durham, North Carolina, USA

Abstract – One must provide information about the conditions under which [the measurement outcome] would change or be different. It follows that the generalizations that figure in explanations [of measurement outcomes] must be change-relating. . . . Both explainers [e.g., person parameters and item parameters] and what is explained [measurement outcomes] must be capable of change, and such changes must be connected in the right way [1]. Rasch's unidimensional models for measurement tell us how to connect object measures, instrument calibrations, and measurement outcomes. Substantive theory tells us what interventions or changes to the instrument must offset a change to the measure for an object of measurement to hold the measurement outcome constant. Integrating a Rasch model with a substantive theory dictates the form and substance of permissible conjoint interventions. Rasch analysis absent construct theory and an associated specification equation is a black box in which understanding may be more illusory than not. The mere availability of numbers to analyze and statistics to report is often accepted as methodologically satisfactory in the social sciences, but falls far short of what is needed for a science.

Keywords: predictive theory, causality, construct validity

1. INTRODUCTION

The vast majority of psychometric thought over the last century has had as its focus the item. Shortly after Spearman's (1904) original conception of reliability as whole instrument replication proved to be difficult when there existed little understanding of what psychological instruments actually measured. The lack of substantive theory made it difficult indeed to "clone" an instrument – to make a genetic copy. In the absence of a substantive theory the instrument maker does not know what features of test items are essential to copy and what features are incidental and cosmetic [2]. So, faced with the need to demonstrate the reliability of psychological instruments but lacking a substantive construct theory that would support instrument cloning early psychometrics took a fateful step inward. Spearman (1910) proposed estimating reliability as the correlation between sum scores on odd and even items of a single instrument. Thus was the instrument lost as a focus of psychometric study and the part score and inevitably the item became ascendant. The spawn of this inward misstep is literally thousands of instruments with non–exchangeable metrics populating a landscape devoid of unifying psychological theory. And, this is so because…"The route from theory or law to measurement can almost never be traveled backwards" [3].

There are two quotes that when taken at extreme face value open up a new paradigm for measurement in the social sciences:

It should be possible to omit several test questions at different levels of the scale without affecting the individuals [readers] score [measure][4].

… a comparison between two individuals [readers] 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 [5].

Both Thurstone and Rasch envisioned a measurement framework in which individual readers could be compared independent of which particular reading items were instrumental for the comparison. Taken to the extreme we can imagine a group of readers being invariantly ordered along a scale when there is not a single item in common. No two readers are exposed to the same item. This would presumably reflect the limit of "omitting" items and making comparisons "independent of the items" used to make the comparison. Compare a fully crossed data collection design in which each item is administered to every reader with a design in which items are nested in persons, i.e. items are unique to each person. Although easily conceived it is immediately clear that there is no data analysis method that can extract invariant reader comparisons from the second design type data. But is this not exactly the kind of data that is routinely generated say when parents report their child's weight on a doctor's office form? No two children (except for siblings) share the same bathroom scale nor potentially even the same underlying technology and yet we can consistently and invariantly order all children in terms of weight? What is different is that the same construct theory for weight has been engineered into each and every bathroom scale even though the specific mechanism (digitally recorded pressure vs. spring driven analog recording) may vary. In addition, the measurement unit (pounds or kilograms) has been consis-

tently maintained from bathroom scale to bathroom scale. So, it is substantive theory and engineering specifications not data that is used to render comparable measurements from these disparate bathroom scales. We argue that this illustrates the dominant distinguishing feature between physical science and social science measurement. Social science measurement does not, as a rule, make use of substantive theory in the ways that the physical sciences do.

Validity theory and practice suffers from an egalitarian malaise, all correlations are considered part of the fabric of meaning and like so many threads each is treated equally. Because we live in a correlated world, correlations of 0.00 are rare, non-zero correlations abound and it is an easy task to collect a few statistically significant correlates between scores produced by virtually any human science instrument and other meaningful phenomena. All that is needed to complete our validity tale is a story about why so many phenomena are correlated with the instrument we are making. And so it goes hundreds and thousands of times per decade, dozens of new instruments are islands unto themselves accompanied by dozens of hints of connectivity whispered to us through dozens of middling correlations. This is the legacy of the nomological network [6]. May it rest in peace!

Validity, for us, is a simple straightforward concept with a narrow focus. It answers the question "What causes the variation detected by the instrument?" The instrument (a reading test) by design comes in contact with an object of measurement (a reader) and what is recorded is a measurement outcome (count correct). That count is then converted into a linear quantity (a reading ability). Why did we observe that particular count correct? What caused a count correct of 25/40 rather than 20/40 or 30/40? The answer (always provisional) takes the form of a specification equation [7] with variables that when experimentally manipulated produce the changes in item behavior (empirical item difficulties) predicted by the theory. In this view validity is not about correlations or about graphical depictions of empirical item orderings called Wright maps [8]. It is about what is causing what? Is the construct well enough understood that its causal action can be specified? Clearly our expectation is unambiguous. There exist features of the stimuli (test or survey items) that if manipulated will cause changes in what the instrument records (what we observe). These features of the stimuli interact with the examinee and the instrument records the interaction (correct answer, strong agreement, tastes good etc.). The window onto the interaction between examinee and instrument is fogged up. We can't observe directly what goes on in the mind of the examinee but we can dissect and otherwise manipulate the item stimuli, or measurement mechanism, and observe changes in recorded behavior of the examinee [9]. Some of the changes we make to the items will matter (radicals) to examinees and others will not (incidentals). Sorting out radicals (causes) from incidentals is the hard work of establishing the validity of an instrument [2]. The specification equation is an instantiation of these causes (at best) or their proxies (at a minimum).

Typical applications of Rasch models to human science data are thin on substantive theory. Rarely is there an a priori specification of the item calibrations (i.e constrained models). Instead the analyst estimates both person parameters and item parameters from the same data set. For Kuhn this practice is at odds with the function of measurement in the "hard" sciences in that almost never will substantive theory be revealed from measurement [3]. Rather "the scientist often seems rather to be struggling with facts [e.g. raw scores], trying to force them to conformity with a theory he does not doubt" [3]. Here Kuhn is talking about substantive theory not axioms. The scientist imagines a world and formalizes these imaginings as a theory and then makes measurements and checks for congruence between what is observed and what theory predicted: "Quantitative facts cease to seem simply the 'given'. They must be fought for and with, and in this fight the theory with which they are to be compared proves the most potent weapon". It's not just that unconstrained models are less potent; they fail to conform to the way science is practiced and most troubling they are least revealing of anomalies [10].

Andrich [10] makes the case that Rasch models are powerful tools precisely because they are prescriptive not descriptive and when model prescriptions meet data, anomalies arise [10]. 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 sets fail to meet the invariance requirements we do not respond by, say, relaxing the invariance requirements through addition of an item specific discrimination parameter to improve fit; rather, we examine the observation model and imagine changes to that model that would bring the data into conformity with the Rasch model requirements.

A causal Rasch model (item calibrations come from theory not the data) is doubly prescriptive [9]. First, it is prescriptive regarding the data structures that must be present:

"The comparison between two stimuli [text passages] should be independent of which particular individuals [readers] were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class [prose] were or might also have been compared. Symmetrically, a comparison between two individuals [readers] should be independent of which particular stimuli within the class considered [prose] were instrumental for [text

passage] comparison; and it should also be independent of which other individuals were also compared, on the same or on some other occasion" [5].

Second, Causal Rasch Models (CRM) [22-23] prescribe that item calibrations take the values imposed by the substantive theory. Thus, the data, to be useful in making measures, must conform to both Rasch model invariance requirements and substantive theory invariance requirements as represented in the theoretical item calibrations. When data meet both sets of requirements then those data are useful not just for making measures of some construct but are useful for making measures of that precise construct specified by the equation that produced the theoretical item calibrations. We note again that these dual invariance requirements come into stark relief in the extreme case of no connectivity across stimuli or examinees. 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 Harry Potter novel and answer questions and I read a Lord of the Rings novel and answer questions, how is it possible that from these disparate experiences an invariant comparison of our reading abilities is realizable? How is it possible that you can be found to read 250L better than I and, furthermore, that you had 95% comprehension and I had 75% comprehension of our respective books. Given that seemingly nothing is in common between the two experiences it seems that invariant comparisons are impossible, but, recall our bathroom scale example, different instruments qua experiences underlie every child's parent reported weight. Why are we so quick to accept that you weigh 50lbs less than I do and yet find claims about our relative reading abilities (based on measurements from two different books) inexplicable. The answer lies in well developed construct theory, instrument engineering and metrological conventions.

Clearly, each of us has had ample confirmation that the construct WEIGHT denominated in pounds and kilograms can be well measured by any carefully calibrated 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 (as described in the Rasch quotes) but 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 reported 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. Finally, by convention (i.e. the written or unwritten practice of a community) in the U.S. we denominate weight in pounds and ounces. The use of pounds and ounces is arbitrary as is evident from the fact that most of the world has gone metric, 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 READING ABILITY does not enjoy a commonly adhered to construct definition, nor a widely promulgated set of instrument specifications nor a conventionally accepted unit of measurement, although, the Lexile Framework for Reading [11] promises to unify the measurement of READING in a manner precisely parallel to the way unification was achieved for LENGTH, TEMPERATURE, WEIGHT and dozens of other useful attributes [21].

A causal (constrained) Rasch model [24] 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 an unconstrained Rasch model. First, with the measurement model and the substantive theory fixed it is self evident that anomalies are to be understood as problems with the data ideally leading to improved observation models that reduce unintended dependencies in the data. Recall that 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. He did not propose parameterizing barometric pressure so that the boiling point of water at sea level would match the model expectations at 3,000 feet above sea level. Second, with both model and construct theory fixed it is obvious that our task is to produce measurement outcomes that fit the (aforementioned) dual invariance requirements. 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. Analogously, finding that not all reading item types 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, anomalous behaviors of various item types 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 as a single class of reading tasks can be regularized by the joint actions of the Lexile theory and a Rasch model is sufficient evidence for the reality of the reading construct.

2. MODEL AND THEORY

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

$$Expected\ raw\ score =: \sum_i \frac{e^{(b-di)}}{1 + e^{(b-di)}}$$

The measurement outcome is the dependent variable and the measure (e.g., person parameter, b) and instrument (e.g., the parameters di 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 parameters 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). The symbol =: exhumed by Judea Pearl can be read as *manipulation of the right hand side via experimental intervention will cause the prescribed change in the left hand side 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 [9]. 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.

Causal models (assuming they are valid) are much more informative than probability models: "A joint distribution tells us how probable events are and how probabilities would change with subsequent observations, but a causal model also tells us how these probabilities would change as a result of external interventions. . . . Such changes cannot be deduced from a joint distribution, even if fully specified." [13]

A satisfying answer to the question of how an instrument works depends on understanding how to make changes that produce expected effects. Identically structured examples of two such narratives include (a) a thermometer designed to take human temperature and (b) a reading test.

2.1. The NexTemp® Thermometer

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 and is easily read.

2.2. The Lexile Framework for Reading®

Text complexity is predicted from a construct specification equation incorporating sentence length and word commonality 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 .93. Available technology for measuring reading ability employs computer-generated items built "on-the-fly" for any continuous prose text. 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. When the item generation protocol is uniformly applied, the only active ingredient in the measurement mechanism is the choice of text complexity.

In the temperature example, if we uniformly increase or decrease the amount of soluble additive in each cavity, we change the correspondence table that links the number of cavities that turn black to degrees Fahrenheit. Similarly, if we increase or decrease the text demand (Lexile) of the passages used to build reading tests, we predictably alter 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 [1]. 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" [12].

3. DISTINGUISHING FEATURES OF CAUSAL RASCH MODELS

Clearly the measurement model we have proposed for human sciences mimics key features of physical science measurement theory and practice. Below we highlight several such features.

1. The model is individual centered. The focus is on explaining variation within person over time.

Much has been written about the disadvantages of studying between person variation with the intent to understand within person causal mechanisms [14, 15]. Molenaar [16] has proven that only under severely restrictive conditions can such cross level inferences be sustained. In general in the human sciences we must build and test individual centered models and not rely on variable or group centered models (with attendant focus on between person variation) to inform our understanding of causal mechanisms. Causal Rasch models are individually centered measurement models. The measurement mechanism that transmits variation in the 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 (the second ergodicity condition of homogeneity) [16]. Note, however, that the fact that there are different developmental pathways that led you to be taller than me and me to be a better reader than you does not mean that the attributes of height and reading ability are somehow necessarily different attributes for both of us.

2. In this framework the measurement mechanism is well specified and can be manipulated to produce predictable changes in measurement outcomes (e.g. percent correct).

For purposes of measurement theory we don't need a sophisticated philosophy of causal inference. For example, questions about the role of human agency in the intervention/manipulation based accounts of causal inference are not troublesome here. All we mean by the claim that the right hand side of Equation 1 causes the left hand side is that experimental manipulation of each will have a predictable consequence for the measurement outcome (expected raw score). Stated more generally all we mean by x causes y is that an intervention on x yields a predictable change in y. The specification equation used to calibrate instruments/items is a recipe for altering just those features of the instrument/items that are causally implicated in the measurement outcome. We term this collection of causally relevant instrument features the "measurement mechanism". It is the "measurement mechanism" that transmits variation in the attribute (e.g. temperature, reading ability) to the measurement outcome (number of cavities that turn black or number of reading items answered correctly).

Two additional applications of the specification equation are: (1) the maintenance of the unit of measurement independent of any particular instrument or collection of instruments [17], and (2) bringing non-test behaviors (reading a Harry Potter novel, 980L) into the measurement frame of reference.

3. Item parameters are supplied by substantive theory and, thus, person parameter estimates are generated without reference to or use of any data on other persons or populations.

It is a feature of the Rasch model that differences between person parameters are invariant to changes in item parameters, and differences between item parameters are invariant to change in person parameters. These invariances are necessarily expressed in terms of differences because of the one degree of freedom over parameterization of the Rash model, i.e. locational indeterminacy. There is no locational indeterminacy in a causal Rasch model in which item parameters have been specified by theory

4. The quantitivity hypothesis [19] can be experimentally tested by evaluating the trade-off property for the individual case. A change in the person parameter can be off-set or traded-off for a compensating change in text complexity to hold comprehension constant. The trade-off is not just about the algebra in Equation 1. It is about the consequences of simultaneous intervention on the attribute (reader ability) and measurement mechanism (text complexity). Careful thinking about quantitivity makes the distinction between "an attribute" and "an attribute as measured." The attribute "hardness" as measured on the Mohs scale is not quantitative but as measured on the Vickers scale (1923) it is quantitative. So, it is confusing to talk about whether an attribute, in and of itself, is quantitative or not. If an attribute "as measured" is quantitative then it can always be represented as merely ordinal. But the obverse is not true. 21st century science still uses the Mohs scratch test which produces more-than-less-than statements about the "hardness" of materials. Pre 1923 it would have been inaccurate to claim that hardness "as measured" was a quantitative attribute because no measurement procedure had yet been invented that produced meaningful differences (the Mohs scratch test produces meaningful orders but not meaningful differences). The idea of dropping, with a specified force, a small hammer on a material and measuring the volume of the resulting indentation opened the door to testing the quantitivity hypothesis for the attribute "hardness". "Hardness" as measured by the falling hammer passed the test for quantitivity and correspondence tables now exist for re expressing mere order (Mohs) as quantity (Vickers).

Michel [19] states "Because measurement involves a commitment to the existence of

quantitative attributes, quantification entails an empirical issue: is the attribute involved really quantitative or not? If it is, then quantification can sensibly proceed. If it is not, then attempts at quantification are misguided. A science that aspires to be quantitative will ignore this fact at its peril. It is pointless to invest energies and resources in an enterprise of quantification if the attribute involved is not really quantitative. The logically prior task in this enterprise is that of addressing this empirical issue (p.75)."

As we have just seen we cannot know whether an attribute is quantitative independent of attempts to measure it. If Vickers company had Michel's book available to them in 1923 then they would have

looked at the ordinal data produced by the Mohs scratch test and concluded that the "hardness" attribute was not quantitative and, thus, it would have been "misguided" and "wasteful" to pursue his hammer test. Instead Vickers and his contemporaries dared to imagine that "hardness" could be measured by the hammer test and went on to confirm that "hardness as measured" was quantitative.

Successful point predictions under intervention necessitate quantitative predictors and outcomes. Concretely, if an intervention on the measurement mechanism (e.g. increase the text complexity of a reading passage by 250L) results in an accurate prediction of the measurement outcome (e.g. how many reading items the reader will answer correctly) and if this process can be successfully repeated up and down the scale then text complexity, reader ability and comprehension (success rate) are quantitative attributes of the text, person and reader/text encounter respectively. Note that, if say, text complexity was measured on an ordinal scale (think Mohs) then making successful point predictions about counts correct based on a reader/text difference would be impossible. Specifically, successful prediction from differences requires that what is being differenced has the same meaning up and down the respective scales. Differences on an ordinal scale are not meaningful (will lead to inconsistent predictions) precisely because "one more" means something different depending on where you are on the scale.

Note that in the Rasch model performance (count correct) is a function of an exponentiated difference between a person parameter and an instrument (item) parameter. In the Lexile Framework for Reading (LF) Equation 1 is interpreted as :

Comprehension = Reader Ability – Text Complexity
(success rate)

The algebra in Equation 1 dictates that a change in reader ability can be traded-off for an equal change in text complexity to hold comprehension constant. However, testing the "quantitivity hypothesis" requires more than the algebraic equivalence in a Rasch model. What is required is an experimental intervention/manipulation on either reader ability or text complexity or a conjoint intervention on both simultaneously that yields a successful prediction on the resultant measurement outcome (count correct). When manipulations of the sort just described are introduced for individual reader/text encounters and model predictions are consistent with what is observed the quantitivity hypothesis is sustained. We emphasize that the above account is individual centered as opposed to group centered. The LF purports to provide a causal model for what transpires when a reader reads a text. Nothing in the model precludes averaging over readers and texts to summarize evidence for the "quantitivity hypothesis" but the model can be tested at the individual level. So, just as pressure and volume can be traded off to hold temperature constant or volume and density can be traded off to hold mass constant so can reader ability and text complexity be traded off to hold comprehension constant. Following Michel [19] we note that a trade-off between equal increases (or decrements) in text complexity and reader ability "identifies equal ratios directly" and "Identifying ratios directly via trade-offs results in the identification of multiplicative laws between quantitative attributes. This fact connects the theory of conjoint measurement with what Campbell called derived measurement" [19].

Garden variety Rasch models and IRT models are in their application purely descriptive. They become causal and law like when manipulations of the putative quantitative attributes produce changes (or not) in the measurement outcomes that are consistent with model predictions. If a fourth grade reader grows 100L in reading ability over one year and the text complexity of her fifth grade science textbook also increases by 100L over the fourth grade year textbook then the forecasted comprehension rate (whether 60%, 70%, or 90%) that that reader will enjoy in fifth grade science remains unchanged. Only if reader ability and text complexity are quantitative attributes will experimental findings coincide with these model predictions. We have tested several thousand students' comprehension of 719 articles averaging 1150 words. Total reading time was 9794 hours and the total number of unique machine generated comprehension items was 1,349,608. The theory based expectation was 74.53% correct and the observed 74.27% correct.

4. CONCLUSION

This article has considered the distinction between a descriptive Rasch model and a causal Rasch model. We have argued for the importance of measurement mechanisms and specification equations. The measurement model proposed and illustrated (using NexTemp thermometers and the Lexile Framework for Reading) mimics in several important ways physical science measurement theory and practice. We plead guilty to "aping" the physical sciences and despite the protestations of Michell [19] and Markus and Boorsboom [20] do not view as tenable any of the competing go forward strategies for the field of human science measurement.

REFERENCES

[1] J. Woodward, "Making things happen", Oxford University Press, New York, NY, 2003.

[2] S.H. Irvine, P.C. Kyllonen, *Item generation for test development*, Lawrence Erlbaum Associates, Inc, Mahwah, New Jersey, 2002.

[3] T.S. Kuhn, "The function of measurement in modern physical science", *Isis*, **52**(168), 161-193, 1961.

[4] L.L Thurstone, "The scoring of individual performance", *Journal of Educational Psychology*, **17**, 446-457, 1926.

[5] G. Rasch, "On general laws and the meaning of measurement in psychology", *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, IV.*, University of California Press, Berkeley, California, 321-334, 1961.
[6] L.J. Cronbach, P.E. Meehl, "Construct validity in psychological tests", *Psychological Bulletin*, **52**, 281-302, 1955.
[7] A. J. Stenner, M. Smith, D. Burdick, "Toward a theory of construct definition", *Journal of Educational Measurement*, **20**(4), 305-316, 1983.
[8] M. Wilson, *Constructing measures: An item response modeling approach,* Lawrence Erlbaum Associates, Inc, Mahwah, New Jersey, 2004.
[9] A.J. Stenner, M.H. Stone, D. Burdick, "The Concept of a Measurement Mechanism", *Rasch Measurement Transactions*, **23** (2), 1204-1206, 2009.
[10] D. Andrich, "Controversy and the Rasch model: a characteristic of incompatible paradigms?", *Medical Care*, 42, 1-16, 2004.
[11] A.J. Stenner, H. Burdick, E. Sanford, D.S. Burdick, "How accurate are Lexile text measures?", *Journal of Applied Measurement*, **7**(3), 307- 322, 2006.
[12] K. Lewin, *Field theory in social science: Selected theoretical papers*, Harper & Row, New York, NY, 1951.
[13] J. Pearl, "Causality: Models, reasoning, and inference", Cambridge University Press, Cambridge, MA, 2000.
[14] J.W. Grice, *Observation oriented modelling*, Elsevier, New York, NY, 2011.
[15] D.H. Barlow, M.K. Nock, M. Hersen, *Single case experimental designs (3rd ed.)*, Pearson, Boston, MA, 2009.
[16] P.C.M. Molenaar, "A manifesto on psychology as ideographic science: Bringing the person back into scientific psychology, this time forever", *Measurement: Interdisciplinary Research and Perspective*, **2**, 201-218, 2004.
[17] A.J Stenner, D.S. Burdick, "Can psychometricians learn to think like physicists?", *Measurement*, **9**, 62-63, 2011.
[18] G. Karabatsos, "The Rasch Model, additive conjoint measurement, and new models of probabilistic measurement theory", *Journal of Applied Measurement*, **2**, 389–423, 2001.
[19] J. Michell, *Measurement in psychology: A critical history of a methodological concept*, Cambridge University Press, New York, NY, 1999.
[20] K.A. Markus, D. Borsboom (2011). "Reflective measurement models, behavior domains, and common causes", *New Ideas in Psychology*, In Press, 2011
[21] A.J. Stenner, M.H. Stone, "Generally objective measurement of human temperature and reading ability: Some corollaries", *Journal of Applied Measurement*, **11**(3), 244-252, 2010.
[22] D.S. Burdick, M.H. Stone, A.J. Stenner, "The combined gas law and a Rasch reading law", *Rasch Measurement Transactions*, **20**(2), 1059-1060, 2006.
[23] A.J. Stenner, D.S. Burdick, M.H. Stone, "Formative and reflective models: Can a Rasch analysis tell the difference?", *Rasch Measurement Transactions*, **22**(1), 1152-1153, 2008.
[24] A.J. Stenner, M.H. Stone, D.S. Burdick, "Indexing vs measuring", *Rasch Measurement Transactions*, **22**(4), 1176-1177, 2009.

Authors:
A. Jackson Stenner, MetaMetrics, Inc. (27713, Durham, North Carolina,USA) +1-919-547-3402, jstenner@lexile.com.
Mark Stone, Adler School of Professional Psychology (60602, Chicago, Illinois, USA) +1-312-662-4000, mstone@adler.edu.
Donald Burdick, MetaMetrics, Inc. (27713, Durham, North Carolina, USA) +1-919-547-3400, dburdick@lexile.com.