Building a Productive Trading Zone in Educational Assessment Research and Practice

William P. Fisher Jr. and Mark Wilson

University of California, Berkeley, USA

Post to:
William P. Fisher Jr.
University of California, Berkeley, USA
Address first author:
Email: wfisher@berkeley.edu

Abstract

Markedly diverse viewpoints are in play in many varied contexts, from science to the classroom to the marketplace. Perhaps surprisingly, both divergent dissonances and convergent harmonies are routinely found together in productive real-world systems. The value of generalizable assessment outcomes hinges on their being both innovative and standardized. These apparently opposite tensions can be reconciled in terms of boundary objects, entities shared by different communities that use and view them quite differently. Further, science has long been seen as taking place on a continuum from everyday thinking and acting to formal logic and methods, so it should not be surprising to find this range manifest as well in psychometric research. We describe methods and results in which psychometrically modeled exemplars known as construct maps and Wright maps function as boundary objects and serve as a basis for productive analogies in educational assessment by (a) preserving relational structures, (b) making isomorphic mappings between systems, and (c) facilitating systematicity, understood as mapping systems of higher order relational structures (Nersessian & Chandrasekaran, 2009). In this conceptual context, we present an application of the BEAR Assessment System and its accompanying software, facilitating translations of relational structures across systems in support of practical alliances of teaching, policy-making, assessment and curriculum development, psychometrics, and information technology (IT).

Keywords: Please include them.

Psychology and the social sciences are marked by their lack of consensus on measurement methods and standards, and, not coincidentally, by a superficial «cookbook» approach to measurement applications. Works emerging over a period of decades (Bakker, van Dijk, & Wicherts, 2012; Berkson, 1938; Bolles, 1962; Coats, 1970; Cohen, 1994; Guttman, 1985; Michell, 1999; Roberts, 1994; Taagepera, 2008; Wilson, 1971) document alarmingly wide ranges of variation in the methods deemed acceptable, disagreements about basic concepts, and broad indifference to the scientific and practical advantages of approaches requiring experimental tests and theoretical predictions instead of unexamined assumptions and purely empirical descriptions. The importance of measurement is universally recognized in principle, and soundly-based and practicable methods have been available in the research literature, textbooks, and software for decades. However, in practice, it is rare to find tests, assessments, or surveys designed with explanatory theories of the construct measured, calibrated in invariant units with estimable uncertainties, and informed by qualitative interpretations of both consistent and inconsistent response patterns. The overall, and quite dismaying, impression gained from a review of this literature is that, even though much good work has been done, including mathematical proofs, inspired teaching, readily available software, simple and practical methods, reproduced empirical results, and persuasive theoretical explanations, this sum total has not been effective in dispelling the fundamental disagreements about social science measurement mentioned above, nor in successfully promoting a framework that embodies the three important qualities that are needed (also mentioned above).

Perhaps one way towards an answer is to look beyond the traditional limits of «measurement» to consider the broader scope within which the measures operate. In this regard, it is relevant to note that, over the last several decades, historical and social studies of science (Galison, 1999; Hutchins, 1995, 2014; Latour, 1987, 1993a; Nersessian, 2006, 2012) have effectively revived the question that Hayek (1948, p. 54) raised as the «central question of all social sciences: How can the combination of fragments of knowledge existing in different minds bring about results which, if they were to be brought about deliberately, would require a knowledge on the part of the directing mind which no single person can possess?» This question arose for Hayek (1948, p. 88) in conjunction with Whitehead's (1911, p. 61) observation that

civilization does not advance via original thinking so much as it does by means of technologies that enable people to successfully execute operations they do not understand and could not accomplish by themselves. No single person, for instance, has the capacity to put automobiles on the road, to bring electric light into homes, or to make a commodities market function. Each of these advances comes about with no single director, through the combined efforts of persons with wide varieties of expertise, from laborers, suppliers, clerks, and consumers to scientists, engineers, and mathematicians to financiers and economists to educators and legislators. Detailed descriptions of how these very diverse groups have historically coordinated and aligned their activities in productive applications of science and technology (Galison, 1997; Latour, 1993a; Miller & O'Leary, 2007; Star & Griesemer, 1989) suggest new possibilities for improving the quality of methods employed in psychology and the social sciences (Fisher, 2000, 2005, 2009; Fisher & Stenner, 2011, 2013a).

Boundary objects for measurement: The construct map and the Wright map

One way in which contemporary history and philosophy of science frames these issues is in terms of *trading zones* (Galison, 1997, 1999; Galison & Stump, 1996) and the *boundary objects* implicated in translation networks (Star & Griesemer, 1989; Woolley & Fuchs, 2011) within those zones. Trading zones are forums in which ideas can be safely exchanged in a manner akin to commercial neutral grounds noted by ethnographers as emerging between unfriendly neighbors who have products they wish to buy and sell. The two groups often invest the objects traded with entirely different meanings and values. Such objects, residing as they do at the boundaries between different groups, are termed boundary objects, and are defined in terms of translations based in systematically structured analogies.

This broad perspective on social science research suggests that we should raise the question here of what such an approach might look like in psychometric research; in other words, what boundary objects might be suitable to ground a diverse array of players, such as those mentioned above, for educational assessment: teachers, assessment developers, psychometricians, information technology (IT) experts, curriculum developers, policy-makers, and others? In this paper, we describe one particular approach to the establishment of a trading zone using boundary objects —based on the boundary-riding concepts of the *construct map*

(Wilson & Sloane, 2000) and the *Wright map* (Wright & Stone, 1979), and a way of organizing them referred to as *construct modeling* (Wilson, 2005).

First, a construct map provides a concrete representation of the theoretical expectations (hopefully buttressed by empirical outcomes) for the nature of the construct in an assessment or survey. Construct maps define particularly useful points along a continuum of the construct, providing a coherent and substantive definition for the content of the assessment (Wilson, 2005). A construct map is a well-thought-out, theoretically justified, and researched ordering of qualitatively different points of performance focusing on one characteristic. Construct maps are derived in part from research into the underlying structure of the domain and in part from professional judgments about what constitutes higher and lower levels of performance or competence, but are also informed by empirical research into how people think and act in practice (Kilpatrick, Swafford, & Findell, 2001).

An example of a construct map is shown in Table 1, which was developed as part of a chemistry assessment project at the University of California, Berkeley called «Perspectives of Chemists» (Claesgens, Scalise, Wilson, & Stacy, 2009). The project attempted to embody understanding of chemistry from a novice to expert level of sophistication in the form of a set of construct maps, and Table 1 shows the *Matter* strand, which is concerned with describing atomic and molecular views of matter. Note, however, that it is not a typical content description, as found in most textbooks, but instead describes how a student's view of matter progresses from a continuous, real-world view, to a particulate view, and then builds in sophistication. There are many ways to display a construct map and this one includes almost all the typical components, reading from the left: a label for each level, then a summary of the content that students are addressing at that level, followed by a description of the student thinking at that level, and completed by examples of items that a student at that level might be asked to attempt. Some maps go beyond this and also include examples of student responses to the items for each level. Clearly, the existence of a construct map does indeed satisfy the first part of our three criteria, that the design should include explanatory theories of the construct. However, this alone is not enough —it is not even the complete first part, as a construct map does not in itself provide a

measure of the construct. That will be represented by the other boundary object, the Wright map.

Table 1
Perspectives of chemists framework, *matter* variable

Level of success	Big ideas	Descriptions of level	Item exemplars
10-12 Construction Why composition, structure, properties, and amounts? (Using models)	The composition, structure, and properties of matter are explained by varying strengths of interactions between particles (electrons, nuclei, atoms, ions, molecules) and by the motions of these particles.	Students are able to reason using normative models of chemistry, and use these models to explain and analyze the phase, composition, and properties of matter. They are using accurate and appropriate chemistry models in their explanations, and understand the assumptions used to construct the models.	a) Composition: How can we account for composition? b) Structure: How can we account for 3-D structure? (e.g., crystal structure, formation of drops,) c) Properties: How can we account for variations in the properties of matter? (e.g., boiling point, viscosity, solubility, hardness, pH, etc.) d) Amount: What assumptions do we make when we measure the amount of matter? (e.g., non-ideal gas law, average mass)
7-9 Formulation How can we think about interactions between molecules? (Multirelational)	The composition, structure, and properties, of matter are related to how electrons are distributed among atoms.	Students are developing a more coherent understanding that matter is made of particles and the arrangements of these particles relate to the properties of matter. Their definitions are accurate, but understanding is not fully developed so that student reasoning is limited to causal instead of explanatory mechanisms. In their interpretations of new situations students may over-generalize as they try to relate multiple ideas and construct formulas.	a) Composition: Why is the periodic table a roadmap for chemists? (Why is it a «periodic» table?) How can we think about the arrangements of electrons in atoms? (e.g., shells, orbitals) How do the numbers of valence electrons relate to composition? (e.g., transfer or sharing) b) Structure: How do connections between atoms (bonds) and motions of atoms explain 3-D structure? (diamond rigid, water flows, air invisible) c) Properties: How can matter be classified according to bonds? (ionic solids dissolve in water, covalent solids hard, matter phase) d) Amount: How can one quantity of matter be related to another? (e.g., mass/mole/number, ideal gas law, Beer's law)
4-6 Recognition How do chemists describe matter? (Unirelational)	Matter is categorized and described by various types of subatomic particles, atoms, and molecules.	Students explore the language and symbols used by chemists to describe matter. They relate numbers of electrons, protons, and neutrons to elements and mass, and the arrangements and motions of atoms to composition and phase. Ways of thinking about matter are limited to relating one idea to another at a simplistic level of understanding.	a) Composition: How does the periodic table show trends? How are elements, compounds, and mixtures classified by letters and symbols? b) Structure: How do the arrangements and motions of atoms differ in solids, liquids, and gases? c) Properties: How can the periodic table be used to predict properties? d) Amount: How do chemists keep track of quantities of particles? (e.g., number, mass, volume, pressure, mole)
1-3 Notions What do you know about matter?	Matter has mass and takes up space. It can be classified according to how it occupies space.	Students articulate ideas about matter, and use experience, observation and logical reasoning to provide evidence. Focus is largely on macroscopic (not particulate).	a) Composition: How is matter distinct from energy, thoughts, feelings? b) Structure: How do solids, liquids, and gases differ from one another? c) Properties: How can you use properties to classify matter? d) Amount: How can you measure the amount

			of matter?

Second, for a construct map, which is an expression of an intention, to be useful for measurement, it must be augmented empirically, producing another version of the map referred to as a Wright map. Empirical support requires development of (a) a set of items that embody the construct, in terms of a person's responses to the items; (b) a plan for transforming those responses into data; and (c) a method of calibrating an instrument from those item responses so that they can be used as the empirical representation of the construct. In the particular approach to this effort that we describe, the empirical representation of the construct is termed the Wright map.

A Wright map, corresponding to the construct map shown in Table 1, is shown in Figure 1, where the left side of this map shows the measured distribution of students who responded to the *Matter* items, and the right side shows the calibrated difficulty of a subset of six of the tasks. Estimation of these student and item locations was based on the Rasch model (Rasch, 1960), which provides the mapping of both the items and the students onto the same logit scale. As can be seen, the student responses have been separated into two separate segments of the logit scale, thus facilitating the allocation of students below (approximately) -0.05 logits to level 1 on the construct map, i.e., to the *Notions* level, and the students above that (highlighted) to level 2 on the construct map, the *Recognition* level. For this particular group of students, only the first two levels of the map were extant. The analysis also generates fit statistics to flag inconsistent response vectors, and other indices for how well levels specified by the model fit the data. Tables of reliability coefficients and standard errors are generated, and inter-rater comparisons can also be made. The Wright map is the result of the successful Rasch scaling of the construct, and this, in combination with the construct map, does indeed satisfy the other portions of our trio of important qualities: that the instrument be designed with explanatory theories of the construct measured, calibrated in invariant units with estimable uncertainties, and can be informed by qualitative interpretations of both consistent and inconsistent response patterns. It is important to note that the successful Rasch scaling itself is insufficient —it is only when the two boundary objects are in synchronicity (as was illustrated in Table 1) that the three qualities are satisfied.

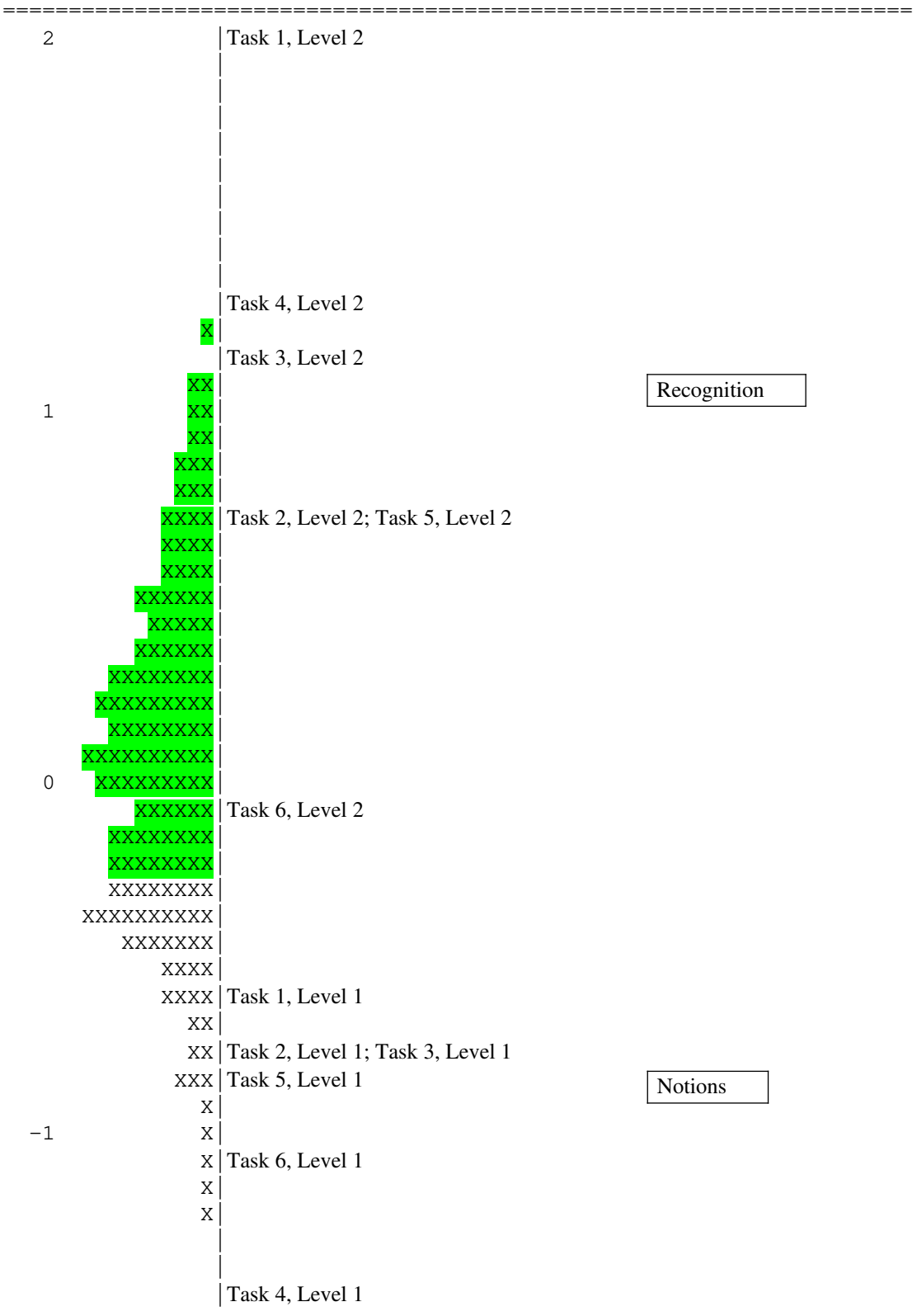

Figure 1. Wright Map of the ChemQuery Matter Variable (Partial Credit, Generalized-Item Thresholds). *Numbers at left are in a measurement unit called the «logit», or log of the odds, with higher numbers indicating better student

performance.

Third, these two representations are coordinated together in a cycle known as construct modeling, which is illustrated in Figure 2. In this figure, there are two intermediate steps between the construct map and the Wright map, specifically, the *items design* and the *outcome space*, corresponding (a) to the design of items that are intended to engender responses that can be interpreted as being indicative of specific levels of the construct map, and (b) to the schema for the valuing of those responses into the construct map levels (and possibly other categories as well). This cycle of instrument development iterates until sufficient consistency is reached between the intentions and the empirical results (see Wilson (2005) for details on this), and the instrument is then ready for the investigation of its reliability and validity evidence, and, eventually, for direct use.

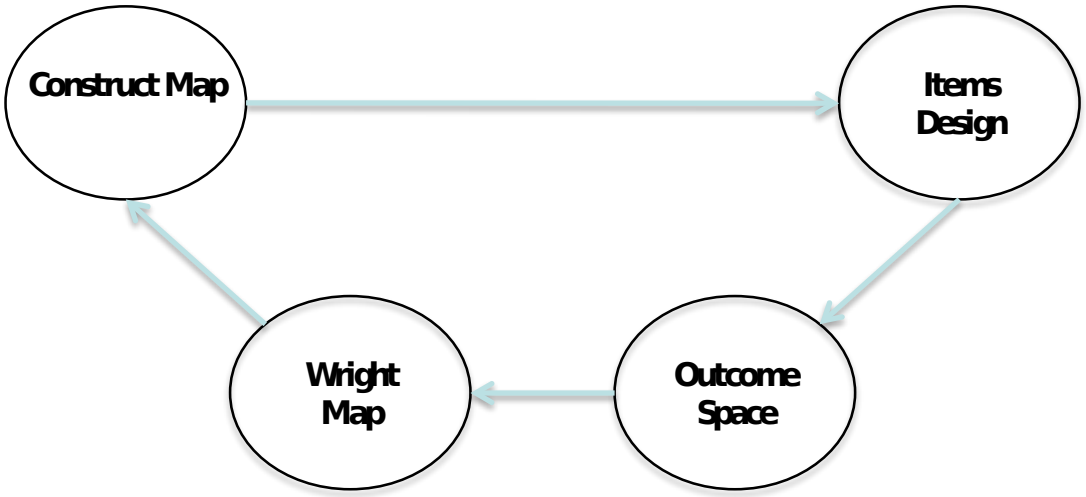

Figure 2. Construct modeling.

Images such as these have proven valuable in improving the reliability, validity, and utility of psychometric measures across a wide range of fields, from education (Black, Wilson, & Yao, 2011) to health care (Best, 2008; Ewert, Allen, Wilson, Üstün, & Stucki, 2010; Smith, 2005; Wilson, Allen, & Li, 2006) and psychology (Dawson, Xie, & Wilson, 2003; Kaiser & Wilson, 2000). Successes to date prompt further inquiry as to if and how construct maps may

function as boundary objects in trading zones, and how their use in this regard might be expanded and enhanced.

These boundary objects can be seen as being placed within a trading zone, as shown in Figure 3, illustrating examples of specific passage points and allies for the context of educational assessment, as it is envisaged in the construct modeling approach. This is an instantiation of the Star and Griesemer (1989) figure shown in Figure 4, although it is somewhat more complex, as it affords two boundary objects rather than one, in the manner of Nersessian's (2012, p. 227) figure showing the interconnections between a lab's problems, researchers, and strategies.

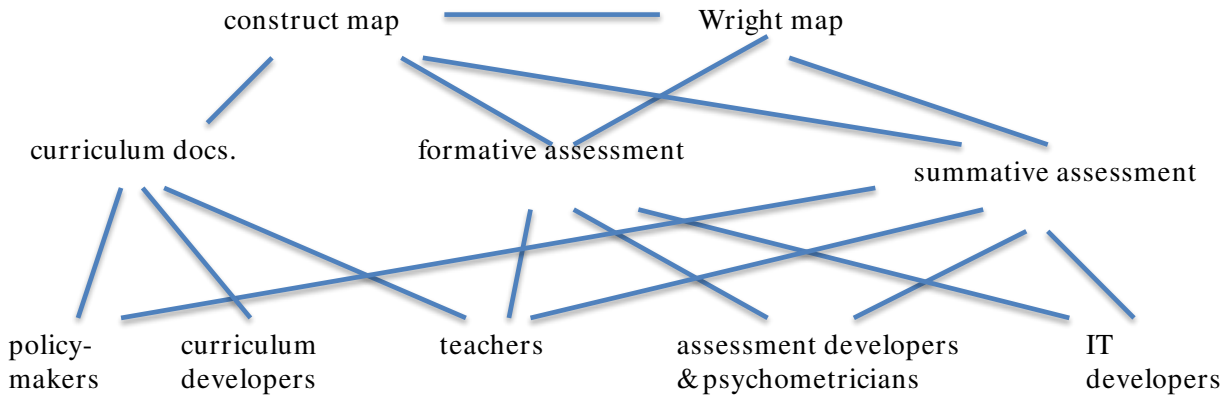

Figure 3. A translation network with two boundary objects.

For example, examining Figure 4, we can see how the two boundary objects can inform the whole process of instruction and assessment. As an example of an episode from this domain, suppose that a teacher is using the results of an assessment to plan the next step in instruction. Now, the background for such an event is complex, but one can surmise (following from left to right in Figure 4) that the topics being taught were decided upon through a policy-development process among policy-makers, curriculum developers and teachers (among others), and this led to a set of curriculum documents. In this process, construct maps (perhaps deployed in complicated relationships within a learning progression) can provide a key communication tool for the curriculum developers and teachers. Further deep background is evident in that (looking to the right-hand side of Figure 4) the assessments the teacher might use to inform instruction (i.e., formative assessments) had been produced by assessment developers, following the

construct maps, which were developed by assessment experts and psychometricians. Aided by the psychometricians, the curriculum developers devised a method for scoring responses that is sensitive to potential misconceptions as well as to establishing prerequisites for individualized instruction. Of course, the teacher will want to assess whether the students had learned what had been taught, and for this the teacher would want to use a summative assessment.

This can be made most powerful through a systematic connection between the two forms of assessment, and this is provided by the Wright map, on which can be displayed not only the formative and summative assessment items, but also student locations on these two sets of items. Note that the policy-makers will have a hand in specifying (at least the motivation for) the summative assessments, given their legitimate concern for monitoring student progress. Each stakeholder in the process has expertise in one or more complementary areas, and knows little or nothing about the technical aspects of others' contributions. Even if an individual has multiple areas of expertise, there would be insufficient time and resources for to accomplish the entirety of what is routinely done through the coordination and alignment of the field-organizing activities.

The trading zone, boundary objects and translations via analogy

In science, boundary objects, often represented as images of various kinds (Daston, 2004; Daston & Galison, 1992, 2007; Dear, Hacking, Jones, Daston, & Galison, 2012; Galison, 2008; Ihde, 1998, 2012), are adaptable and translatable across different perspectives. No one point of view dominates all of the others, since the effectiveness of each perspective's operational definitions is realized only in terms of the overall social projection of the object within that perspective. Scientists think and act in relation to boundary objects in ways not qualitatively different from the ways children learn through play (Nersessian, 1996) and the ways everyday thinking and acting relate to conversational objects (Nersessian & Chandrasekaran, 2009). In science, children's play, and everyday thinking, reasoning is model-based and situated in distributed networks facilitating active imitative analogies via trial and error (Nersessian, 2002, 2006, 2008, 2012).

What does this mean for psychology and education? Galison's (1999) concept of the trading zone as an area in which ideas can be exchanged and value can be obtained without

assuming reduction to a common universal point of view provides an apt metaphor for organizing a positive program of research and practice in psychology and the social sciences. Each particular neighborhood in such a community may include people sharing a general perspective on a process, outcome, or goal. Consider the field of assessment in education, where significant types of role-players, such as teachers, assessment developers, psychometricians, IT experts, curriculum developers, policy-makers, and others, each have specific interests that, while they are related, differ in important ways.

The problem is how these interests might be jointly advanced more effectively than current methods allow. A significant clue as to how this problem might be approached lies in the difficulties experienced in educating end users, such as teachers, theoreticians, or experimental researchers, about the technical specifics of psychometrics. Communicating the complexity of the models, the estimation processes, the uncertainty and model fit evaluations, and the interpretation of the results can cause considerable frustration for all parties. Conversely, psychometricians may find the problems experienced by teachers or formulated by substantive researchers to be as incomprehensible as the latter groups may find the math. In this case, what often happens, instead of mutually-informative dialogues between the various groups, is something «like the parallel play that Piaget described in preschoolers: They talk (and play) in each other's company rather than *to* and *with* each other» (Bond & Fox, 2015, p. 299). The developmental analogy here raises the question as to whether researchers can be expected to mature in a manner akin to the way preschoolers eventually come to share in more reciprocal relationships. Might there be another way of mediating relationships between different areas of technical expertise without expecting the kind of shared mastery exhibited when everyone is fluent in the same language game? Following Nersessian's (1996) sense of the playful learning through trial and error that both children and scientists engage in, might there be a productive developmental pathway for educational research stakeholders to follow that does not necessarily involve more talking (and playing) to and with each other than is already the norm?

Clues as to how this problem might be addressed are provided by Galison's (1997) overview of the philosophical paradigm shifts of recent decades. Historically, in education as in many other fields, the objective status of data has been assumed to provide a compelling

rationale and logical basis for coordinating activities across the somewhat convergent and somewhat divergent perspectives of the various stakeholder groups. As has been abundantly demonstrated in the anti-positivist literature (Kuhn, 1970; Toulmin, 1961, 1982), this positivist or modern stress on data as primary rarely conforms to actual practice. Attention can be focused on data only insofar as some theoretical concept is deployed —i.e., data becomes «data», only in light of a theoretical expectation. Even if no explicit theory is available, sharing and communicating what has been noticed about things in the world requires ideas and concepts stable enough to mediate relationships in a regular, predictable (though perhaps revisable) way. The anti-positivist perspective, in turn, has been found deficient by being locked into a relativistic stance emphasizing historical, cultural, and linguistic dependencies, which is ultimately counterproductive (Latour, 1991).

The post-positivist (Galison, 1999), unmodern (Dewey, 2012), or amodern (Latour, 1990, 1991, 1993b, 2010) perspective offers a new alternative that conceptualizes calibrated instruments as being brought to bear relative to *both* data and theory, making it possible to accommodate different perspectives and interests without compromising the pragmatic need to articulate a logical and productive program of research and development. Instruments measuring in the shared language of a common framework make phenomena reproducible and available for close study, while also contextualizing the positive value of anomalous observations, or the lack thereof. Instruments designed and interpreted as embodied boundary objects then function as a kind of contextualized, inchoate, or potential universal (Ricoeur, 1992, p. 289), an «embedded relationality» that serves as «the prophylaxis for both relativism and transcendence» (Haraway, 1996, pp. 439-440). Though theory, instruments, and experimental data change over time, at varying tempos, local communications frame individual experiences relative to global standards in a manner analogous to the way everyday language works. Translations between different groups' practical understandings of phenomena take place via analogies in the trading zones at the boundaries separating the different communities where strategic alliances are formed.

Figure 4 (adapted from Star and Griesemer, 1989, p. 390) provides a schematic view of how different stakeholder groups ally themselves relative to shared purposes via translations and boundary objects. This view modifies the perspective originally advanced by Callon (1985),

Latour (1987, 1993a), and Law (1985). The untestable assumption concerning the existence of a single overarching abstract object, termed an «immutable mobile» by Latour (1987), is dropped in favor of a Quinean pragmatist focus on the observable decision points effected via translation into each group's vocabulary, concepts, and processes. Diverse stakeholders participate concurrently in organizing a field, even though the objects of inquiry specified in each field inhabit separate social worlds. Boundary objects are adaptable across these worlds to the extent that the methods through which they are brought into play via translation are standardized within each stakeholder group's processes. Translation cannot be effected in a loose, laissez-faire kind of relativism, but requires the rigor of repeated reconstructions of the common object, where each translation incorporates elements of all the others. The interests of each group are operationalized in and advanced by the coordinated processes to the extent they are transparently incorporated into all of the rest as obligatory participants in the overall community.

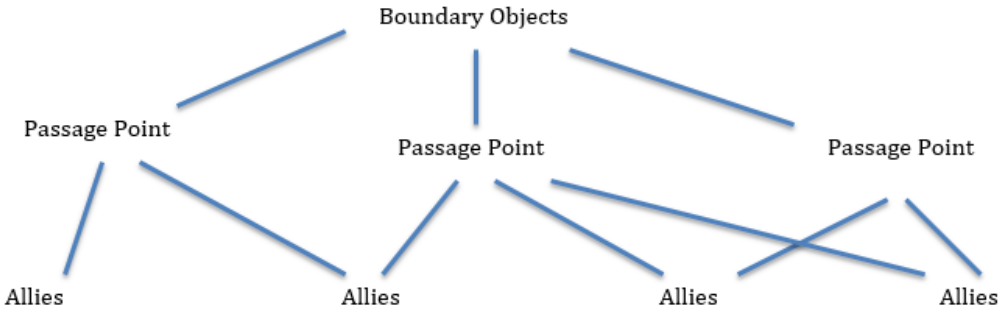

Figure 4. Trading zone alliances and translations framing a field's collective intelligence (adapted from Star and Griesemer, 1989, p. 390).

Star and Griesemer (1989, p. 390) thus note that the coherence of independent sets of translations must be indifferent to the particular processes producing them. The indefinite numbers of ways in which actors in each group of stakeholders might make their work necessary to the other groups results in an indeterminate number of possible translations. Our translation of what is meant by the *coherence* of boundary objects within the measurement domain is in terms effectively equivalent to requiring measurement invariance in a psychometric model. New possibilities for advancing educational research and practice are made apparent in the contrast of

positivist and post-positivist approaches to the way different groups' interests are translated to each other.

For instance, education is premised on the idea that training in particular problems in the classroom can be effective despite variation in realizing that ideal across students, teachers, schools, tests, and curricula, and despite the fact that no particular set of problems can ever represent every possible problem that might be encountered in the real world. Though circumstances are changing, contemporary educational research and practice tends to define boundary objects and standardized processes in two rather different ways. The most traditional way is to use locally-developed assessments and locally-determined curriculum content. The paradigm is positivist, in the sense that the objective facts (counts and percentages) of correct and incorrect responses to particular sets of assessment questions are deemed sufficient to the determination of results. Translation is thus encumbered with near-insurmountable problems in comparability, since the meaning of scores changes in unknown ways with each assessment content and context, and across students and curricula. The expense and trouble that would be encountered in trying to translate locally-defined test scores across classrooms, grades, schools, districts, and regions would far outweigh any value that could conceivably be obtained.

The so-called «modern» approach is to develop so-called «standardized tests» (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; Plake & Wise, 2014) as the assessment boundary object and to adopt «educational standards» (Troia & Olinghouse, 2013) as the definition of the curriculum boundary object. Again, these are explicitly positivist in their intent. However, each has its own translation problems. The standardized test, as it is commonly defined, following the classical test theory (CTT) paradigm, uses a common set of items (perhaps extended beyond one single test form through the technique of population-based equating) to establish a norm-referenced interpretation of the test scores. This certainly generates an interpretation that has clear meaning (such as «student X is at the *nt*h percentile in population *P*»), and also generates estimates of uncertainty for such statements. However, what it fails to do is to give a clear interpretation of what this means in terms of the curriculum itself

(however it is represented), and it also fails to establish translatability to other populations of stakeholders beyond the classroom or the particular assessment involved.

The typical use of educational standards is as a list-wise definition of the domain of the associated standardized test. These lists of standards have attracted criticism (see, for example, Pellegrino, Wilson, Koenig, & Beatty, 2014; Williamson, Fitzgerald, & Stenner, 2013) of two different sorts: (a) that the standards that are included in the list tend to focus on student's content knowledge rather than the processes that must be learned to master a domain of knowledge and (b) that they are presented as individual atoms of this knowledge rather than being linked in sequences of standards that are intended to describe the passages of students towards the most sophisticated levels, which are referred to as *learning progressions* (or, sometimes, learning trajectories; Black, et al., 2011). Both of these are serious translation issues, as they mean that the resulting assessment is not fully mapping to the curriculum, nor is it interpretable in a criterion-referenced way by the teachers who will be teaching the curriculum.

Alternative post-positivist (unmodern or amodern) possibilities are offered in contemporary psychometric research through a construct mapping approach (Wilson, 2005; Wilson & Sloane, 2000). Probabilistic psychometric models supporting construct mapping operationalize the education ideal by parameterizing each facet in the overall design, allowing for (a) testing of hypotheses concerning the representativeness of the student and item samples, (b) theorizing about the coherence of the questions asked and answers received, and (c) qualitative characterizations of both consistent and inconsistent measures. Assessments are created to be consistently structured, meaning that comparisons are interpretable in common terms across different collections of items and different samples of students. Mathematical models of this kind articulating boundary objects from the psychometric perspective provide new opportunities for translating across stakeholder groups, with the assessment content articulating the boundary object for the curriculum developer, the end results for the teacher, and the learning progression for the student. Given advances in electronic communications, networking, computing, analytic algorithms, etc., creative applications of psychometric models can lead to important new developments in translating advanced concepts and methods across the diverse stakeholder groups interested in advancing the quality of educational outcomes, which may

extend outside of the education domain specifically to sustainability accounting, ecological economics, forensic and legal metrology, etc.

The problem for all actors in the translation network is how to reduce their local uncertainty as to what the boundary object is and what it means for them, without alienating the other groups of stakeholders. Allowing a different translation that may better —or only seemingly better— embody their interests will enable it to become a new obligatory passage point. Translations thus take place via standardized processes that coordinate different perspectives relative to the boundary object at the obligatory points of passage.

Standardization ought to be less an imposition of externally determined, arbitrary constraints than a way of setting up analogies across stakeholders' own perspectives. These analogies must be capable of replicating the extension of everyday thinking's model-based reasoning processes accomplished in the natural sciences, an extension that (a) preserves relational structures, (b) makes isomorphic mappings between systems, and (c) achieves systematicity, understood as mapping systems of higher order relational structures (Nersessian & Chandrasekaran, 2009, p. 186). Construct mapping and psychometric modeling aid in setting up, checking, and implementing analogies useful in standardizing the terms for negotiating obligatory passage points in translation networks. In effect, the end goal and result of construct mapping is an ability to say student A relates to item Y in the same way student B relates to item Z, at, say, a 50-50 odds of success. Alternatively, the goal might be to say that student A is to student B as item Y is to item Z (or as the assemblage of skills needed to succeed on item Y is to the skills needed for success on Z).

Either way, the psychometric translation of the boundary object as a mathematical model, the analytic translation as manipulated data, the curricular translation as a learning progression, and the instructional translation as what to teach this student next must all function as analogues of each other. Passage points between stakeholder groups become obligatory when they embody the accepted standard for coordinating activities. In this context, a new perspective emerges on the value of basing measurement design principles in Rasch's separability theorem. In the same way the natural sciences have extended everyday model-based reasoning into an integrated combination of explanatory theory, experimental hypothesis tests, and instruments calibrated in

standard units (Nersessian, 2006, 2008, 2012; Nersessian & Chandrasekaran, 2009), so, too, might psychometrics better fulfill its potential as a science (as opposed to an array of statistical techniques) (Wilson, 2013a) by more systematically making use of these strategies.

Approaches to measurement and instrument calibration focused on the identification and scaling of invariant, unidimensional constructs (Rasch, 1960; Wilson, 2005; Wright, 1977) specify the simplest possible relational structures capable of supporting analogous interpretations across stakeholder groups. This contrasts with statistical models that incorporate variation in item difficulty depending on the location and consistency of responses (DeMars, 2010, p. 16), since these preempt and defeat the identification, mapping, and preservation of relational structures across stakeholder groups. Similarly, Rasch-based test equating methods (Engelhard & Osberg, 1983; Masters, 1985; von Davier, 2010) connect different tests and assessments in a larger network that effectively comprises isomorphic mappings between systems. Finally, Rasch-based construct mapping (Wilson, 2005) facilitates item design, response scoring, and mathematical modeling, including explanatory modeling (De Boeck & Wilson, 2004, 2014; Stenner & Fisher, 2013; Stenner & Smith, 1982) that achieves the systematicity characteristic of higher order relational structures.

Traditional test and survey research not engaged in construct mapping of this kind scores responses irrespective of any overtly conceptualized boundary object, assuming that the objectivity of the response processing, and the transparency of adding up scores, suffices to compel the appropriate concepts, methods, and organizational processes. Standards are then imposed from without on the basis of external social authority and do not emerge from within the experiences of each stakeholder group as authentic expressions. The value of the construct mapping method and the associated probabilistic models hinges on the capacity to advance different groups' genuine interests more effectively than they could be using the traditional methods of classical test theory and norm-referenced standardized testing.

It is important to stress that groups participating in translation networks do not try to assimilate, dissolve, or eradicate groups with differing priorities, cultures, or languages. Instead, they need the multiplicity, instability, marginality, and multicenteredness that foster creative syntheses of divergent originality and convergent conformity (Berg & Timmermans, 2000, p.

38), although this may not always be clear to the participants in every group. Historical studies suggest, then, that the ongoing success of science capitalizes on a balance between divergent, oppositional thinking and convergent, unified thinking (Edwards, Mayernick, Batcheller, Bowker, & Borgman, 2011; Galison & Stump, 1996; Woolley & Fuchs, 2011). The positivist emphasis on observation, the instrumentalist emphasis on technology, and the anti-positivist emphasis on theory as unifying frames of reference have given way in the unmodern possibility of an «intercalated» perspective allowing data, instruments, and theory to act as independent but interrelated factors given varying weights across and within different social networks (Ackerman, 1985; Galison, 1999; Ihde, 1991). As Golinski (2012, p. 35) observes, «Practices of translation, replication, and metrology have taken the place of the universality that used to be assumed as an attribute of singular science». Researchers in any given field typically reside within distinct communities focused on specific issues involving instrumentation, experiment, or theory. There may never or very rarely be any significant communications between these groups beyond the way each appropriates processes and outcomes from the others in its own terms (Galison, 1999).

Facilitating teacher management of assessments

Following on from the discussion above, it can be seen that the commonplace event of a teacher carrying out a student assessment occurs within a very complex web of interlocking boundary objects and passage points, as well as with an accompanying cast of allies. Mostly, of course, this all occurs without the teacher being aware of any of its complexity, let alone actually invoking any of it. However, while one would not want to add to the complexity of a teacher's task, there are good reasons to want this background to be available to the teacher, and, indeed for the teacher to at least be familiar with some of it. For example, having as strong a connection as possible between the results of summative and formative assessments would facilitate the rational use of those results in planning the next steps in instruction. Of course, this is what the Wright map/construct map pair is specifically designed for, and ways to use these maps have been well-documented (e.g, Black, et al, 2011).

Collecting the results of these assessments into a data-accessible form, using the Wright map to assign student locations based on that information, and helping teachers to interpret the

maps are all highly complex activities, and need to be available (almost) instantaneously to be of real help to teachers. This is why there is another set of allies, the IT developers, illustrated in Figure 4, because without a comprehensive data collection and analysis system linked up to a guided interpretative system, a teacher engaged in the typical classroom will be simply overwhelmed by the demands for data entry, handling, and analysis, as well as the operation of the interpretative system. Thus, it is imperative that these materials and ideas be implemented in an IT system that is designed to work consistently within the framework illustrated in Figure 4.

Our response to this is the UC Berkeley Evaluation and Assessment Research (BEAR) Center's online formative assessment system, the BEAR Assessment System Software (BASS; Scalise et al., 2007; Scalise & Wilson, 2011; Torres Irribarra, Freund, Fisher, & Wilson, 2015; Wilson, Scalise, Galpern, & Lin, 2009), which is explicitly designed to facilitate self-directed assessments and has been in development over the last 12 years. Supported by funding from the U.S. National Science Foundation (NSF) and the U.S. federal government's Institute for Education Science (IES), BASS incorporates the principles of the construct modeling approach described above, and is currently being used in several U.S. states for integrated assessment and instruction in various areas of STEM education. The BEAR Assessment System (BAS) includes four building blocks and associated tools for constructing quality assessments: Construct Maps, Items Design, the Outcome Space, and the Measurement Model (see Table 1). These building blocks map to the National Research Council's Assessment Triangle, developed by the National Research Council's (NRC) Committee on the Foundations of Assessment (National Research Council, Division of Behavioral and Social Sciences and Education, Center for Education, & Mathematics Learning Study Committee, 2001).

Table 2
BEAR Assessment System principles, building blocks and products

Principle	Building block	Product
Assessment should be based on a developmental perspective	Construct	Construct map
A match between teaching practice and what is assessed	Items design	Items
Teachers must be the managers of the system,	Outcome space	Scoring guides and exemplars

with the tools to use it efficiently and effectively		
Evidence of quality in terms of reliability, validity, and evidence of fairness	Measurement model	Wright maps

BASS is a web-delivered system intended to be used by assessment and curriculum development teams working with classroom teachers in designing, developing, and delivering formative assessments based on empirically grounded and theoretically validated learning progressions (Wilson, 2004, 2009, 2013a). Using the software, teachers are able to monitor and report student progress in a way that makes it easy to enhance this progress. Results are delivered in a framework that makes science, technology, engineering, and mathematics (STEM) standards meaningful and achievable. One of the distinguishing features of BASS is its incorporation of advanced educational measurement models for assessment. The online-accessible system allows teachers to accurately diagnose students' comprehension and learning needs by providing real-time assessment, logging, analysis, feedback, and reporting.

BASS enables teachers and researchers to modify and further develop assessments, by utilizing, modifying or enhancing construct definitions, item designs, instrument assembly, assessment delivery, data collection, response scoring, statistical modeling, validity and reliability analysis, and reporting of results. Figure 5 shows an overview of the structure of BASS.

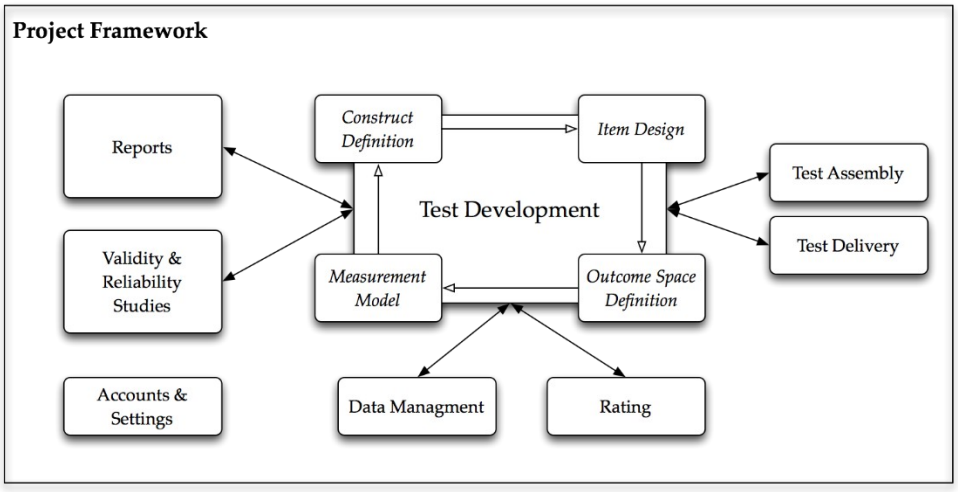

Figure 5. Overview of the BEAR Assessment System Software (BASS) modules.

The first four modules of BASS (Construct Definition, Item Design, Outcome Space Definition, Measurement Model) deploy the functionality of the four foundational BEAR building blocks. The fifth and sixth modules (Test Assembly and Test Delivery) allow creation and delivery of assessment instruments based on the inputs of the test development loop. The seventh module is dedicated to data management, automatic scoring, and recoding, while the eighth (Rating) provides functionality for all scoring that requires human judgment. The ninth module (Validity and Reliability Studies) includes the software features that will be required for validity and reliability analyses. The tenth module (Reports) provides the reporting capabilities needed both in terms of the types of plots and tables available, as well as the different kinds of data that would in principle be useful to access through these reports. Finally, the last module (Accounts and Settings) contains functionality related to administration of the system, including managing settings and preferences and assigning permissions and user access. Although these modules are at different levels of development and polish, most of them are already functional and several have been used in more than one prior or ongoing project.

This system is designed to provide teachers a formative assessment delivery system that accesses the cognitive processes students engage in as they construct responses to problems that require critical thinking. Another purpose of the BASS is to advance teachers' understandings of the uses and techniques of formative assessment, and in particular the integration of formative assessment with the interpretive context laid out through construct maps and their relationship to standards and learning progressions (Wilson, Scalise, Galpern, & Lin, 2009).

Discussion

To be understood and useful in ways enabling teachers, curriculum developers, policy-makers, psychometricians, and assessment developers to advance the field of education as a whole, assessments must do more than merely speak to each stakeholder group in its own language. They must first enable each group to see itself and its interests through the eyes of each of the other groups, and second, enable each group to engage substantively with the others by making the products of its processes available to them in ways that help advance their interests. These translations are the dialogues through which common languages are worked out and put to work in realizing outcomes of collective efforts unattainable by each group alone.

Psychometrics, for instance, must pose certain challenges to itself and its partners in research to fulfill its role in the reflective community. In the project described above, challenges faced by psychometrics could be classified as standard and non-standard. Standard challenges are ones that are usually posed even in a positivist, modern context that prioritizes evidence over theory and instrument. These include defining variables well, creating items in a design-specific way, developing sound coding and scoring systems, applying uni- and multidimensional models to interrogate the data appropriately, and additional efforts that ought to be standard, such as making reports useful for teachers, helping teachers design and adapt assessments by providing them with the tools they need to be effective.

Non-standard challenges to psychometrics include incorporating the metric in multidimensional models, representing and modeling links across dimensions, developing new models that represent well the latent continuum and/or latent classes in learning progressions, and representing and modeling longitudinal change across individual learning progressions.

Challenges to the curriculum developers from the construct modeling approach include the need for greater conceptual precision than is currently typical, along with a need for an increased transparency and explicitness of goals. Instructional fidelity evaluations are charged with much the same tasks, but also can be provided with tools for transcending merely obvious outputs and outcomes. Also, professional development opportunities can be expanded for

teachers in terms of new ways to link teacher practices to large-scale assessment and (even) accountability indices.

BASS is built on the idea of measurement scales that stand for generalized learning structures. The stability and instability of these structures will become better understood in the context of education practice as networks of allies expand by translating and replicating boundary objects relative to common obligatory passage points. Scales representing stable learning structures will eventually be linked together in common systems, not unlike the metrological systems of weights and measures informing commerce and the natural sciences (Fisher, 2000, 2009). Common metrological systems like these will eventually impact physical, institutional, and information resources in ways that will make educators and employers better able to work together to identify and meet human resource and a wide variety of other needs.

Galison (1997, pp. 844-845; also see Fisher, 2011) suggests this direction of inquiry. With no reference to Hayek's (1948, p. 88) statement of the problem (see above), Galison seeks a new analogy capable of informing models of disunified science's translation, replication, and metrology networks. He points to recent technologies that are more reliable and useful in a somewhat disordered form than they are when rigidly ordered (such as amorphous crystals in electronics and laminated materials in structural engineering). In this vein, Galison recalls Wittgenstein's metaphor of concepts as intertwined fibers in a thread that is stronger than it would be if it were formed from only one single continuous fiber. Galison finds this metaphor insufficient to the dynamic processes involved and calls for a non-mechanical metaphor expressing the coordination of the «different symbolic and material actions [through which] people create the binding culture of science». Berg and Timmermans (2000) independently concur, saying that the stability and reach of the medical decision networks they studied were «not due to more (precise) instructions: the protocol's logistics could thrive only by parasitically drawing upon its own disorder» (p. 56). Similar, apparently paradoxical interrelations of order and disorder in speech (Moskowitz & Dickinson, 2002), interpretation theory (Rasch, 1992), and visual perception (Riani & Simonotto, 1994) suggest that a basis for a productive analogy might be found in the physical phenomenon of stochastic resonance (Fisher, 1992, 2011). Characterized by noise-induced order (Dykman & McClintock, 1998; Matsumoto & Tsuda,

1983; Schimansky-Geier, Freund, Neiman, & Shulgin, 1998), stochastic resonance presents intriguing parallels to the model of disunified science sought by Galison, especially when interpreted in terms of nonlinear control theory (Repperger & Farris, 2010).

Heene (2013) suggests potential limitations posed by an analogy from stochastic resonance. Heene's challenge attributes the poor quality of measurement in psychology to the failure of researchers to subject hypotheses of their constructs' quantitative status to experimental falsification. The general disunity of science and the lack of adequate translations through obligatory passage points do not figure as relevant factors for Heene, and they were not mentioned in the brief commentary on stochastic resonance (Fisher, 2011) that he cited. But contrary to Heene's (2013, p. 2) assertions, the analogy to stochastic resonance in no way requires presupposing an extrapolation of micro-level phenomena to macro-level phenomena, nor is the analogy advanced primarily as a justification for the probabilistic nature of item response models. Instead, as is only vaguely suggested by Fisher's (2011) brief commentary, following Maxwell's method of physical analogy (Nersessian, 2002), the point is to use a physical phenomenon as a suggestive point of departure for theory development and experimental evaluation. Maxwell (1965/1890, pp. 155-166, 159-160; Black, 1962, pp. 226-227; Boumans, 2005, pp. 24-25) used analogies to avoid both premature adoption of an explanatory theory and distraction by the analytical subtleties into which researchers can easily be led by mathematical methods. In the same way that Maxwell's model of a frictionless fluid worked to advance electromagnetic theory and its practical application, perhaps stochastic resonance could aid the advancement of measurement theory and practice.

Heene (2013, p. 3) also states that «no experimental evidence currently exists which shows why and how such system-inherent error might occur in the item response process». This assertion ignores the longstanding recognition of the attenuation paradox (Loevinger, 1954; Masters, 1988; Sitgreaves, 1961), in which the removal of stochastic variation results in a deterministic Guttman structure that provides no information useful in estimating the distances between person or item locations. The possibility that Rasch's probabilistic models tap into a general structural phenomenon creating the appearance of deterministic patterns echoes proofs of irreducible randomness even in arithmetic, elementary number theory and Newtonian physics

(Chaitin, 1994). Duncan (1984, p. 220) accordingly observes: «It is curious that the stochastic model of Rasch, which might be said to involve weaker assumptions than Guttman uses [in his deterministic models], actually leads to a stronger measurement model». Where Guttman requires a Procrustean conformity with expectations, such that all observations below an ability measure, for instance, indicate successes, and all of those above the measure, failures, Rasch's «weaker» assumptions allow for play in the observations. Mildly unexpected successes or failures (in the range of 60/40, or 40/60, odds) do not contradict the overall pattern. And even quite markedly unexpected anomalies can be instructionally useful qualitative guides to special needs or strengths.

A larger point here is that no models are true, and data never fit them perfectly (Box, 1979, p. 202; Rasch, 1980/1960, pp. 37-38, 2011/1973). No real triangle ever satisfies the Pythagorean theorem, just as there are no mathematical pendula involving heavy points suspended from weightless strings in a vacuum. As Butterfield (1957, p. 17) put it, «...we do not in real life have perfectly spherical balls moving on perfectly smooth horizontal planes—the trick lay in the fact that it occurred to Galileo to imagine these». In this context, it is interesting that Heene takes the refusal to engage with experimental falsifiability in measurement research as the primary explanation as to why quantification is so often of such poor quality in psychology. It may be that Heene, like Michell (2004), assumes that quantity is something that exists in and of itself in the real world, ontologically prior to its discovery. This would seem to deny the historical fact that measured quantities come into language and social usage via developmental processes. If that denial is taken as the norm, then models and laws are not understood as unrealistic ideals only approximated by measures, and falsification demands demonstration of a quantitative status that can be inferred as temporally preceding the experimental framework in which it is manifest. Though strict empiricism may be satisfied with this, in the absence of an historical, developmental perspective, one is left with nothing but 20/20 hindsight: success in the struggle for an explanatory theory and an additive unit can only mean that the construct had been quantitative all along.

From our point of view, this modern prioritization of data over the historical emergence and development of theory and instrumentation is fatally flawed. As elaborated by Duhem,

Quine, Feyerabend, and others, falsifiability alone is insufficient to the task of justifying theoretical explanations. A variety of economic, moral, political, and other interests compete in augmenting or contradicting the weight of evidence relative to sustaining or changing people's beliefs about things. If Copernicus, Galileo, and Einstein had been strict empiricists convinced solely by the falsification of their hypotheses, it is unlikely we would know their names today because they would not have persisted in believing their theories in the face of contrary evidence. Because of the ways in which scientific innovators are able to ally with others in networks advancing the interests of many diverse stakeholders (see Latour, 1993a, on Pasteur, for instance), it seems likely that social factors involving trading zones, translation networks, stakeholder alliances, and boundary objects will play significant roles in advancing the cause of improved measurement in psychology. On the other hand, if it turns out that the modernist focus on data and falsification turns out in some as yet unknown way to be true, concern with these social factors will be rendered pointless.

Finally, another related potential shortcoming that may undermine the use of construct maps as boundary objects emerges when Heene (2013, p. 4) holds the following: «It is not necessarily wrong to develop mathematical models independently from empirical observations. But, it is also not at all self-evident that empirical insights will result from such models». In opposition to Heene's unstated assumption, it is also not the case that empirical insights follow most assuredly from mathematical models developed from empirical observations. The general independence of mathematical models from empirical observations in the history of science has been a recognized philosophical problem for decades. «Russell speaks of cases 'where the premises of science turn out to be a set of presuppositions neither empirical nor logically necessary,' and in a remarkable passage, Karl R. Popper confesses very plainly to the impossibility of making a science out of only strictly verifiable and justifiable elements» (Holton, 1988, p. 41). Contrary to Heene's presuppositions as to the primacy of data, Butterfield (1957) points out that:

> The law of inertia is not the kind of thing you would discover by mere photographic methods of observation —it required a different kind of thinking-cap, a transposition in the mind of the scientist himself; for we do not actually see ordinary objects continuing their rectilinear motion in that kind of empty space (pp. 16-17).

Instead, what has repeatedly happened in the history of science is that analogies involving projections of geometric concepts and functions have had remarkable success in finding empirical traction, cognitive accessibility, and socioeconomic purchase. This only follows through, then, to a significant degree, on Kant's (1965, pp. 20-21) point: to not merely follow «nature's leading-strings» but to recognize that «reason has insight only into that which it produces after a plan of its own». Though other problems in need of close attention emerge here (Fisher, 2003), Rasch's appropriation of Maxwell's method of analogy (Fisher, 2010) appears to open up a similar realm of geometric possibilities in psychometrics (Fisher & Stenner, 2013a).

The possible relevance of stochastic resonance as a useful analogy is also supported by studies of collective intelligence, which have found that successful and productive fields of research and practice incorporate divergent opening and bridging activities in their organizing activities, along with convergent defining and bounding activities, grounding the whole continuum in reflective practice (Woolley & Fuchs, 2011). Thus, beyond reflective practitioners (Schon, 1983), reflective communities are needed to coordinate and align diverse interests in ways that capitalize on systematic disunities (Berg & Timmermans, 2000; Fischer, Giaccardi, Eden, Sugimoto, & Ye, 2005; Haraway, 1996), where the members of different stakeholder groups have only partial understandings of the data, theory, and instruments that help them advance their interests.

Construct maps, Wright maps, and the construct modeling method in general can provide images capable of intermediating the diverse needs of translation networks. To be able to perform this function effectively, they will need to embody system-level analogies between different stakeholder groups' perspectives. Efforts to date in persuading and educating researchers and the public as to the value of rigorous and meaningful approaches to measurement have largely failed to provoke anything akin to the revolution that might have been expected (Cliff, 1992), leading some to hold that such a revolution cannot happen (Trendler, 2009). One of the widely held assumptions supporting that contention concerns a supposed incapacity for causal manipulations, which is contradicted by longstanding research results (DeBoeck & Wilson, 2004, 2014; Embretson, 2010; Fischer, 1973, 1983; Stenner, Fisher, Stone, & Burdick, 2013; Stenner & Smith, 1982). Even so, it is likely nonetheless still true that the successful

application of explanatory and predictive models is also insufficient in itself to the larger task of broad-scale improvements to the quality of measurement in psychology. Success in generalizing the realization of high-quality measurement and the communication of meaningful quantities will require efforts that go beyond proofs and evidence to thickly-elaborated social and economic alliances across stakeholder groups, whose activities must be coordinated and aligned relative to obligatory passage points in a systematically constructed translation network.

In the manner of metrologically traceable instrumentation, quantitative measuring units and their associated qualitative concepts and vocabularies will require not just empirical evidence and theoretical explanations but also systematic alliances capable of advancing every implicated stakeholder group's interests more effectively than those groups could advance their interests alone. Work in this direction is proceeding (e.g., Fisher & Stenner, 2013b; Mari & Wilson, 2013; Pendrill & Fisher, 2015; Wilson, 2013b; Wilson, Mari, Maul, & Torres Irribarra, 2015). But instead of being the fulfillment of positivist dreams of abstract universals —or of the anti-positivist nightmare of incommensurate local dependencies— the standards incorporated in translation networks for psychology and education will likely be more realistically described as a highly demanding prescription for hard work, with more promise for productive results than has previously been conceivable.

References

Ackermann, J. R. (1985). *Data, instruments, and theory: A dialectical approach to understanding science*. Princeton, New Jersey: Princeton University Press.

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

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543-554. doi: 10.1177/1745691612459060

Berg, M., & Timmermans, S. (2000). Order and their others: On the constitution of universalities in medical work. *Configurations, 8*(1), 31-61.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *American Statistical Association Journal, 33*(201-204), 526-536.

Best, W. R. (2008). A Rasch model of the Crohn's Disease Activity Index (CDAI): Equivalent levels of ranked attribute and continuous variable scales. In J. N. Cadwallader (Ed.), *Crohn's disease: Etiology, pathogenesis and interventions* (Chapter 5). New York: Nova Science Publishers, Inc.

Black, M. (1962). *Models and metaphors*. Ithaca, New York: Cornell University Press.

Black, P., Wilson, M., & Yao, S. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research & Perspectives, 9,* 1-52. doi: 10.1080/15366367.2011.591654

Bolles, R. D. (1962). The differences between statistical hypotheses and scientific hypotheses. *Psychological Reports, 11*, 639-645.

Bond, T., & Fox, C. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences.* (3rd ed.). New York: Routledge.

Boumans, M. (2005). *How economists model the world into numbers*. New York: Routledge.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer, & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201-235). New York: Academic Press, Inc.

Butterfield, H. (1957). *The origins of modern science* (Rev. ed.). New York: The Free Press.

Callon, M. (1985). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St. Brieuc Bay. In J. Law (Ed.), *Power, action and belief: Sociological Review Monograph No. 32* (pp. 196-230). London: Routledge & Kegan Paul.

Chaitin, G. J. (1994). Randomness and complexity in pure mathematics. *International Journal of Bifurcation and Chaos, 4*(1), 3-15.

Claesgens, J., Scalise, K., Wilson, M., & Stacy, A. (2009). Mapping student understanding in chemistry: The perspectives of chemists. *Science Education, 93*(1), 56-85. doi: 10.1002/sce.20292

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science, 3,* 186-190.

Coats, W. (1970). A case against the normal use of inferential statistical models in educational research. *Educational Researcher, 3*, 6-7.

Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist, 49*, 997-1003.

Daston, L. (Ed.). (2004). *Things that talk: Object lessons from art and science.* New York: Zone Books.

Daston, L., & Galison, P. (1992). The image of objectivity. *Representations, 40,* 81-128.

Daston, L., & Galison, P. (2007). *Objectivity.* Cambridge, MA: MIT Press.

Dawson, T. L., Xie, Y., & Wilson, M. (2003). Domain-general and domain-specific developmental assessments: Do they measure the same thing? *Cognitive Development*, *18*(1), 61-78.

De Boeck, P., & Wilson, M. (2014). Multidimensional explanatory item response models. In S. P. Reise, & D. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 252-271). New York: Routledge.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer-Verlag.

Dear, P., Hacking, I., Jones, M. L., Daston, L., & Galison, P. (2012). Objectivity in historical perspective. *Metascience, 21*(1), 11-39. doi: 10.1007/s11016-011-9597-2

DeMars, C. (2010). *Item response theory (N. Beretvas, Series Ed.). Series in Understanding Statistics*. New York: Oxford University Press.

Dewey, J. (2012). *Unmodern philosophy and modern philosophy* (P. Deen, Ed.). Carbondale, Illinois: Southern Illinois University Press.

Duncan, O. D. (1984). *Notes on social measurement: Historical and critical*. New York: Russell Sage Foundation.

Dykman, M. I., & McClintock, P. V. E. (1998). What can stochastic resonance do? *Nature, 391*(6665), 344.

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science, 41*(5), 667-690. doi: 10.1177/0306312711413314

Embretson, S. E. (2010). *Measuring psychological constructs: Advances in model-based approaches.* Washington, DC: American Psychological Association.

Engelhard, G., Jr., & Osberg, D. (1983). Constructing a test network with the Rasch measurement model. *Applied Psychological Measurement, 7*(3), 283-294.

Ewert, T., Allen, D. D., Wilson, M., Üstün, B., & Stucki, G. (2010). Validation of the International Classification of Functioning Disability and Health framework using multidimensional item response modeling. *Disability and Rehabilitation*, 32(*17*), 1397-1405. doi: 10.3109/09638281003611037

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

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48*(1), 3-26.

Fischer, G.H., Giaccardi, E., Eden, H., Sugimoto, M., & Ye, Y. (2005). Beyond binary choices: Integrating individual and social creativity. *International Journal of Human-Computer Studies, 63,* 482-512.

Fisher, W. P. Jr. (1992). Stochastic resonance and Rasch measurement. *Rasch Measurement Transactions, 5*(4), 186-187.

Fisher, W. P. Jr. (2000). Objectivity in psychosocial measurement: What, why, how. *Journal of Outcome Measurement, 4*(2), 527-563.

Fisher, W. P. Jr. (2003). Mathematics, measurement, metaphor, metaphysics: Parts I & II. *Theory & Psychology, 13*(6), 753-828.

Fisher, W. P. Jr. (2005). Daredevil barnstorming to the tipping point: New aspirations for the human sciences. *Journal of Applied Measurement, 6*(3), 173-179.

Fisher, W. P. Jr. (2009). Invariance and traceability for measures of human, social, and natural capital: Theory and application. *Measurement, 42*(9), 1278-1287. doi:10.1016/j.measurement.2009.03.014

Fisher, W. P. Jr. (2010). The standard model in the history of the natural sciences, econometrics, and the social sciences. *Journal of Physics: Conference Series, 238*(012016), 1-5. doi: 10.1088/1742-6596/238/1/012016

Fisher, W. P. Jr. (2011). Stochastic and historical resonances of the unit in physics and psychometrics. *Measurement: Interdisciplinary Research & Perspectives, 9*(1), 46-50. doi: 10.1080/15366367.2011.558789

Fisher, W. P., Jr., & Stenner, A. J. (2011). Integrating qualitative and quantitative research approaches via the phenomenological method. International Journal of Multiple Research Approaches, 5(1), 89-103.

Fisher, W. P. Jr., & Stenner, A. J. (2013a). On the potential for improved measurement in the human and social sciences. In Q. Zhang, & H. Yang (Eds.), *Pacific Rim Objective Measurement Symposium 2012 Conference Proceedings* (pp. 1-11). Berlin, Germany: Springer-Verlag.

Fisher, W. P. Jr., & Stenner, A. J. (2013b). Overcoming the invisibility of metrology: A reading measurement network for education and the social sciences. *Journal of Physics: Conference Series, 459*(012024), 1-6. doi: 10.1088/1742-6596/459/1/012024

Galison, P. (1997). *Image and logic: A material culture of microphysics.* Chicago: University of Chicago Press.

Galison, P. (1999). Trading zone: Coordinating action and belief. In M. Biagioli (Ed.), *The science studies reader* (pp. 137-160). New York: Routledge.

Galison, P. (2008). Image of self. In L. Daston (Ed.), *Things that talk: Object lessons from art and science* (pp. 256-294). New York: Zone Books.

Galison, P., & Stump, D. J. (1996). *The disunity of science: Boundaries, contexts, and power.* Palo Alto, California: Stanford University Press.

Golinski, J. (2012). Is it time to forget science? Reflections on singular science and its history. *Osiris, 27*(1), 19-36.

Guttman, L. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis, 1*, 3-10.

Haraway, D. J. (1996). Modest witness: Feminist diffractions in science studies. In P. Galison & D. J. Stump (Eds.), *The disunity of science: Boundaries, contexts, and power* (pp. 428-441). Stanford, California: Stanford University Press.

Hayek, F. A. (1948). *Individualism and economic order.* Chicago: University of Chicago Press.

Heene, M. (2013). Additive conjoint measurement and the resistance toward falsifiability in psychology. *Frontiers in Psychology, 4*(246). doi: 10.3389/fpsyg.2013.00246

Holton, G. (1988). *Thematic origins of scientific thought: Kepler to Einstein* [Revised ed.]. Cambridge, Massachusetts: Harvard University Press.

Hutchins, E. (1995). *Cognition in the wild.* Cambridge, Massachusetts: MIT Press.

Hutchins, E. (2014). The cultural ecosystem of human cognition. *Philosophical Psychology, 27*(1), 34-49. doi: 10.1080/09515089.2013.830548

Ihde, D. (1991). *Instrumental realism: The interface between philosophy of science and philosophy of technology*. Bloomington, Indiana: Indiana University Press.

Ihde, D. (1998). *Expanding hermeneutics: Visualism in science*. Evanston, Illinois: Northwestern University Press.

Ihde, D. (2012). *Experimental phenomenology: Multistabilities.* (2nd ed.). Albany, New York: SUNY Press.

Kaiser, F. G., & Wilson, M. (2000). Assessing people's general ecological behavior: A cross-cultural measure. *Journal of Applied Social Psychology, 30*, 952-978.

Kant, I. (1965). *Critique of pure reason*. New York: St. Martin's Press.

Kuhn, T. S. (1970). *The structure of scientific revolutions*. Chicago, Illinois: University of Chicago Press.

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

Latour, B. (1990). Postmodern? No, simply amodern: Steps towards an anthropology of science. *Studies in History and Philosophy of Science, 21*(1), 145-71.

Latour, B. (1991). The impact of science studies on political philosophy. *Science, Technology, & Human Values, 16*(1), 3-19.

Latour, B. (1993a). *The Pasteurization of France.* Cambridge, Mass.: Harvard University Press.

Latour, B. (1993b). *We have never been modern.* Cambridge, Massachusetts: Harvard University Press.

Latour, B. (2010). A compositionist manifesto. *New Literary History, 41,* 471-490.

Law, J. (Ed.). (1985). *Sociological review monograph. Vol. 32: Power, action and belief.* London: Routledge & Kegan Paul.

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51*, 493-504.

Mari, L., & Wilson, M. (2013). A gentle introduction to Rasch measurement models for metrologists. *Journal of Physics Conference Series, 459*(1). doi: 10.1088/1742-6596/459/1/012002

Masters, G. N. (1985). Common-person equating with the Rasch model. *Applied Psychological Measurement, 9*(1), 73-82.

Masters, G. N. (1988). Item discrimination: when more is worse. *Journal of Educational Measurement, 25*(1), 15-29.

Matsumoto, K., & Tsuda, I. (1983). Noise-induced order. *Journal of Statistical Physics, 31*(1), 87-106.

Maxwell, J. C. (1965/1890). *The scientific papers of James Clerk Maxwell* (W. D. Niven, Ed.). New York: Dover Publications.

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

Miller, P., & O'Leary, T. (2007). Mediating instruments and making markets: Capital budgeting, science and the economy. Accounting, Organizations, and Society, 32(7-8), 701-734.

Moskowitz, M. T., & Dickinson, B. W. (2002). Stochastic resonance in speech recognition: Differentiating between /b/ and /v/. *Proceedings of the IEEE International Symposium on Circuits and Systems, 3*, 855-858.

National Research Council, Division of Behavioral and Social Sciences and Education, Center for Education, & Mathematics Learning Study Committee (2001). *Adding it up: Helping children learn mathematics* (J. Kilpatrick, J. Swafford, & B. Findell, Eds.). Washington, DC: National Academy Press.

Nersessian, N. J. (1996). Child's play. *Philosophy of Science, 63,* 542-546.

Nersessian, N. J. (2002). Maxwell and «the method of physical analogy»: Model-based reasoning, generic abstraction, and conceptual change. In D. Malament (Ed.), *Reading natural philosophy: Essays in the history and philosophy of science and mathematics* (pp. 129-166). Lasalle, Illinois: Open Court.

Nersessian, N. J. (2006). Model-based reasoning in distributed cognitive systems. *Philosophy of Science, 73,* 699-709.

Nersessian, N. J. (2008). *Creating scientific concepts.* Cambridge, Massachusetts: MIT Press.

Nersessian, N. J. (2012). Engineering concepts: The interplay between concept formation and modeling practices in bioengineering sciences. *Mind, Culture, and Activity, 19,* 222-239.

Nersessian, N. J., & Chandrasekaran, S. (2009). Hybrid analogies in conceptual innovation in science. *Cognitive Systems Research, 10,* 178-188.

Pellegrino, J. W., Wilson, M., Koenig, J. A., & Beatty, A. S. (Eds). (2014). *Developing assessments for the next generation science standards.* Report of the Committee on Developing Assessments of Science Proficiency in K-12. Washington DC: National Academies Press.

Pendrill, L., & Fisher, W. P. Jr. (2015). Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. *Measurement, 71*, 46-55.

Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the Revised AERA, APA, NCME Standards for Educational and Psychological Testing? *Educational Measurement: Issues and Practice, 33*(4), 4-12.

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

Rasch, G. (2011/1973). All statistical models are wrong! Comments on a paper presented by Per Martin-Löf, at the Conference on Foundational Questions in Statistical Inference, Aarhus, Denmark, May 7-12, 1973. *Rasch Measurement Transactions, 24*(4), 1309.

Rasch, W. (1992). Injecting noise into the system: Hermeneutics and the necessity of misunderstanding. *SubStance, 21*(1), 61-76.

Repperger, D. W., & Farris, K. A. (2010). Stochastic resonance —a nonlinear control theory interpretation. *International Journal of Systems Science, 41*(7), 897-907.

Riani, M., & Simonotto, E. (1994). Stochastic resonance in the perceptual interpretation of ambiguous figures: A neural network model. *Physical Review Letters, 72*(19), 3120-3123.

Ricoeur, P. (1992). *Oneself as another*. Chicago, Illinois: University of Chicago Press.

Roberts, F. S. (1994). Limitations on conclusions using scales of measurement. In A. Barnett, S. Pollock, & M. Rothkopf (Eds.), *Operations research and the public sector* (pp. 621-671). Amsterdam, The Netherlands: Elsevier.

Scalise, K., Bernbaum, D. J., Timms, M., Harrell, S. V., Burmester, K., Kennedy, C. A., & Wilson.M. (2007). Adaptive technology for e-learning: Principles and case studies of an emerging field. *Journal of the American Society for Information Science and Technology, 58*(14), 2295-2309.

Scalise, K., & Wilson, M. (2011). The nature of assessment systems to support effective use of evidence through technology. *E-Learning and Digital Media, 8,* 121-132.

Schimansky-Geier, L., Freund, J. A., Neiman, A. B., & Shulgin, B. (1998). Noise induced order: Stochastic resonance. *International Journal of Bifurcation and Chaos, 8*(5), 869-879.

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action.* New York: Basic Books.

Sitgreaves, R. (1961). A statistical formulation of the attenuation paradox in test theory. In H. Solomon (Ed.), *Studies in item analysis and prediction* (p. 17-28). Stanford, CA: Stanford University Press.

Smith, E. V. Jr. (2005). Representing treatment effects with variable maps. In N. Bezruczko (Ed.), *Rasch measurement in health sciences* (pp. 247-259). Maple Grove, MN: JAM Press.

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, 'translations,' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science, 19,* 387-420.

Stenner, A. J., Fisher, W. P. Jr., Stone, M. H., & Burdick, D. S. (2013). Causal Rasch models. *Frontiers in Psychology: Quantitative Psychology and Measurement, 4*(536), 1-14. doi: 10.3389/fpsyg.2013.00536

Stenner, A. J., & Fisher, W. P. Jr. (2013). Metrological traceability in the social sciences: A model from reading measurement. *Journal of Physics: Conference Series, 459*(012025). Retrieved from http://iopscience.iop.org/1742-6596/459/1/012025.

Stenner, A. J., & Smith, M. (1982). Testing construct theories. *Perceptual and Motor Skills, 55,* 415-426.

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

Torres Irribara, D., Freund, R., Fisher, W. P. Jr., & Wilson, M. (2015). Metrological traceability in education: A practical online system for measuring and managing middle school mathematics instruction. *Journal of Physics Conference Series, 588*(012042). doi:10.1088/1742-6596/588/1/012042

Toulmin, S. E. (1961). *Foresight and understanding: An enquiry into the aims of science*. London, England: Hutchinson.

Toulmin, S. E. (1982). The construal of reality: Criticism in modern and postmodern science. *Critical Inquiry, 9*, 93-111.

Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory & Psychology, 19,* 579-599.

Troia, G. A., & Olinghouse, N. G. (2013). The Common Core State Standards and evidence-based educational practices: The case of writing. School Psychology Review, 42(3), 343-357.

von Davier, A. (Ed.). (2010). *Statistical models for test equating, scaling, and linking*. (Statistics for Social and Behavioral Sciences). New York: Springer.

Whitehead, A. N. (1911). *An introduction to mathematics*. New York: Henry Holt and Co.

Williamson, G. L., Fitzgerald, J., & Stenner, A. J. (2013). The common core state standards' quantitative text complexity trajectory: Figuring out how much complexity is enough. *Educational Researcher, 42*(2), 59-69.

Wilson, M. (Ed.). (2004). *National society for the study of education yearbooks. Vol. 103, Part II: Towards coherence between classroom assessment and accountability*. Chicago, Illinois: University of Chicago Press.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Wilson, M. R. (2009). Measuring progressions: Assessment structures underlying a learning progression. Journal of Research in Science Teaching, 46, 716-730.

Wilson, M. R. (2013a). Seeking a balance between the statistical and scientific elements in psychometrics. *Psychometrika, 78*(2), 211-236. doi: 10.1007/s11336-013-9327-3

Wilson, M. R. (2013b). Using the concept of a measurement system to characterize measurement models used in psychometrics. *Measurement, 46,* 3766-3774. doi:10.1016/j.measurement.2013.04.005

Wilson, M., Allen, D. D., & Li, J. C. (2006). Improving measurement in behavioral sciences using item response modeling: introducing item response modeling. *Health Education Research, 21*(Supplement 1), 4-18.

Wilson, M., Mari, L., Maul, A., & Torres Irribarra, D. (2015). A comparison of measurement concepts across physical science and social science domains: Instrument design, calibration, and measurement. *Journal of Physics Conference Series, 588*(012034). doi:10.1088/1742-6596/588/1/01203

Wilson, M., Scalise, K., Galpern, A., & Lin, Y.-H. (2009). *A guide to the Formative Assessment Delivery System (FADS).* Berkeley: University of California Berkeley, Berkeley Evaluation & Assessment Research Center.

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

Wilson, T. P. (1971). Critique of ordinal variables. *Social Forces, 49*, 432-444.

Woolley, A. W., & Fuchs, E. (2011). Collective intelligence in the organization of science. *Organization Science, 22*(5), 1359-1367.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*(2), 97-116.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.