PSYCHOMETRIKA—VOL. 78, NO. 2, 211–236
APRIL 2013
DOI: 10.1007/S11336-013-9327-3

# SEEKING A BALANCE BETWEEN THE STATISTICAL AND SCIENTIFIC ELEMENTS IN PSYCHOMETRICS

MARK WILSON

UNIVERSITY OF CALIFORNIA, BERKELEY

In this paper, I will review some aspects of psychometric projects that I have been involved in, emphasizing the nature of the work of the psychometricians involved, especially the balance between the statistical and scientific elements of that work. The intent is to seek to understand where psychometrics, as a discipline, has been and where it might be headed, in part at least, by considering one particular journey (my own). In contemplating this, I also look to psychometrics journals to see how psychometricians represent themselves to themselves, and in a complementary way, look to substantive journals to see how psychometrics is represented there (or perhaps, not represented, as the case may be). I present a series of questions in order to consider the issue of what are the appropriate foci of the psychometric discipline. As an example, I present one recent project at the end, where the roles of the psychometricians and the substantive researchers have had to become intertwined in order to make satisfactory progress. In the conclusion I discuss the consequences of such a view for the future of psychometrics.

Key words: psychometrics, test theory, test construction.

## 1. The Roles of Psychometricians

Like many young people first starting out along the journey that eventually led to psychometrics, I had my first real experiences of psychometrics during my doctoral studies. There, at the University of Chicago, I was ensconced in a moderately large cohort of graduate students, and had some of my formative experiences at the "coal-face" of psychometrics, working alongside my fellow doctoral students. One of the sayings (in fact it was a joke) of that group was "Data comes in boxes," that is, we do not go out and get the data, it comes to us ready for data cleaning and analysis. Of course, today, this expression is somewhat anachronistic—today one might say "Data comes on disks," or perhaps "Data comes in over the cloud." But the sentiment is the same, we (the psychometricians) do not have responsibility for "the data" before it is indeed "data." To put it another way, psychometricians are concerned with "Test Theory" (i.e., with analyzing the test data), but not so much with "Test Construction" (i.e., with what happens before the data is produced). In a similar vein, Borsboom (2006) has seen this as a part of the "splendid isolation" of psychometrics.

The perceived role of the psychometrician, then, is to consult with the data collectors about the nature and structure of the data set, and the scientific research questions that motivated its construction. Then the psychometrician selects, or in some cases, develops, statistical models for analyzing the data set and reporting back in ways that inform the research questions. This process may iterate a number of times and the research questions may evolve as time goes by. Results tend to get published in (at least) two areas: (a) results concerning the modeling/statistical issues are published in psychometrics and other methodology journals, and (b) results for the research questions are published in substantive journals related to the topic area of the research.

This "standard story" about what psychometricians do is thus associated with particular roles concerning psychometrics for these two sorts of journals. Psychometrics journals build up

Requests for reprints should be sent to Mark Wilson, University of California, Berkeley, Berkeley, CA, USA. E-mail: markw@berkeley.edu

an armamentarium of psychometric methods and/or techniques for developing new methods that will make the analyses mentioned above both systematic and efficacious. Substantive journals, as part of their reporting on scientific results, disseminate new (and hopefully better) methods to their specific communities of substantive researchers. Note that this is only a minor role for the substantive journals themselves, their major role, of course, is to disseminate the results of those research questions mentioned above. Their role concerning psychometrics is, in their own terms, a relatively minor one.

In the next segment, I illustrate this "standard story" using an experience from my own career. I chose an example of my own work, not to advertise it, but rather so that I can be as free as possible in my observations, which would not be the case if I were to use someone else's research.

## 2. Interlude I: The Saltus Model[1]

The *saltus* model ("saltus" is Latin for leap) (Wilson, 1989) is a special case of a confirmatory mixture item response model with linear restrictions on the relations among sets of item difficulties for the different (latent) subpopulations. It was originally designed for the investigation of developmental stages. This model was developed as a method for detecting and analyzing discontinuities in performance that are hypothesized to occur as a result of rapidly occurring person growth (e.g., Fischer, Pipp, & Bullock, 1984). Such discontinuities are often theorized to occur as the result of progression through developmental stages or levels. Thus, the model is built upon the assumption that the subpopulations are ordered in some way (as developmental stages in children are assumed to be ordered), and that groups of items become predictably easier (or, perhaps less often, more difficult) for subpopulations further along the developmental continuum.

The saltus model was developed partly in response to developmental theories of cognition such as those posited by Piaget. Theories with similar structure, but perhaps different substantive focus, are described by the many neo-Piagetian researchers, and by other educational and psychological researchers who use stage-based theories. For example, Siegler (1981) developed sets of items regarding which side of a balance scale would go down, when one placed different combinations of weights and distances from the fulcrum on the two sides of the balance scale. These sets of items changed predictably in difficulty for different age groups of children as the children progressed through Piagetian-based stages. Some groups of items became easier and some more difficult, while others remained the same. The developmental stages of the children thus resulted in relative shifts in the probability that certain groups of items would be answered correctly. Thus, this can be seen as a case where substantive scientific theory of a reasonably comprehensive nature is available (as distinct from the more general situation as perceived by Borsboom (2006), where "there is a shortage of substantive theory that is sufficiently detailed to drive informed psychometric modeling"). Of course, as Borsboom observes, there remain many psychometric modeling issues that are not driven by the theory. The saltus model is designed to be relevant for use with such sets of items (see Wilson, 1989; and Draney, 1996). A more general mixture model, such as Rost's (1990) mixture Rasch model, would require the estimation of a difficulty parameter for each item within each developmental stage (if the items are dichotomous); the saltus model can accommodate many developmental theories by estimating one difficulty for each item, plus a small number of additional parameters to describe the changes associated with developmental stage.

[1]Note that the material in this example is based principally on the paper by Draney and Wilson (2004).

The saltus model is based on the assumption that there are $H$ developmental stages. A different set of items represents each one of these stages, such that only persons at or above a stage are fully equipped to answer the items associated with that stage correctly. The saltus model assumes that all persons in stage $h$ answer all items in a manner consistent with membership in that stage. However, persons within a stage may differ in proficiency. In a Piagetian context, this means that a child in, say, the concrete operational stage is always in that stage, and answers all items accordingly. The child does not possess formal operational development for some items and concrete operational development for others. However, some concrete operational children may be more proficient at answering items than are other concrete operational children. To describe the model, suppose that, as in the partial credit model (Masters, 1982), the random variable $X_{ni}$ indicates the $n$th person's response to item $i$. Items have $J_i + 1$ possible response alternatives indexed $j = 0, 1, \ldots, J_i$. The difference in difficulty between any two consecutive item levels is referred to as a step, as in Masters' representation of the model. The parameter indicating step $j$ for item $i$ is indicated by $\beta_{ij}$; the vector of all $\beta_{ij}$ for item $i$ by $\boldsymbol{\beta}_i$. In the saltus model, a person is characterized by a proficiency parameter $\theta_n$ and an indicator vector for stage membership $\boldsymbol{\phi}_n$. If there are $H$ potential stages, $\boldsymbol{\phi}_n = (\phi_{n1}, \ldots, \phi_{nH})$, where $\phi_{nh}$ takes the value of 1 if person $n$ is in stage $h$ and 0 if not. Only one of the $\phi_{nh}$ is theoretically nonzero. As with $\theta_n$, values of $\boldsymbol{\phi}_n$ are not observable.

Just as persons are associated with one and only one stage, items are associated with one and only one stage. Unlike person stage membership, however, which is unknown and must be estimated, item stage is assumed to be known *a priori*, based on the theory that was used to produce the items. We denote item stage membership by the indicator vector $\boldsymbol{b}_i$. As with $\phi_n$, $\boldsymbol{b}_i = (b_{il}, \ldots, b_{iH})$, where $b_{ik}$ takes the value of 1 if item $i$ belongs to item stage $k$, and 0 otherwise. The set of all $\boldsymbol{b}_i$ across all items is denoted by the matrix $\boldsymbol{b}$. The equation

$$P(X_{ni} = j \mid \boldsymbol{\theta}_n, \boldsymbol{\phi}_{nh}, \boldsymbol{\beta}_i, \boldsymbol{\tau}_{hk}) = \frac{\exp \sum_{s=0}^{j} (\boldsymbol{\theta}_n - \boldsymbol{\beta}_{is} + \boldsymbol{\tau}_{hk})}{\sum_{t=0}^{J_i} \exp \sum_{s=0}^{j} (\boldsymbol{\theta}_n - \boldsymbol{\beta}_{is} + \boldsymbol{\tau}_{hk})} \tag{1}$$

indicates the probability of response $j$ to item $i$ for person $n$, where it is assumed that it is known that person $n$ is in class $h$. The saltus parameter $\tau_{hk}$ describes the additive effect—positive or negative—for people in stage $h$ on the item parameters of all items in stage $k$. In a developmental context, this often takes the form of an increase in probability of success as the person achieves the stage at which an item is located, indicated by $\tau_{hk} > 0$ when $h \geq k$ (although this need not be the case). The saltus parameters can be represented together as an $H$-by-$H$ matrix $\boldsymbol{T}$. The probability that an examinee with stage membership parameter $\boldsymbol{\phi}_n$ and proficiency $\boldsymbol{\theta}_n$ will respond in category $j$ to item $i$ is given by

$$P(X_{ni} = j \mid \boldsymbol{\theta}_n, \boldsymbol{\phi}_n, \boldsymbol{\beta}_i, b_i, \boldsymbol{T}) = \prod_h \prod_k P(X_{ni} = j \mid \boldsymbol{\theta}_n, \boldsymbol{\phi}_{nh} = 1, \boldsymbol{\beta}_i, \boldsymbol{\tau}_{hk})^{\phi_{nh} b_{ik}}. \tag{2}$$

Assuming conditional independence, the modeled probability of a response vector is

$$P(X_{ni} = x_{ni} \mid \boldsymbol{\theta}_n, \boldsymbol{\phi}_n, \boldsymbol{\beta}_i, b_i, \boldsymbol{T}) = \prod_h \prod_k \prod_i P(X_{ni} = x_{ni} \mid \boldsymbol{\theta}_n, \boldsymbol{\phi}_{nh} = 1, \boldsymbol{\beta}_i, \boldsymbol{\tau}_{hk})^{\varphi_{nh} b_{ik}}. \tag{3}$$

The model requires a number of constraints on the parameters. For item step parameters, we use two traditional constraints: first, $\beta_{i0} = 0$ for every item, and second, the sum of all the $\beta_{ij}$ is set equal to zero. Some constraints are also necessary on the saltus parameters. The set of constraints usually chosen is the same as that used by Mislevy and Wilson (1996), and allows one to interpret the saltus parameters as changes relative to the first (lowest) developmental stage. Two sets of constraints are used. First $\tau_{h1} = 0$; thus, the difficulty of the first stage of items is held constant for all person groups; changes in the difficulty of items representing higher stages are interpreted with respect to this first stage of items for all person stages. Also $\tau_{1k} = 0$; thus,

items as seen by person stages higher than 1 will be interpreted relative to the difficulty of the items as seen by persons in the lowest developmental stage.

As in Mislevy and Wilson (1996), the EM algorithm (Dempster, Laird, & Rubin, 1977) can be used to estimate the structural parameters for the model. Empirical Bayes estimation is then used to obtain estimates of the probabilities of stage membership for each subject, as well as proficiency estimates given membership in each stage. A person is classified into the stage for which that person's probability of membership is highest; however, it is possible to investigate the confidence with which we classify persons with various sorts of response patterns into that stage. Software for this purpose has been developed by Draney and Jeon (2011).

### *2.1. Saltus: An Evaluation*

All in all, I see this as a relatively good example of the "standard story" that was mentioned above. The story starts with a particular scientific theory (which was, at least at one point, very prominent), and describes the development of a psychometric model that addresses several of the critical features of the theory, embodying the features of the theory into model parameters that allow one to express the theory quantitatively, and hence, allowing one to describe how data relates to the theory, and to test whether those features are empirically supported.

And it has certainly had an impact. For example, there have been some 20 publications that are focused on the model and/or applications of the model (see Appendix for a list of these). However, the impact on the field of psychology has been very little—of the 21 publications, only five have been in substantive journals (Acton, Kunz, Wilson, & Hall, 2005; De Boeck, Wilson, & Acton, 2005; Demetriou & Efklides, 1989; Demetriou & Kyriakides, 2006; Pirolli & Wilson, 1998).

Maybe this is just a story of poor communication—I have not succeeded in getting notice of this work among psychologists. But I have made efforts, and the fact that three of the publications just cited have me as a co-author speak to that. Now, there are some caveats to this—stage theories are no longer a "hot topic" in psychology, in fact they are rather passé—so maybe one would not expect psychologists to pick up the saltus model and use it. But there is still a "Jean Piaget Society" that has annual meetings that attracts hundreds of participants each year. So, there should be some interest from that group. Altogether, it seems a puzzling, and rather alarming outcome.

Another argument could be that the saltus model was just way too complex for the psychologists to take up. This may indeed be true. However, the saltus model is *much less* complex than many models that are being promulgated today, which feature not only mixture distributions, but also multidimensionality, multilevel perspectives, and complex discrete modeling. Which might give one pause.

A third argument could be that this occurrence of an idea being proposed, and not taken up, is in fact a common one in science. Lots of good ideas get generated in psychometrics, some survive for a while, few for a long time—occasionally an idea that not taken up gets resurrected. Many factors influence these, not just scientific issues, but also factors such as social dynamics, what is fashionable at a certain time, and even coincidence. This phenomenon of the waxing and waning of ideas can be seen as being akin to "natural selection," and hence one should not be overly concerned if it does not proceed in a specific direction.

## 3. The Roles of Psychometricians, Continued

Thus, to continue the story above, we can see the logical path: Psychometricians, having mastered the armamentaria mentioned above (and exemplified via *saltus*), now develop more

and more complex models (and families of complex models) to add to the armamentarium. To publish their products (i.e., these more and more complex models), psychometricians must cast around for suitable data to illustrate the new methods. The newer and yet newer models are published in psychometrics journals, and psychometrics and related edited volumes. But seldom do they get published elsewhere.

One of the consequences of this is that many new psychometric models (even while embodying interesting substantive hypotheses) are beyond the reach of typical substantive researchers. Hence, substantive researchers and users (e.g., research psychologists and people in the production divisions of testing companies) tend towards the following:

(a) they maintain "standard measurement approaches" that are not well-aligned with more recent developments in psychometrics (see Borsboom, 2006, for interesting observation of such), and/or
(b) they develop approaches that derive from alternative perspectives to psychometrics (e.g., non-quantitative perspectives, item-focused approaches).

As an example of the former consequence, consider the typical training of research psychologists in measurement. As Borsboom (2006) observes: "every introductory textbook on psychological research methods starts and ends its section on measurement with ideas and concepts that are based on classical test theory." Looking beyond that, the most commonly used graduate level text in measurement for research psychologists is the venerable book by Nunnally (Nunnally & Bernstein, 1994). Examining the contents of the latest edition, one sees for example, that item response theory gets 20 pages (out of over 700). And many research psychologists do not even take a course at the standard of Nunnally's.

A second consequence, and one that in a sense works in the opposite way, is associated with the recent burgeoning of generalized statistical modeling tools, such as *gllamm*, various R packages such as *lmer*, etc. These hold great possibilities for psychometrics, but also, disconcertingly, much potential risk as well. As psychometricians pursue the more complex modeling armamentarium described above, they are finding a ready resource in these software tools, one that we cannot turn away from. However, the virtues of these tools, their ease of use and great range of possible application due to the variety of models that can be readily generated and estimated, means that the "home territory" of psychometrics is no longer a difficult field for others to enter. Sophisticated software users, and particularly statisticians, with no needed background in measurement etc., can analyze data and carry out just the same sort of statistical modeling studies that psychometricians tend to specialize in. That is, if a field of study is primarily defined by its tools, and if those tools are mainly the same as the tools of another discipline, then what is the difference between those two fields? This leads one to wonder: Is there a psychometrics that is different from (a minor branch of) statistics?

My view is that we should not (and cannot) shy away from the developments in statistical modeling. I think that we psychometricians know a great deal that relates to that special liminal domain that is at the interface between statistics and (various subdomains of) science, although we often dismiss it as "mere practice" using terms such as "Test Construction." It is a domain that most statisticians would like to ignore—epitomized by the "data comes in boxes" idea that this essay started with. If we get swallowed by statistics, then that special knowledge about measurement will wither (I will return to this in the next section). But in order to avoid this, we have to come to know ourselves and our roles more completely than we currently do.

## 4. An Alternative View of what Psychometricians Should Do and Should Study

This alternative view, with its four parts, is based primarily on my own *practice* of psychometrics as it has developed over the last 30 years, and hence, I will prioritize the pragmatic aspects

of the parts. Thus, this might look like no more than "what we do in consulting"—and that is a very good thing, as it grounds the ideas in what we are best at—but each part is also, in and of itself, a different field of study, though, of course, each is linked as part of this unified view of psychometric practice. Note, I have typically called these parts "building blocks" (Wilson, 2005). The four building blocks can be thought of as answers to a series of questions, as follows.

**Question 1** (What is it that we want to measure?). Psychometricians work with substantive researchers to conceptualize and establish the nature of the constructs that are to be measured. This is consistent with Borsboom's (2006) observation that measurement consists of "devising a model structure to relate an observable variable to a theoretical attribute." The answer to Question 1 is just that "theoretical attribute." This leads us to the study of:

(a) Good practice in developing conceptualizations *with* substantive researchers, and
(b) General approaches for creating/developing the *idea* of a construct, including the study of their advantages and disadvantages in particular scientific contexts.

**Question 2** (How should we prompt observations about the construct?). Psychometricians devise methods (i.e., instruments) that tap into those constructs. This leads us to the study of:

(a) Good practice in developing items/instruments/systems with substantive researchers, and
(b) General classes of types of items/instruments, including their advantages and disadvantages in particular scientific contexts.

**Question 3** (How should we code the results of the observations?). Psychometricians investigate how to relate the outcomes of the items/instruments to the underlying construct. This leads us to the study of:

(a) Good practice in developing binning/coding/scoring in collaboration with substantive researchers, and
(b) General classes of types of outcome spaces[2] including their advantages and disadvantages in particular scientific contexts.

**Question 4** (How should we model the data arising from the coding?). Psychometricians select and/or develop mathematical/statistical/logical models to analyze resulting data. This leads us to the study of:

(a) Good practice in developing new statistical models, paying close attention to embodying answers to the scientific questions of substantive researchers in the guise of estimable parameters of the models, and
(b) General classes of useful models including their advantages/disadvantages in particular scientific contexts.

This last makes it clear that the suggested approach is not to do *less* modeling—the other three building blocks constitute an *addition* to the current topics we generally study. One way to see this is to think of the first three building blocks as constituting a particular way to structure "Test Construction", and then the fourth could be considered as "Test Theory." However, the crucial aspect of this is the necessity for these to be integrated into a unifying logic. For this idea to be more than just a hobby-horse of psychometricians, it is crucial that at each step the

[2] An *outcome space* is a set of qualitatively described categories for recording and/or judging how respondents have responded to items (Marton, 1981; Wilson, 2005).

DaD6: Integrate case with aggregate perspectives.
DaD5: Consider the data in aggregate when interpreting or creating displays.
DaD4: Recognize or apply scale properties to the data.
DaD3: Notice or construct groups of similar values.
DaD2: Focus on individual values when constructing or interpreting displays of data.
DaD1: Create displays or interpret displays without reference to goals of data creation.

FIGURE 1.
The levels of the Data Display construct map.

psychometricians are sharing their work and conclusions with substantive researchers who are constructing tests. Hence, this message is relevant to both psychometricians, and to those who construct tests. And, in fact, if it were to be successful, then those two groups would become much less distinct.

Taken together, these four questions, with their associated building blocks, constitute a cycle of measurement practice that culminates in an outcome variable (sometimes called a "progress variable"). This set of steps may repeat several (even many!) times before a satisfactory conclusion is reached, with many occasions for feedback from one building block to another—In terms of a field of studies, this implies a need to study the operation of the whole cycle as well as its constituent parts.

In the next section, I will again digress, to provide an example of how this cycle of psychometric practice can work, providing concrete exemplifications of the four building blocks to help the reader conceptualize what a field of study centered around each building block might look like.

### *4.1. Interlude II: Measuring Middle School Students' Understanding of Data Display (DaD)*

As an example of an instrument development such as is outlined above, consider the work of the ADM project, a collaboration between researchers at Vanderbilt University (led by Rich Lehrer), and researchers at the BEAR Center at Berkeley (for further information, see Lehrer, Kim, Ayers, & Wilson, 2013). We have been working on assessments intended to support the development of middle school students' understandings of concepts and practices of data modeling. Data modeling refers to the invention and revision of models of chance to describe the variability inherent in particular processes. The construction of the assessments involved analysis of core concepts and practices of data modeling that were useful to students for learning, to teachers for developing their instructional practices, and to assessment developers for developing assessments. In all, seven constructs span the range of the curriculum, but for now, I will concentrate on just one: Data Display, which taps into students' representational competence.

**Question 1** (What is it that we want to measure?). Our initial steps towards defining the construct we eventually called Data Display (abbreviated "DaD"), involved the substantive description of "milestones" along the hypothesized DaD construct—generically called a *construct map*. The DaD construct map (see Figure 1), shows this sequence describing students' perceptions of data, specifically the ways they might think about interpreting or creating a display (e.g., a graph) as a way to improve their understanding of the situation. The main change as they go up the construct map is that there is a shift from a case-specific to an aggregate perspective of the data display, and the highest level describes an integration of the two perspectives. The milestones along the way that we have chosen to describe the construct are expressed in terms of displays of a single variable, but in general the progression shown is important to most forms of display.

At the lowest milestone labeled DaD1, students interpret displays as collections of values (numbers or categories), but they tend not to connect the form of the display to the intent of

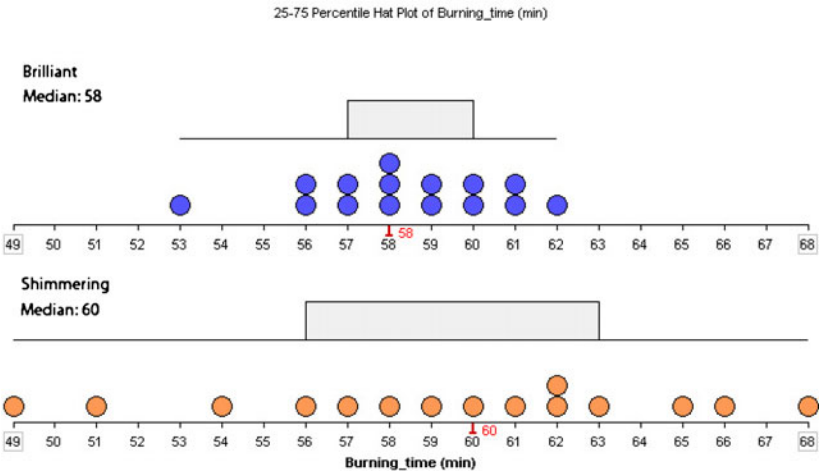
the designer of the display, such as by thinking about the question that motivated the display in the first case. At the next milestone, at DaD2, students interpret displays by thinking about specific cases. For example, they pay attention to what is common among the values (e.g., repeated values), the relative order of cases, or what makes them special (e.g., outliers). By DaD3, the students have begun to think about aggregates of cases: For example, when they are putting together a display, they might group together sets of similar values of cases, or when they are interpreting a display, they might identify clumps in the data. However, in doing so, they will often include "bins" (intervals) that are not of uniform size or that do not employ a continuous scale.

The milestone DaD4 signifies a move upward in sophistication, where they employ scale in thinking about grouping of data. When developing groupings of values, they make sure that the intervals are all the same, to make interpretation simpler. When interpreting a display, they note features in displays that allow for the highlighting of trends and they understand how a continuous scale might allow them to see "holes" or "gaps" in data. DaD5 marks a continuation of this shift involving quantification of the aggregates. For example, students might indicate the percentage of values in different groupings, or they may employ statistics to quantify aggregate qualities, such as spread, and then show that on the display. At the highest milestone, DaD6, students bring the case-perspective and the aggregate-perspective together. They can see cases as representative of segments of the data, and they can (informally) use data trends at the aggregate level to interpret individual cases (e.g., pointing out clusters of cases that might be inconsistent with the rest).

**Question 2** (How should we prompt observations about the construct?). Given that one has developed a construct map, such as in Figure 1, then the main feature of any item that one designs to prompt observations of that construct is that the responses can be mapped back to the levels of the construct map in a reasonably straightforward way. Of course, there will be many other characteristics that will need to be specified, such as the reading level of the text, and the interest–level among the population that is targeted, but these will vary from case to case. Consider for example, the sample item shown in Figure 2, labeled "Shimmering Candles." In this item, the focus of interest, of course, lies in the explanation that the student provides. The technique used is to ask the student to commit to a judgment about which candle company's product is better, then to see how sophisticated their explanations are in interpreting the data displays and the information they impart to support their argument. As such, there is an expectation that this item should prompt responses at all levels of the DaD construct, as shown in Figure 1. Of course, when designing an item, this is only a prediction—and it remains just that until we actually gather responses to the item and see how these relate to the construct map.

**Question 3** (How should we code the results of the observations?). A summary of the results that were obtained when we collected student responses is shown in Figure 3. The first two columns show the levels of the DaD construct, and the third shows examples of student responses. The results, as is most common, do not entirely match to our expectations—some things we might have expected do not show up, some things are more complicated than we expect. First of all, the fourth level, DaD4 does not appear at all. This is because of the type of question that was asked—the scale features were already embedded into the two example graphs, and hence, there is no way to observe different student understanding regarding use of scale. One conclusion from this is that in order to tap into DaD4, comparison-type items would need to include examples that differed in their use of scale (of course, other item-types may also tap into DaD4). Another notable feature is that there are two levels labeled "DaD5." What has happened is that the raters have observed that students have responded in two different ways both of which are judged to be at level DaD5, but that are systematically different. In fact, they have also judged that those labeled "DaD5b" are somewhat higher on the DaD construct than those labeled "DaD5a." A third notable feature

Shimmering Candle Company claims that their candles burn longer, on average, than candles made by the Brilliant Candle Company. Testers for Consumer News burned 15 Shimmering candles and 15 Brilliant candles and recorded the number of minutes that each candle burned. The plot below shows the burning time and the middle-50 percent of the data around the median.




1. Is the claim made by the Shimmering Candle Company, that their candles burn longer, supported by the test results? Circle your answer.

**Yes** **No**

Why do you think so?

FIGURE 2.
An example item related to the DaD construct: "Shimmering Candle."

is that the item has prompted responses at lower levels than are recorded in Figure 1. This is, of course, to be expected, as there are inevitably students who either fail to respond ("M"), or who simply "get it wrong." However, there was a little more that was noticed by the raters, that some students had responses with no relevant features (labeled "NL(i)", where NL = "no level"), some gave responses (labeled "NL(ii)") that were below DaD1 but that were at least related to the item. Although this was not a part of the overall scheme, it was decided to maintain this distinction in the coding.

**Question 4** (How should we model the data arising from the coding?). In the ADM project, the primary goal of selecting a measurement model has been to relate the student ability on each construct to the levels of the construct, in order to make the results easily interpretable, in particular, via graphical feedback to the students and their teachers. Thus, we require (ideally) that the order of item difficulties is the same for all respondents and second, we require that the order of respondents is the same for any subset of items. To accomplish this, we use a polytomous extension of the Rasch model (Rasch, 1961, 1980) because of its positive qualities for developing graphical interpretive tools (Wilson, 2005). Rasch-based modeling provides a convenient way to develop estimates of person proficiency and item difficulty using the same scale. This subsequently facilitates the interpretation of person proficiency estimates using criteria from the item content. An example of such a scale (known as a "Wright map") is shown in Figure 4—the estimation was carried out using the ConQuest software (Adams, Wu, & Wilson, 2012). In this example the histogram "on its side" on the left hand side shows the distribution of student locations estimated

| Levels | Response exemplars | Example of student response |
|---|---|---|
| DaD5b | Student quantifies aggregate property of the display using one or more of the following: ratio or proportion or percent. | • "50 % of Brilliant candles are between 57 and 60, and 50 % of Shimmering candles are between 56 and 63."* |
| DaD5a | Student recognizes that a display provides information about the data as a collective. | • No, b/c the majority of their burning times overlap.*<br>• No, b/c even though their media was larger, they were also more spread out.* |
| DaD3 | Notice or construct groups of similar values from distinct values. | • Yes, b/c the most of the Brilliant candles were between 57–59, while most of the shimmering candles were between 61–63.* |
| DaD2 | Concentrate on specific data points without relating these to any structure in data. | • "Because the last candle burned 6 min. more than Brilliant's last candle."<br>• "They have more with larger numbers."<br>• "The median is 58 for the other company, but Shimmering's median is 60."<br>• "Because Shimmering candle has more candles that burnt 58 long."<br>• "I think this because I added them up and got 875 minutes for Brilliant Candles and 953 for Shimmering Candles."*<br>• "Because I calculated their means." |
| DaD1 | Create or interpret data displays without relating to the goals of the inquiry.<br>Makes a numerical assertion that is present in the data but not related to assertions about burning time. | • "There are 30 measurements."* |
| NL(ii) | Responses are marginally related to the topic.<br>Either tells the topic of the display without giving any quantitative information or provide a qualitative assertion not supported by the display. (May include misinterpretation of the display, e.g., confuse $x$–$y$ axis.) | • "Because it's not the highest." |
| NL(i) | Irrelevant response | • "I don't know." |
| M | Missing response | |

FIGURE 3.
The outcome space related to the "Shimmering Candles" item.

by the model. To the right of that, the category thresholds for each item are shown, grouped by their respective DaD levels from Figure 3 (shown at the bottom)—each item threshold is indicated by a code for the item (for example "Shimmering Candles" is abbreviated "Candle"). The threshold locations are generally arrayed in a way that we would expect, with the sets showing a steady rise in mean difficulty from DaD1 to DaD5. The pattern becomes less strong at either end: at the lower end, the behavior of the NL(ii) thresholds is understood as being mainly due to factors relating to students background rather than their knowledge about Data Display; at the upper end, we note that few to no students are located up beyond the Dad5 thresholds, hence we must leave the accurate mapping of the DaD6 thresholds to a further data collection (and,

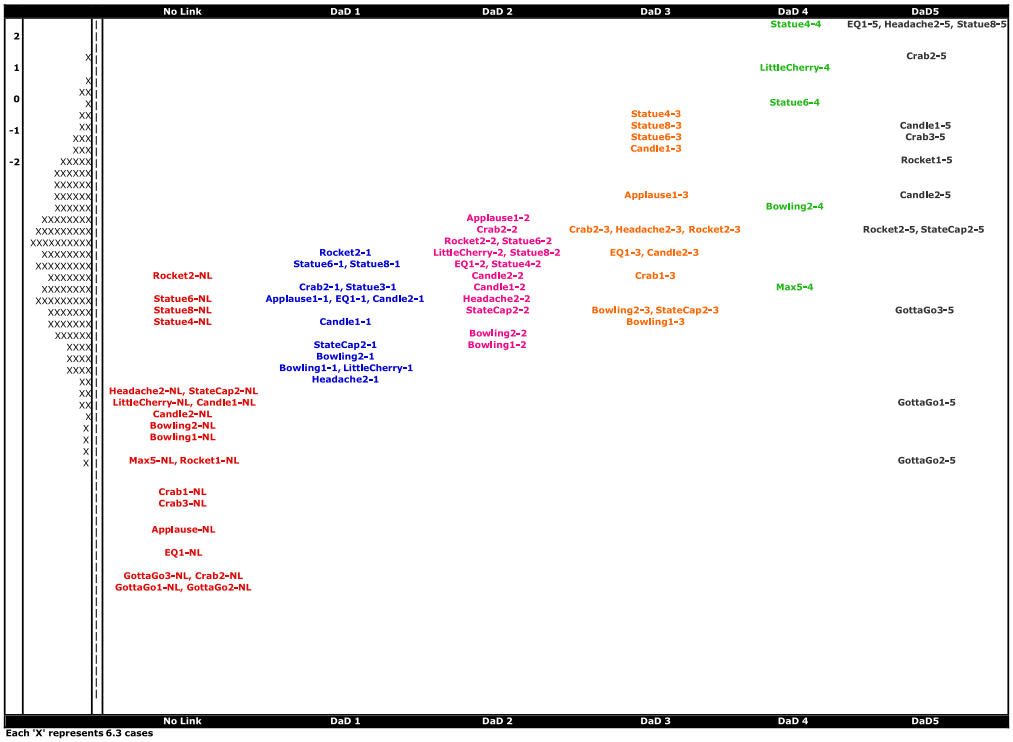



FIGURE 4.
Wright map of the DaD items.

in addition, the "GottaGo" item was eventually left out of the ADM item bank). This summary Wright map is then used as a basis for many different reports made available to the teachers and students about class and student progress.

The account above lays out the process of instrument development in a seemingly linear fashion, but that is really only an abstraction used to help explain the logic underlying the development process.[3] In fact, the process is more like a spiraling cycle, with the development iterating through the four questions, roughly in the order presented above, but also with many complications to that ordering. Our approach, then, is to fit the data to the construct (and to learn about both the construct and the measurement from when it does and does not fit well), rather than seeking a model to fit the data. Clearly, this makes the quality of the items design and the actual item development critical.

## 5. The Roles of Psychometricians, Continued (Again)

The use of these questions and their associated building blocks as a guide for constructing measures has been illustrated in the previous section. However, that does not illustrate the argument that we could also be using this same framework as a structure for our discipline. Put another way, although the previous section has illustrated each of the points under 1(a), 2(a), 3(a) and 4(a) (i.e., about psychometric *practices*) on page 12 above, it has not illustrated the

[3]This system, called the *BEAR Assessment System* (BAS), is described in Wilson (2005).

points under 1(b), 2(b), 3(b) and 4(b) (i.e., about the object of *psychometric studies*). Thus, that is the subject of this section and the next. In the next several paragraphs, I discuss the nature and contexts of possible studies that would be suitably placed into each of the four categories mentioned above. And, in the subsequent section, I illustrate these using a contemporary example in educational measurement.

The topic of *conceptualization of measurement variables* is probably the least quantitative of the fields of study. Psychometricians have typically shied away from this area of our work, possibly for exactly this reason. Yet, it is the most important, for without adequate conceptualization, all else is empty—in particular, the whole concept of "validity evidence" becomes moot, as there is no substance to validate to. It is also the aspect of our work that is most deeply embedded in the substantive scientific area in which the measurement is to take place. And this makes generalizations here difficult, as the idea of a variable will inevitably be the property of the scientific area wherein it arises. Yet, it has been my experience that, mostly, the scientists who inhabit that area do not have a clear idea of what a measurement variable could be. (And here I must note that the areas in which I work are almost entirely confined to the social sciences.) The native approaches that I observe when substantive scientists describe a new measurement variable tend to veer wildly from the strictly operational (e.g., the discussion of constructing the test is limited to describing "test specifications", where the substantive variable being measured is never defined except in an indirect way), to the loose and heterodox (e.g., the concept is conceived to be so complex and interconnected, one could never model it). Note that I am not implying that psychometricians need to be studying how to develop the substantive concepts (this rightly belongs in the domain itself), but rather the ways that one can find to attune those substantive concepts so that they are measurable. This is an area that I think needs our best efforts, as the current approach is largely hit-or-miss, and the cost involved in establishing whether one has indeed hit or missed is considerable. The generalized techniques described below under "outcome spaces" do help out with this, as they aim to give internal structure to a measurement variable, but they do not help provide the establishing *idea*.

In terms of *instrument development*, and specifically, of items design (Wilson, 2005), again, the development of generalized approaches is hampered (inevitably) by the need for the items to be grounded in the application areas of the substantive discipline. Nevertheless, some useful generalities have been developed. A good example, in my view is the typology of interview types described by Patton (1980). This ranges from the "informal conversational interview" to the "standardized open-ended interview." Patton gives general guidelines for which types are useful for different sorts of concepts, as well as for different sorts of social contexts. (For a generalization of this to broader item-types, see Chapter 3 of Wilson, 2005 and Scalise & Gifford, 2008.) The technique of "cognitive laboratories" (e.g., American Institutes for Research, 2000) is a recent example of the type of empirical work that can help with item development. In addition, there are various other techniques in response process studies (AERA/APA/NCME, 1999) available for investigating item types, such as think alouds, exit interviews, reaction time studies, eye movement studies and various treatment studies, where, for example, the respondents are given certain sorts of information before they are asked to respond.

In terms of the *outcome space*, that is, the relations between the item responses and the construct map (also termed coding and scoring), there have been some useful and interesting generalized schemes developed in the areas of psychology and education. One example is in the field of phenomenography, a method of constructing an outcome space for a cognitive task based on a detailed analysis of student responses. Phenomenographic analysis has its origins in the work of Marton (1981) who describes it as:

> a research method for mapping the qualitatively different ways in which people experience, conceptualize, perceive, and understand various aspects of, and phenomena in, the world around them (Marton, 1986, p. 31).

Phenomenographic analysis usually involves the presentation of an open-ended task, question, or problem designed to elicit information about an individual's understanding of a particular phenomenon. Most commonly, tasks are attempted in relatively unstructured interviews during which students are encouraged to explain their approach to the task or conception of the problem. Researchers have applied phenomenographic analysis to many learning areas.

A significant finding of these studies is that students' responses invariably reflect a limited number of qualitatively different ways of thinking about a phenomenon, concept or principle (Marton, 1988). A search is made for statements that are particularly revealing of a student's way of thinking about the phenomenon under discussion. These revealing statements, with details of the contexts in which they were made, are excerpted from the transcripts and assembled into a *pool of quotes* for the next step in the analysis. The focus of the analysis then shifts to the pool of quotes. Students' statements are read and assembled into groups:

> Bringing the quotes together develops the meaning of the category, and at the same time the evolving meaning of the category determines which quotes should be included and which should not. This means, of course, a tedious, time-consuming iterative procedure with repeated changes in the quotes brought together and in the exact meaning of each group of quotes (Marton, 1988, p. 198).

The result of the analysis is a grouping of quotes reflecting different kinds of understanding. These groupings become the outcome categories, which are then described and illustrated using sampled student quotes. Outcome categories are "usually presented in terms of some hierarchy: There is a *best* conception, and sometimes the other conceptions can be ordered along an evaluative dimension" (Marton, 1988, p. 195).

A second example is the SOLO (Structure Of the Learning Outcome) taxonomy, a general theoretical framework that may be used to construct an outcome space for a task related to cognition. The taxonomy, which is shown in Figure 5, was developed by Biggs and Collis (1982) to provide a frame of reference for judging and classifying students' responses. The taxonomy is based on Biggs and Collis' observation that attempts to allocate students to Piagetian stages and to then use these allocations to predict students' responses to tasks invariably results in unexpected observations (i.e., 'inconsistent' performances of individuals from task to task). The solution for Biggs and Collis is to shift the focus from a hierarchy of stages to a hierarchy of observable outcome categories: "The difficulty, from a practical point of view, can be resolved simply by shifting the label from the *student* to his *response* to a particular task" (1982, p. 22). Thus the SOLO levels "describe a particular performance at a particular time, and are not meant as labels to tag students" (1982, p. 23). In subsequent work using the taxonomy, several other useful levels have been developed. A problem in applying the SOLO Taxonomy was found—the "Multistructural" level tends to be quite a bit larger than the other levels—effectively, there are lots of ways to be partially correct. In order to improve the diagnostic uses of the levels, several intermediate levels parallel to the Multistructural one have been developed by the Berkeley Evaluation and Assessment Research (BEAR) Center (Wilson, 2005), and hence, three additional levels have been inserted in the same place as the Multistructural level—these are shown in Figure 6.

In terms of the *measurement model*, this is, of course, the category most well-represented in the work of psychometricians. Some thoughts towards a sequencing of developments in this area have been collected together in a volume by the National Research Council (2001, Chapter 4). Hence I will not spend time on it here, except to note that having these four categories of studies implies a fifth category, that is, the integration of the parts. Efforts to study such integration include the building blocks conceptualization discussed above, as well as the evidence-centered design (ECD) approach developed by Mislevy, Steinberg, and Almond (2003).

In the following section, I give an example that looks beyond the use of the four building blocks to measure a single construct.

An *extended abstract* response is one that not only includes all relevant pieces of information, but extends the response to integrate relevant pieces of information not in the stimulus.
A *relational* response integrates all relevant pieces of information from the stimulus.
A *multistructural* response is one that responds to several relevant pieces of information from the stimulus.
A *unistructural* response is one that responds to only one relevant piece of information from the stimulus.
A *pre-structural* response is one that consists only of irrelevant information.

FIGURE 5.
The SOLO Taxonomy.

A *semi-relational* response is one that integrates *some* (but not all) of the relevant pieces of information into a self-consistent whole.
A *multistructural* response is one that responds to several relevant pieces of information from the stimulus, and that relates them together, but that does not result in a self-consistent whole.
A *plural* response is one that responds to more than one relevant piece of information, but that does not succeed in relating them together.

FIGURE 6.
Three additional levels of the revised SOLO Taxonomy (between unistructural and relational levels).

## 6. Interlude III: Measuring Middle School Students' Understanding of Statistical Modeling

In the previous Interlude, we considered the measurement of a single construct, which we somewhat unceremoniously labeled DaD. This is surely the core activity of psychometrics. However, it is very often the case that researchers are interested in more than just one variable: for perfectly good researcher's reasons, they want to measure several constructs at the same time. In fact, in the ADM project, there are six constructs that span the learning statistical modeling curriculum. They are, first, DaD, which has already been described, plus:

2. Meta-Representational Competence (MRC): comparing and considering trade-offs among displays;
3. Conceptions of Statistics (CoS): characterizing distributions;
4. Chance (Cha): theoretical and empirical approaches to estimating probability;
5. Modeling Variability (MoV): constructing models of chance variation; and
6. Informal Inference (InI): model-based views of inference.

These are then rounded out with a seventh construct, Theory of Measurement (ToM: students theories about measuring), which is not formally a part of the statistical modeling itself, but rather is the area to which the statistical modeling is to be applied. These seven constructs have been (exhaustively!) worked through the building blocks system described above, and hence each has well-defined and exemplified levels available, as well as a collection of items for each.

In order to incorporate this multi-construct perspective into the ADM modeling, we next moved to using the multidimensional random coefficients multinomial logit (MRCML) model (Adams, Wilson, & Wang, 1997a, 1997b). This allows us to take advantage of the relationships among the dimensions to borrow information, and hence efficiently measure seven dimensions, which would otherwise prove too time-consuming in a classroom environment. This has allowed us to represent the whole set of dimensions as in Figure 7, where the dimensions are coordinated using a dimensional alignment technique (Schwartz, Ayers, & Wilson, 2010). This extends the

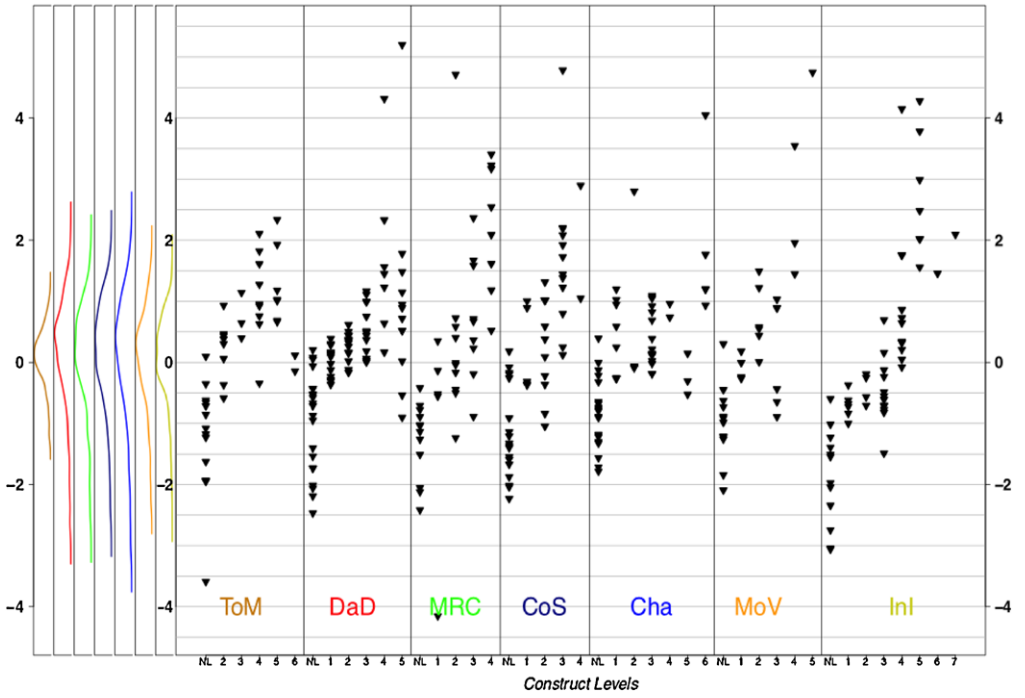



FIGURE 7.
Seven-dimensional Wright map for the ADM constructs.

interpretational reach of the measurements to a profile for each student, allowing a more complex assessment of each student's learning (but, of course, making greater demands on the teacher's interpretational skills).

This was indeed the principal aim of the modeling for the ADM project. However, the cognitive scientists involved were very much concerned that, although having a profile of constructs for each student was indeed one of their aims, they were also interested in another aspect of the modeling that had been invisible to the measurement up to that point. In the design of the ADM curriculum, the curriculum developers have incorporated into their instructional plans certain assumptions about relationships between the levels of different constructs. That is, they have assumed that advancement to one level of a construct, say CoS, was predicated in the successful understanding of a certain level of another construct, say DaD. (Of course, this is besides the obvious requirement that the student had also achieved the previous level of CoS.)

This concept has a broader background in studies in the learning sciences—corresponding to the concept of a "learning progression" (sometimes referred to as a "learning trajectory"). Based on a recent survey of studies in the topic area of learning progressions in science education, the following broad description was given:

> Learning progressions are descriptions of the successively more sophisticated ways of thinking about an important domain of knowledge and practice that can follow one another as children learn about and investigate a topic over a broad span of time. They are crucially dependent on instructional practices if they are to occur (Corcoran, Mosher, & Rogat, 2009, p. 37).

The description is deliberately broad, allowing a wide possibility of usage, but, at the same time, it is intended to reserve the term to mean something more than just an ordered set of ideas, curriculum pieces, or instructional events. As well, the group saw it as a *requirement* that the learning

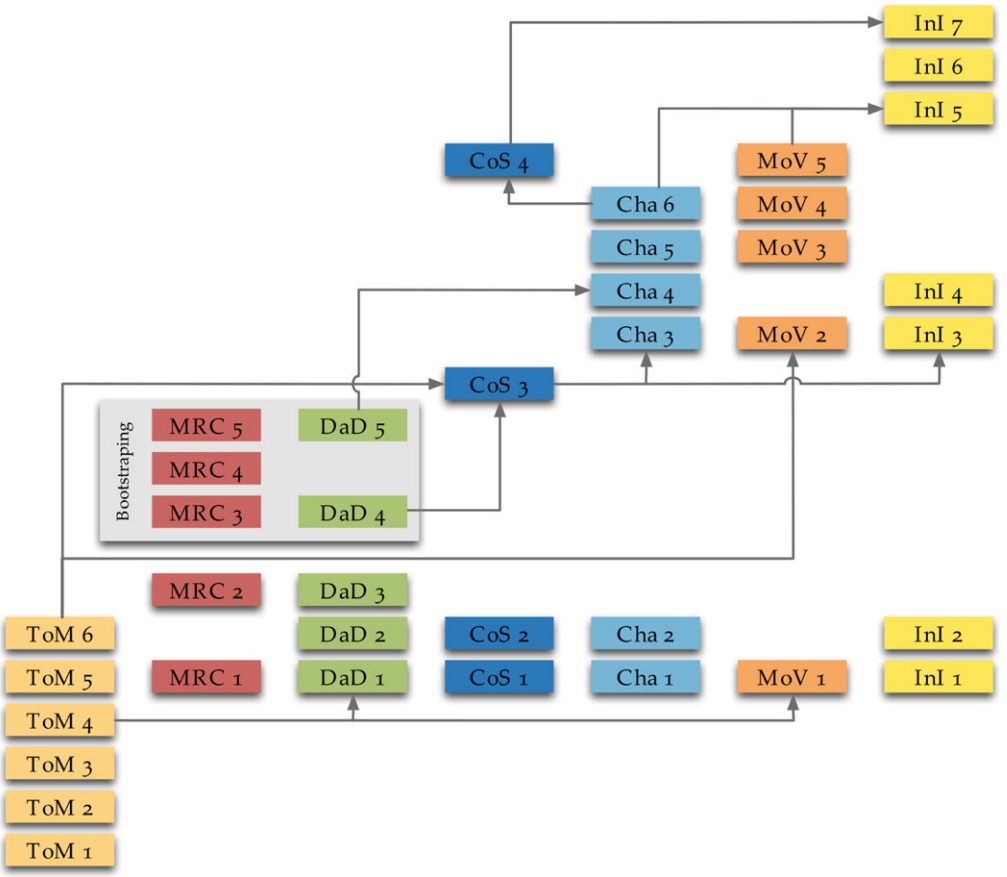



FIGURE 8.
The hypothesized links among the seven ADM constructs (note that the *labels* are explained in the text).

progression should indeed describe the "progress" through a series of levels of sophistication in the student's thinking, but still be broad enough to allow for complications such as non-linearity in the ordering, and the possibility that the order of levels might differ for different subgroups of students. A learning progression must be articulated in ways that coordinate efforts to link across the areas of substantive discipline, learning, instruction, and assessment, and it is the last that we are primarily concerned with here in this paper (although, of course, it is learning that is the ultimate goal of the learning progression). The concept of a learning progression is related to other concepts that have been proposed in the areas of curriculum theory, learning science, and cognitive modeling. Different modeling approaches to these concepts have been suggested by others such as Falmange (e.g., Falmagne & Doignon, 2011), and researchers working within the area of cognitive diagnostic modeling (CDM; e.g., Rupp, Templin, & Henson, 2010). Below, I describe a somewhat different approach grounded in the collaborative work in the ADM project (Wilson, 2009).

Thus, returning to the ADM example, an illustration of the current state of the "link" hypotheses that have been postulated is shown in Figure 8. Note that this figure includes links of the type described above, but also represents something extra—"bootstrapping"—this represents a hypothesis of "mutual support" among levels of two constructs.

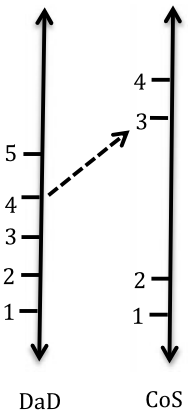



FIGURE 9.
Hypothesized relationship between DaD and Cos.

Looking a bit closer now, let us consider the relationship between DaD3 and CoS3[4] (see Figure 9).

In order to empirically test out this hypothesized link, as well as to investigate the potential impacts on the measurement of the constructs of the multidimensional measurement model implicit in the hypotheses, we have engaged in the development of a class of psychometric models, which we call *structured construct models* (SCMs; Wilson, 2012). The central idea of these models is illustrated in Figure 9, which shows the two constructs mentioned above, one of which (Level 3 of the variable, CoS, is the *target* level) has a level that requires a level of another (Level 4 of the variable, DaD, is the *requirement* level). By this requirement we mean that, for a student to succeed at Level 3 of CoS, s/he must not only have learned the content of Level 2 of CoS (which is the usual developmental hypothesis for a construct map), but also Level 4 of DaD. In considering this diagram, there are three questions that need to be considered:

(a) what is the nature of the levels?
(b) what do we mean by the links?
and
(c) how do we model them jointly?

Different answers to these questions will correspond to different assumptions about the substantive situation. For example, although the modeling in the previous Interlude assumed that the underlying constructs were continua, that is not necessarily the case; instead, the levels of the constructs could also be thought of as ordered latent classes, or, indeed located latent classes. Moreover, the nature of the link can also be thought of in different ways, for example, it could be deterministic, depending simply on the presence or absence of a student in the requirement level, or it could be probabilistic, depending in addition on characteristics of the level and/or the underlying construct.

In terms of a mathematical expression for these possibilities, SCMs can be decomposed into two parts, a model for the responses on each construct and a model for the person locations on each construct (including the effects of any links). Suppose we have a response vector for a person $p$, $\boldsymbol{X}_p = \boldsymbol{x}_p$. Then the likelihood of that response vector is given by

$$\Pr(\boldsymbol{X}_p = \boldsymbol{x}_p) = \sum_{r,t} \pi_{r,t} \prod_i \pi_{(ij|r,t)}, \tag{4}$$

[4]This level of CoS is summarized as: "Consider statistics as measures of characteristics of a sample distribution."

where $\boldsymbol{\pi}_{r,t}$ is the joint probability of person $p$ in requirement class $r$ of construct $\theta^R$ (e.g., DaD4 in Figure 9) and target class $t$ of construct $\theta^T$ (e.g., CoS3 in Figure 9) the sum $\sum_{r,t}$ is over all possible combinations of class membership on the constructs $\theta^R$ and $\theta^T$ (where for continuous constructs, the summation would be an integral), $\pi_{(ij|r,t)}$ is the conditional item response probability of person $p$ in requirement class $r$ of construct $\theta^R$ and target class $t$ of construct $\theta^T$, and the product $\prod_i$ is over the response probabilities of all the items (assuming local independence). Thus, our response to the first question above is that, potentially, the levels may be thought of as (ordered) latent classes, or as parts of continua, although we will focus here on the former. Now turn to the second question: How to conceptualize the hypothesized link. For a latent class approach, one possibility would be to set the joint probabilities for any pair "above" $(r, t)$ to be 0.0:

$$\text{if } r' < r \text{ and } t' \geq t, \text{ then } \boldsymbol{\pi}_{r't'} = 0. \tag{5}$$

However, this might be too restrictive, and hence, one could set this probability to be less than a small number $\varepsilon$. Or, perhaps just that each higher probability was less than that for the pair "below." For a continuous approach, the probabilities could be modeled as governed by a cut point associated with the requirement level $r$, and the probabilities could be modeled in a similar away as for the latent class case, as being discrete at that point, or as a suitable function depending on proximity to the cut point.

In joint work with two current graduate students at the University of California, Berkeley (Ronli Diakow and David Torres Irribarra), we are currently investigating the estimability and usefulness of these SCM models in a range of possible situations, and across some different ways of conceptualizing them. So far we have only pursued the first possibility mentioned above, that the joint probability is zero. Another concern is the appropriate level of measurement for the constructs. Our initial work has focused on assuming that a latent class approach makes sense, as that makes for a straightforward relationship between the measurement and the links (i.e., the former as latent classes, and the latter as links between those classes). This leaves open the question as to whether these latent classes are merely ordered, or located (and ordered, of course), and, if located, whether they are homogeneous or not. We are currently working our way through these possibilities. Lastly, in the interests of consistency with the previous work on construct maps, we have begun considering whether continuous models could also be thought of in this light. This work has been reported at the 2011 and 2012 Psychometric Society meetings (Diakow & Irribarra, 2011; Diakow, Irribarra, & Wilson, 2012b) and at several other recent meetings as well (Diakow, Irribarra, & Wilson, 2011; Diakow, Irribarra, & Wilson, 2012a; Irribarra, Diakow, & Wilson, 2012).

Although this is ongoing work, and not yet complete enough for full publication, I include a brief summary of some partial results, in order to illustrate some of the our approach. We used a subset of the data from Schwartz et al. (2010). It consists of responses to the items from the CoS and Cha constructs for 847 middle school students. There were 18 items for the CoS construct and 23 items for the Cha construct. The items were scored using scoring guides similar to the one in Figure 3. While the levels on the scoring guides were actually more fine-grained than the construct levels as represented in Figure 1 (e.g. the scoring guides distinguished between performances labeled DaD5a and DaD5b), scores were collapsed for analysis according to overall construct level (e.g. DaD5a and DaD5b were both coded as 5). Scores of NL(i) and NL(ii) were coded 0; scores of M were coded as missing. For CoS items, the maximum possible score on any item was 4; for Cha items, the maximum possible item score was 6. However, not every score could be obtained for every item. In particular, the actual maximum possible score varied across items. The results below have been reported in Diakow et al. (2011).

CoS and Cha were selected to illustrate this methodology because, shown in Figure 8, the links between them represent interesting bidirectional relations. The arrows depicted in Figure 8

represent the following two links from one construct to the other: (a) a link from CoS Level 3 to Cha Level 3. A student cannot, in theory, reach Level 3 on the Cha construct unless he/she has reached Level 3 on the CoS construct. (b) A link from Cha Level 6 to Cos Level 4. A student cannot, in theory, reach Level 4 on the CoS construct unless he/she has reached Level 6 on the Cha construct. In order to distinguish between the constructs involved in these links, we refer to the construct from which the link originates as the 'requirement' and to the construct at which the link terminates as the 'target'. For the first link, CoS is the requirement construct and Cha is the target construct; for the second link, Cha is the required construct and CoS is the target construct.

To demonstrate the use of the SCM models, we fit a set of four SCM models (Equations (4) and (5)) that applied the hypothesized constraints between the CoS and Cha constructs as follows:

(a) an unconstrained model,
(b) a model applying the constraint from CoS3 to Cha3,
(c) a model applying the constraint from Cha6 to CoS4, and
(d) a model applying both constraints between CoS and Cha.

For each of these models, we assumed four latent classes for CoS and 6 latent classes for Cha, which corresponds to the theoretical learning progression. In doing so, we tacitly assumed that the latent classes were a correct specification of the levels on each construct. In other words, we assumed both that the number of ordered latent classes matched the number of hypothesized levels for each construct and, more importantly, that each ordered latent class corresponded reasonably well to one of the theoretically defined levels. We made these rather strong assumptions due to the "proof-of-concept" quality of our initial goal for the analysis (but we have also investigated the validity of these assumptions Diakow et al., 2012b). To demonstrate how SCMs could be used to test the hypothesized links, we compared each of the singly constrained models to the unstructured model and the doubly constrained model to each of the singly constrained ones. We took a holistic approach and evaluated the models across a number of aspects. Such a flexible approach was advocated by Spiegelhalter, Best, Carlin, and van der Linde (2002) in their initial presentation of the Deviance Information Criterion (DIC).

To compare the models, we examined (a) the Bayesian Information Criterion (BIC) and number of activated constraints, (b) the joint probabilities of latent classification $\pi_{\text{rt}}$, and (c) the person classification. In terms of more formal criteria, we obtained the BIC and number of activated constraints (AC) for each model. We used Latent Gold (Vermunt & Magidson, 2007) to estimate the models, which calculates BIC as follows:

$$\text{BIC}_{\text{LG}} = -2\log L + (\log P)N_{\text{par}}, \tag{6}$$

where $\log L$ is the log-likelihood evaluated at the maximum likelihood, $P$ is the number of persons (here, always 847), and $N_{\text{par}}$ is the number of parameters. The number of activated constraints (AC) is the number of parameters constrained by the inequality restrictions imposed by the ordered latent class part of the model. In general, a better fitting model is one with lower BIC and fewer activated constraints.

We then examined how the estimated, model-based classification probabilities (i.e. the joint probabilities $\pi_{\text{rt}}$) changed as additional constraints were applied. We graphed the $\pi_{\text{rt}}$ for each model and compared them across the models. The better the constraint is reflected by the data, the less the joint probabilities will differ from the unconstrained model when the constraint is applied. We also considered how the classification of individual persons changed across the models. For each person, we obtained the modal latent classification (i.e. the latent class for which that person had the highest posterior latent class membership probability based on his/her item responses). This is the standard classification method in latent class analysis (Vermunt & Magidson, 2007). We graphically compared how person classification changed across the models as the

TABLE 1.
Comparison of SCM models.

| Model | Parameters | AC | BIC |
|---|---|---|---|
| (a) Unconstrained | 197 | 67 | 12671.07 |
| (b) Cos $\Rightarrow$ Cha | 192 | 64 | 12741.22 |
| (c) Cha $\Rightarrow$ Cos | 184 | 76 | 12626.60 |
| (d) Cha $\Leftrightarrow$ Cos | 187 | 64 | 12764.83 |

different constraints were applied. If a constraint is appropriate given the data, we would expect the modal classification not to change much when it is applied.

In order for the model comparison to be meaningful, we needed to consider the meaning of the latent classes. This involved examining the assumptions mentioned above. We needed to see if the latent classes matched the theoretical levels in the learning progression to know if the constraints being applied matched the hypothesized links between constructs. In general, it is also necessary to evaluate the stability of the latent classes when applying the constraints. If the redistribution of persons across the latent classes alters the meaning of the latent classes, the different models might not be directly comparable. This is also a potential sign that the constraint is inappropriate for modeling the data.

Comparison of models (a) to (d) above, based on BIC, (see Table 1) indicates that model (c) provides the best fit of the four models. This result is to some extent unsurprising when the model results are represented as joint probabilities in Figure 10, where it is possible to see in the upper left panel (the unconstrained model) that the cells where the constraint from Cha over CoS is applies have a very small proportion of people, making it very similar to the pattern of results observed for model (c) in lower left panel. In contrast, the pattern of results of the constraint from CoS over Cha, presented in the upper right panel, shows a clearly different pattern from the unconstrained model. Consequently, the lower right panel, which implements both sets of constraints, also shows a markedly different pattern from the one observed in the unconstrained mode.

Another aspect that can be examined when interpreting these results corresponds to the way in which the cases are redistributed across the classes when the constraints are applied. To find a graphical representation of the movement of persons from one class to another, see Diakow et al. (2011). Overall, the examination of the changes in person classification from model to model shown in Diakow et al. (2011) yields a similar conclusion to the one based on overall model fit, namely, that the constraint from Cha to CoS fits the data considerably better. Nevertheless, these figures also raise some questions regarding how stable the classes are, especially considering that under the model that exhibits the best fit, one of the classes in the Chance learning progression appears to have no members according to the model. Further exploration of these results, along with detailed ordered latent class analysis of the number of classes, and their interpretation according to the expected learning performances defined by the ADM framework, indicate that the proficiency levels of the students in the sample do not capture the full range of the theoretical progress variables. Thus, although the methods have proved useful to reveal interesting features of the data, the ADM project needs to go back and collect more data from students who have mastered higher levels of the constructs, before an adequate test of the hypotheses in Figure 8 can be carried out.

The point, for this essay, is not so much to show complete results of estimating these models (there are far too many details to fit into these pages anyway), but rather to point up the way that pursuing the first three building blocks has opened up a new line of inquiry with respect to the fourth, modeling. In addition, one can see that the speculation about the fourth, in terms of the links, has also brought about consideration of alternative ways to conceptualize (a) the

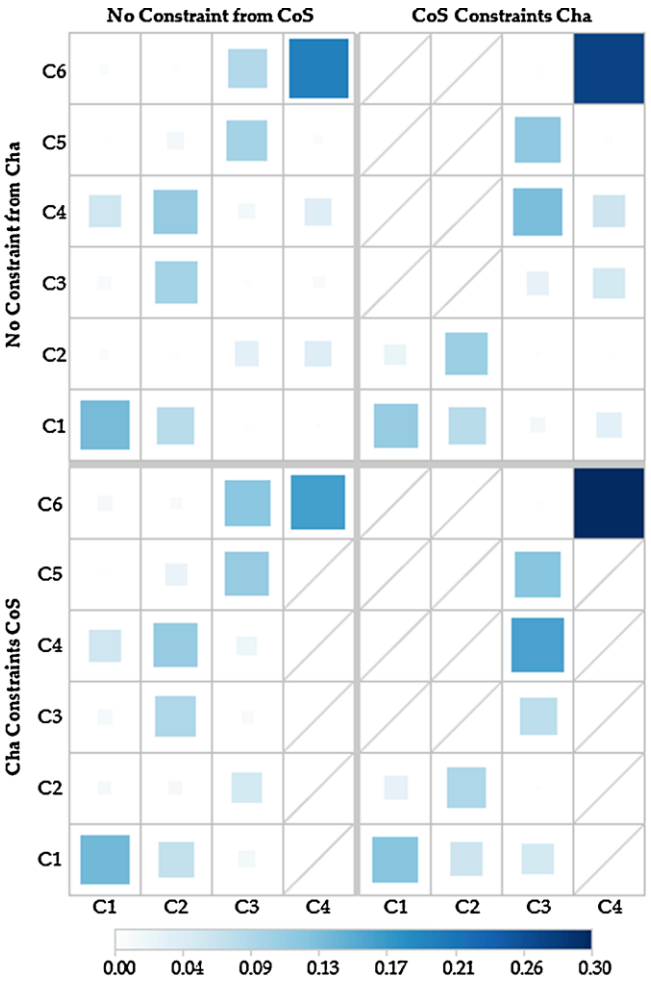



FIGURE 10.
Results of fitting SCM models (a) to (d), expressed as joint probabilities.

first, the construct map, in that a new idea, of bootstrapping has arisen (and is not yet entirely resolved), and (b) the third, the outcome space, which was initially considered a mapping into segments of a continuum, and then re-conceptualized as mapping into a latent class (and perhaps also, back again to think of it as a continuum). Thus, while the representation of the construct clearly depends on its development through the earlier building blocks, and there are ample possibilities to generate alternative modeling strategies, we see possibilities for the influence to then rebound, with modeling possibilities affecting the earlier building blocks. Of course, there are many possible paths of feedback among the building blocks, and across iterations of a given development cycle (see Brown & Wilson, 2011 for a summary of several such possible links).

## 7. Conclusion

The thoughts and examples above are intended to help the field (and myself) think about the roles that psychometricians should see ourselves as playing, and the topics that we write about in our journals. It is my view that the practices we engage in when we are acting as consultants form

a broader and richer range of possible subjects for our discipline than is typically displayed in our journals, our conferences, and other forms of communication. We need to transcend the statistical modeling "box" of our thinking that currently proscribes what is seen as "serious" work. The first step towards this is for us to "see" our work in its entirety. We need to accord the parts of our work that necessarily precede modeling the same standing as we currently give only to modeling. In this paper, by examining my own work and practices over the last 30 years, I have sought to re-evaluate what I see as important about that work. I hope that this will encourage others to do so too. I do not expect that the specific ideas and structures I have laid out here encompass the entirety of what we could be doing as psychometricians. Others will have different ways to structure their ideas about this work, and those thoughts need to be heard also.

One context in which these thoughts and ideas take on a particularly important shade is the design of our entry-level training, effectively the syllabus of the graduate training we give to newcomers to our field. The concerns I have registered above, and the (albeit partial) solutions I have assayed, have, in large part, been driven by my experiences as a teacher, and as someone closely observing students learning our science and craft. I have learned, for instance, that the very best examples that I can use to help a student understand measurement concepts and issues are the examples that the student values most highly, preferably measurement examples with which they themselves are currently deeply engaged, and if possible, examples that they themselves are the primary authors on. My way to enact this in my own classes has been to reduce the prominence of the examples that I bring to the class (although I still have lots of examples that I bring), and raise to the foreground the creative context in which each student is working—thus, in my introductory class (see below), each student creates their own new instrument, and it is used as the context for all of the major learning that takes place in the class.

Beyond tactics such as this, I also have found it useful to change around the order and nature of the typical sequence of psychometrics courses. In what I see in many places, the sequence of classes goes something like this: (a) a sequence of modeling classes, starting with classical test theory, and advancing through more and more complex modeling courses, until the student reaches the current state of play in the professor's work; and, at some point along this sequence, (b) a practical class on "test construction." Clearly, the weight of classes, and the general direction of this plan, leads our students to see modeling as our principal product, and our highest achievement. Instead, I have re-programmed my own classes in psychometrics, and placed first and foremost a foundational class that uses the context of a practical instrument development to help my students experience and appreciate ALL four of the building blocks: conceptual development, instrument design, outcome space design, and modeling. I see this as essential to their future growth as psychometricians, and as providing a launching pad for both a succeeding set of courses, and their adoption of certain areas and topics as their own chosen specialties for their thesis and dissertation work. Subsequent classes are designed to then take on particular issues. Surely, our entry level students often come to us with little background in statistical modeling, hence concentrated classes on modeling are still needed. But beyond this, I have endeavored to base courses on larger and perennial issues that arise as we pursue sound measurement: topics such as multidimensionality, explanatory perspectives, "multilevelness," class vs. continuum, etc. Unfortunately, I have the resources to only address the first few of these on a regular basis, but I attempt others in special "research seminars" when it is feasible. One area in which I fall down is that I do not have the resources to mount courses on the policy-related issues that arise as we work as psychometricians.

Beyond our teaching of newcomers is, of course, our own professional communications—our journals and our conferences. What I see as the bounded nature of our focus for publication and presentation needs to be lifted, so that a broader range of types of psychometric paper can be (and are) published. These would include (a) foundational papers about the philosophy and means of conceptualizing measurable variables, (b) papers focused on specific areas of observation, that would investigate generic types of instrumentation for measuring, (c) papers that get

"under the cover" of coding, scoring, and response processes, to lay out the logic of these crucial processes, and, of course, (d) a continuing stream of articles about rich and relevant models and the issues related to their usage.

Papers that look across these categories are needed too, as well as papers that focus on specific application areas to investigate the affordances of different techniques in particular areas. We seldom write papers like this. One reason is that we do not yet have established genres of and styles for these different types of issues. Thus, exemplary papers are needed for each of these types (and other types not mentioned above). One bright spot is the recent creation of the *Application Reviews and Case Studies* Section of *Psychometrika*. This section, founded by visionary colleagues of ours, offers a home for just the sort of experimentation in researching and writing that is called for above. So far it has yet to draw as much attention as its older sibling, but I am hoping that the ideas in this paper will help with that problem.

Some of this may sound like a diatribe against modeling, but that is far from my intent. My wish is to embed modeling in the natural place in which it sits in our work as psychometricians. Indeed, I see that, by allocating due attention to issues of variable conceptualization, item design and outcome space design, as well as modeling, we will enrich the nature and range of our modeling (as I hope was clearly the case in the third interlude, above). It is my view that efforts to enlarge the scope of the discipline of psychometrics, such as those discussed in the text above, would help improve psychometrics (a) by giving us a deeper and richer set of issues to work on, (b) by enriching our connections to substantive researchers, whose sciences are the well-spring for all of our own work and (c) by giving us a framework to distinguish our work from that of the statisticians.

I hope that this paper has proved stimulating, and that I will be subjected to many complaints and comments about it. We need to be thinking about nature of our work, and maintain our discussion of alternatives and possibilities.

## Acknowledgements

Many colleagues have contributed to the thoughts and ideas presented in this paper—unfortunately, I cannot acknowledge all of you. Hence, I restrict my acknowledgements to two groups. First, those who commented on drafts of the text: Ronli Diakow, Paul De Boeck, Karen Draney, Andy Maul, Roger Millsap, and David Torres Irribarra. Second, those who worked directly on the examples used in the text: for the saltus example, Karen Draney and Bob Mislevy; for the ADM example, Beth Ayers, Kristen Burmester, Tzur Karelitz, Rich Lehrer, David Torres Irribarra, Kavita Seeratan and Bob Schwartz; and for the SCM example, Ronli Diakow, and David Torres Irribarra. Any errors or omissions are, of course, the responsibility of the author.

## Appendix: Publications Related to the Saltus Model (in Chronological Order)

21. Draney, K., & Jeon, M. (2011). Investigating the saltus model as a tool for setting standards. *Psychological Test and Assessment Modeling, 53*(4), 486–498.
20. Draney, K., Wilson, M., Gluck, J., & Spiel, C. (2008). Mixture models in a developmental context. In G.R. Hancock & K.M. Samuelson (Eds.), *Advances in latent variable mixture models* (pp. 199–216). Charlotte: Information Age Publishing.
19. Draney, K., & Wilson, M. (2007). Application of the saltus model to stage-like data: some applications and current developments. In M. von Davier & C. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 119–130). New York: Springer.

18. Draney, K. (2007). Understanding Rasch measurement: the saltus model applied to proportional reasoning data. *Journal of Applied Measurement, 8*.
17. Demetriou, A., & Kyriakides, L. (2006). The functional and developmental organization of cognitive developmental sequences. *British Journal of Educational Psychology, 76*(2), 209–242.
16. Acton, G.S., Kunz, J.D., Wilson, M., & Hall, S.M. (2005). The construct of internalization: conceptualization, measurement, and prediction of smoking treatment outcome. *Psychological Medicine, 35*, 395–408.
15. De Boeck, P., Wilson, M., & Acton, G.S. (2005). A conceptual and psychometric framework for distinguishing categories and dimensions. *Psychological Review, 112*(1), 129–158.
14. Draney, K., & Wilson, M. (2004). Application of the polytomous saltus model to stage-like data. In A. van der Ark, M. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences*. Mahwah: Erlbaum.
13. Fieuws, S., Spiessens, B., & Draney, K. (2004). Mixture models. In P. De Boeck & M. Wilson, (Eds.), *Explanatory item response models: a generalized linear and nonlinear approach* (pp. 317–340). New York: Springer.
12. Pirolli, P., & Wilson, M. (1998). A theory of the measurement of knowledge content, access, and learning. *Psychological Review, 105*(1), 58–82.
11. Wilson, M., & Draney, K. (1997). Partial credit in a developmental context: the case for adopting a mixture model approach. In M. Wilson, G. Engelhard, & K. Draney (Eds.), *Objective measurement IV: theory into practice* (pp. 333–350). Norwood: Ablex.
10. Draney, K.L., & Wilson, M. (1997). *PC-saltus [computer program]*. BEAR Center Research Report, UC Berkeley.
9. Mislevy, R.J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika, 61*(1), 41–71.
8. Draney, K.L. (1996). *The polytomous saltus model: a mixture model approach to the diagnosis of developmental differences*. Unpublished doctoral dissertation, UC Berkeley.
7. Wilson, M. (1994). Measurement of developmental levels. In T. Husen & T.N. Postlethwaite (Eds.), *International encyclopedia of education* (2nd ed., pp. 1508–1514). Oxford: Pergamon Press.
6. Wilson, M. (1993). The "Saltus model" misunderstood. *Methodika 7*, 1–4.
5. Wilson, M. (1990). Measurement of developmental levels. In T. Husen & T.N. Postlethwaite (Eds.), *International encyclopedia of education: research and studies. Supplementary volume 2*. Oxford: Pergamon Press.
4. Demetriou, A., & Efklides, A. (1989). The person's conception of the structures of developing intellect: early adolescence to middle age. *Genetic, Social, and General Psychology Monographs, 115*, 371–423.
3. Wilson, M. (1989). Saltus: a psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105*(2), 276–289.
2. Wilson, M. (1985). *Measuring stages of growth*, ACER occasional paper, No. 19. Melbourne, Australia: ACER.
1. Wilson, M. (1984). *A psychometric model of hierarchical development*. Unpublished doctoral dissertation, University of Chicago.

References

Adams, R.J., Wilson, M., & Wu, M. (1997a). Multilevel item response models: an approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, *22*(1), 47–76.

Adams, R.J., Wilson, M., & Wang, W.C. (1997b). The multidimensional random coefficients multinomial logit. *Applied Psychological Measurement*, *21*, 1–23.

Adams, R.J., Wu, M., & Wilson, M. (2012). *ConQuest 3.0 [computer program]*. Hawthorn, Australia: ACER.

Acton, G.S., Kunz, J.D., Wilson, M., & Hall, S.M. (2005). The construct of internalization: conceptualization, measurement, and prediction of smoking treatment outcome. *Psychological Medicine*, *35*, 395–408.

American Educational Research Association, American Psychological Association, National Council for Measurement in Education (AERA, APA, NCME) (1999). *Standards for educational and psychological testing*. Washington: American Educational Research Association.

American Institutes for Research (2000). *Voluntary national test, cognitive laboratory report, year 2*. Palo Alto: American Institutes for Research.

Biggs, J.B., & Collis, K.F. (1982). *Evaluating the quality of learning: the SOLO taxonomy*. New York: Academic Press.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425–440.

Brown, N.J.S., & Wilson, M. (2011). Model of cognition: the missing cornerstone of assessment. *Educational Psychology Review*, *23*(2), 221–234.

Corcoran, T., Mosher, F.A., & Rogat, A. (2009). *Learning progressions in science: an evidence-based approach to reform* (CPRE Research Report #RR-63). New York: Center on Continuous Instructional Improvement, Teachers College—Columbia University.

De Boeck, P., Wilson, M., & Acton, G.S. (2005). A conceptual and psychometric framework for distinguishing categories and dimensions. *Psychological Review*, *112*(1), 129–158.

Demetriou, A., & Efklides, A. (1989). The person's conception of the structures of developing intellect: early adolescence to middle age. *Genetic, Social, and General Psychology Monographs*, *115*, 371–423.

Demetriou, A., & Kyriakides, L. (2006). The functional and developmental organization of cognitive developmental sequences. *British Journal of Educational Psychology*, *76*(2), 209–242.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, *39*, 1–38.

Diakow, R., & Irribarra, D.T. (2011). *Developing assessments of data modeling and mapping a learning progression using a structured constructs model*. Paper presented at the international meeting of the psychometric society, Hong Kong, July 2011.

Diakow, R., Irribarra, D.T., & Wilson, M. (2011). *Analyzing the complex structure of a learning progression: structured construct models*. Paper presented at the annual meeting of the national council of measurement in education, New Orleans, LA, April 2011.

Diakow, R., Irribarra, D.T., & Wilson, M. (2012a). *Analyzing the complex structure of a learning progression: structured construct models*. Paper presented at the national council on measurement in education annual meeting, Vancouver, Canada, April 2012.

Diakow, R., Irribarra, D.T., & Wilson, M. (2012b). *Evaluating the impact of alternative models for between and within construct relations*. Paper presented at the international meeting of the psychometric society, Lincoln, Nebraska, July 2012.

Draney, K. (1996). *The polytomous saltus model: a mixture model approach to the diagnosis of developmental differences*. Unpublished doctoral dissertation, University of California, Berkeley.

Draney, K., & Jeon, M. (2011). Investigating the saltus model as a tool for setting standards. *Psychological Test and Assessment Modeling*, *53*(4), 486–498.

Draney, K., & Wilson, M. (2004). Application of the polytomous saltus model to stage-like data. In A. van der Ark, M. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences*. Mahwah: Erlbaum.

Falmagne, J.-C., & Doignon, J.-P. (2011). *Learning spaces*. Heidelberg: Springer.

Fischer, K.W., Pipp, S.L., & Bullock, D. (1984). Detecting discontinuities in development: methods and measurement. In R.N. Emde & R. Harmon (Eds.), *Continuities and discontinuities in development*. Norwood: Ablex.

Irribarra, D.T., Diakow, R., & Wilson, M. (2012). *Alternative specifications for structured construct models*. Paper presented at the IOMW 2012 conference, Vancouver, April 2012.

Lehrer, R., Kim, M.-J., Ayers, E., & Wilson, M. (2013, in press). Toward establishing a learning progression to support the development of statistical reasoning. In J. Confrey & A. Maloney (Eds.), *Learning over time: learning trajectories in mathematics education*. Charlotte: Information Age Publishers.

Marton, F. (1981). Phenomenography: describing conceptions of the world around us. *Instructional Science*, *10*, 177–200.

Marton, F. (1986). Phenomenography—a research approach to investigating different understandings of reality. *Journal of Thought*, *21*, 29–49.

Marton, F. (1988). Phenomenography—exploring different conceptions of reality. In D. Fetterman (Ed.), *Qualitative approaches to evaluation in education* (pp. 176–205). New York: Praeger.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement Interdisciplinary Research & Perspective*, *1*, 3–67.

Mislevy, R.J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika*, *61*, 41–71.

National Research Council (2001). *Knowing what students know: the science and design of educational assessment*. Committee on the Foundations of Assessment, J. Pellegrino, N. Chudowsky, & R. Glaser (Eds.), Washington: National Academy Press.

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Patton, M.Q. (1980). *Qualitative evaluation methods*. Beverly Hills: Sage.

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

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (Vol. 4, pp. 321–334).

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press (original work published 1960).

Rost, J. (1990). Rasch models in latent classes: an integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282.

Rupp, A.A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: theory, methods, and applications*. New York: The Guilford Press.

Scalise, K., & Gifford, B.R. (2008). *Innovative item types: intermediate constraint questions and tasks for computer-based testing*. Paper presented at the national council on measurement in education (NCME), session on 'Building adaptive and other computer-based tests', in New York, May 2008.

Schwartz, R., Ayers, E., & Wilson, M. (2010). *Modeling a multi-dimensional learning progression*. Paper presented at the annual meeting of the American educational research association, Denver, CO, April 2010.

Siegler, R.S. (1981). Developmental sequences within and between concepts. *Monograph of the Society for Research in Child Development*, *46*(2, Serial No. 189).

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, *64*, 583–616.

Vermunt, J.K., & Magidson, J. (2007). *Latent GOLD 4.5 syntax module (computer program)*. Belmont, MA: Statistical Innovations.

Wilson, M. (1989). Saltus: a psychometric model of discontinuity in cognitive development. *Psychological Bulletin*, *105*(2), 276–289.

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

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

Wilson, M. (2012). Responding to a challenge that learning progressions pose to measurement practice: hypothesized links between dimensions of the outcome progression. In A.C. Alonzo & A.W. Gotwals (Eds.), *Learning progressions in science*. Rotterdam: Sense Publishers.

*Published Online Date: 26 FEB 2013*