1-1-2009

Measuring measuring: Toward a theory of proficiency with the Constructing Measures framework

Brent M. Duckor
San Jose State University, brent.duckor@sjsu.edu

K Draney

M Wilson

Recommended Citation

Brent M. Duckor, K Draney, and M Wilson. "Measuring measuring: Toward a theory of proficiency with the Constructing Measures framework" *Journal of Applied Measurement* (2009): 296-319.

JOURNAL OF APPLIED MEASUREMENT, *10*(3), 296-319

Measuring Measuring: Toward a Theory of Proficiency with the Constructing Measures Framework

Brent Duckor

Karen Draney

Mark Wilson

University of California, Berkeley

This paper is relevant to measurement educators who are interested in the variability of understanding and use of the four building blocks in the Constructing Measures framework (Wilson, 2005). It proposes a uni-dimensional structure for understanding Wilson's framework, and explores the evidence for and against this conceptualization. Constructed and fixed choice response items are utilized to collect responses from 72 participants who range in experience and expertise with constructing measures. The data was scored by two raters and was analyzed with the Rasch partial credit model using ConQuest (1998). Guided by the 1999 Testing Standards, analyses of validity and reliability evidence provide support for the construct theory and limited uses of the instrument pending item design modifications.

Requests for reprints should be sent to Brent Duckor, Dept. of Secondary Education, SH 436, College of Education, San Jose State University, One Washington Square, San Jose, CA 95192-0074, USA, e-mail: Brent.Duckor@sjsu.edu.

Our aim in this article is to share research on how expertise develops in classrooms and professional communities that use the Constructing Measures framework (Wilson, 2005). As measurement professionals and educators, we are particularly interested in identifying learning progressions in areas such as the development of construct theory, the conceptualization of the items design and scoring procedures, the application of measurement models, and the uses of validity and reliability evidence for making arguments to support (or challenge) an instrument's uses. This study is as much a philosophical endeavor as it is an experimental study. We believe the study of qualitative differences in individual thinking about the role of "building blocks" of measurement, assessment, and testing must build upon the work of philosophers and experts in measurement (National Research Council, 2001) and that we can hypothesize about how experts might differ from novices. Yet the study of the emergence and development of conceptions that advance the first principles of measurement, assessment, and testing in education and the social sciences must also embrace the learner as s/he strives to understand the field. In our judgment, one is not naturally born an expert in measurement who has access to a specialized intelligence. Nor do we believe that the student of measurement is an empty vessel whose training in the field can be reduced to the mastery of a series of repetitive technical procedures or the mere building up of an association of quantitative skills. Rather, we believe that the learner progresses in her/his understanding of the field through a learning-by-doing process that involves constructing (and replacing) more sophisticated concepts with less sophisticated ones over time. Unfortunately, in the quest to provide professional knowledge to other disciplines, we have taken for granted the development and structure of learning for the learner, and hence the cogntive development of expertise in and for our own field.

Having established the framework of what we see as fundamental variables along with professionals in the field of measurement **should** be developed, we lean to the empirical study. The results from the empirical study may help us to differentiate among more and less sophisticated ways of thinking about measurement. For over a century, experts in the field of measurement, assessment, and testing have concerned themselves with the construction and validation of instruments designed to measure human traits, aptitudes, and skills. Psychologists have used sophisticated measurement tools such as factor analysis to advance our understanding of variables such as intelligence. Education experts have employed novel assessment techniques and unique item designs to better understand the nature of variables such as children's mathematical ability or language skills. Researchers and practitioners have sought to measure a host of constructs related to human behavior and, to do so, they have employed their knowledge of measurement tools, procedures, and principles.

Despite substantial progress in measuring human variables in the domains of education and psychology, we know very little about the development of measurement expertise among individuals in the field of measurement itself (either in the university classroom or in professional settings). We have largely taken for granted the experts' use and application of the tools of educational measurement, assessment, and testing and thus ignored how that expertise develops in communities of practice (Wenger, 1998) that use these tools. Not surprisingly, our picture of both expert and novice parts of the learning trajectory in the field of psychological and educational measurement itself is largely anecdotal and impressionistic. We do, however, have pictures of the history of measurement expertise as reflected in the individual development of particular theorists (Thurstone, 1925; Stevens, 1946; Rasch, 1960; Wright, 1968).

Some experts have identified a broad range of misconceptions about measurement and testing that persist in public policy discourse (Braun and Mislevy, 2005; Popham, 2000, 2004), but these observations have not yet been the subject of rigorous empirical study. It is worth noting that, at this time, we do not possess, as a research community, a body of literature that engages us in the study of measurement "expertise," "proficiency" or related concepts. Nor do we have, as a professional body, a set of technically calibrated assessments of the major domains of measurement knowledge that might be used to warrant decisions about an individual's capacity to practice in the field. Recent federal legislation (NCLB, 2001) has spurred increased demand for more expertise in test development but measurement specialists have not yet met this with meaningful or consistent measures of individual proficiency in the field of measurement itself.

Toward a Theory of Measurement Expertise

This article examines a body of measurement, assessment, and testing knowledge that is concerned with the principled construction and design of instruments in the social sciences. It addresses the problem of defining what constitutes that knowledge in part by turning to the definitions, principles and, ultimately, theoretical framework proposed by a national committee of experts (NRC, 2001) charged with explaining what excellent contemporary practice in test and assessment design entails. While certain features of what we define as measurement knowledge will be familiar to those in the Rasch measurement field (Fischer and Molenaar, 1995), it is worth noting that we see knowledge of the principles of Rasch measurement (e.g., the focus on the underlying variable, background in specific objectivity and so forth) as being useful but not sufficient for understanding how best to construct measures in education and the social sciences.

The work of the NRC Committee (2001) suggests that experts use sophisticated concepts, procedures, and practices to construct test instruments and to validate the scores and interpretations derived from them. In particular, the NRC committee's findings suggest that a useful tool for understanding how experts use their knowledge would be the "assessment triangle." As shown in Figure 1, the corners of the triangle represent three key aspects that serve as a framework for thinking about the foundations of assessment and their interrelationships (NRC, 2001, p. 44).

The first vertex of the triangle, "Cognition," refers to the model of cognition or learning in a given domain under study. [Note that the more general term "construct" (AERA, APA, NCME,

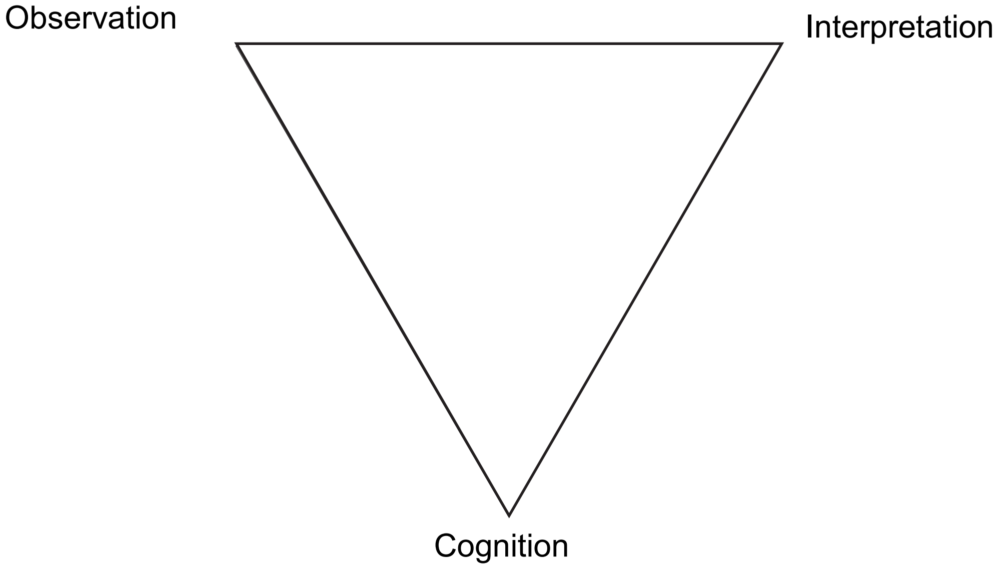


Figure 1. Assessment triangle (adapted from NRC, 2001).

1999) would be located at this vertex.[1]] This corner of the triangle articulates the explicit theory or hypothesis that describes the human variable being measured. Experts have many ways of depicting the cognition corner in an evidence-based framework for measuring, including domain representations (Khatri et al., 2006), facets (Minstrell, 2000), construct mapping (Wilson, 2005), predicted response patterns (Siegler, 1976, 1998), learning progressions (Catley, Lehrer, and Reiser, 2004) and other schematic representations.

The second vertex of the assessment triangle, "Observation," describes the set of prompts, tasks, and their contents that are expected to elicit demonstrations of the construct or latent trait under study. Measurement experts and other professionals often refer to the contents of this vertex generically as "the items." The tasks or items that human subjects are asked to respond to in an assessment, measurement, or testing situation are not arbitrarily chosen (NRC, 2001, pp. 47): items are designed and deliberately chosen to represent the cognitive model or construct under investigation. In Wilson's (2005) framework discussed below, the "items design" (observation corner) is linked to the construct map (cognition corner): it provides the content that leads to the validity evidence for the cognitive model of learning. Experts have various strategies for developing item sets, and they may have preferences for certain types of items (e.g., mapping sentences for facet theories of meaning, concept maps for schematic knowledge, or fixed-choice for declarative knowledge), depending on the content domain or theory of cognition.

The third vertex of the assessment triangle, "Interpretation," examines the data collected based on the observation corner, in particular, the ways in which the evidence relates to the construct(s) from the cognition corner. The NRC (2001) committee sees this last corner as encompassing "all the methods and tools used to reason from fallible observations" (p. 48) back to the cognition vertex. This vertex is often referred to as "score interpretation" (AERA, APA, NCME, 1999), which, in the case of educational or psychological testing, is constructed from numbers generated by quantitative models. Psychometric experts use statistical techniques to investigate the expectations or hypotheses developed about the constructs or traits under investigation. The experts' need to transfer from raw observations to codes often resides outside the reach of psychometrics, although the success of the psychometic expertise depends entirely on the success of this aspect of the work. They use psychometric approaches based on classical test theory (Spearman, 1904), item response theory (Rasch, 1960; Lord and Novick, 1968; Wright and Masters, 1982), and generalizability theory (Cronbach, Gleser, Nanda and Rajaratnam, 1972) as tools for examining the nature and structure of observations from items.

With the use of such schematic mental models, measurement developers can attend to the cognitive, observational and interpretive features of the instruments they design in order to draw more consistent and meaningful inferences about the behavior of human subjects. In this article we focus on how a diverse range of measurement professionals demonstrate proficiency with an approach to measurement knowledge and pratice that is sympathetic to the one outlined in the NRC committee report. The Constructing Measures (CM) framework outlined by Wilson (2005) is consistent with the NRC triangle and is based on a similar "building blocks" model of test design. The CM framework empasizes the use of evidence marshalled on behalf of an instrument to make meaningful and consistent statements about variables such as a person's skills or abilities, based on their responses to items. Working in the item response modeling tradition, the CM framework utilizes several schematic representations or mental models to advance individual understanding and practice of measurement of latent variables in education and the social sciences.

[1] We define construct broadly to include cognitive, attitudinal, and behavioral measures. In our interpretation of this element of the assessment triangle, the term "cognition" refers to the hypothetical structure of **any** latent variable, not just those variables concerned with cognition or learning outcomes for students in the field of education.

Background on Constructing Measures Framework and the Definition of the CM variable

There are a variety of frameworks available to the researcher who seeks to describe the structure of latent variables in a content or knowledge domain. In order to make the process of measurement consistent with the topic of measurement, we employed the construct modeling (Wilson, 2005) approach which uses Wilson's "building blocks" method. This approach allows us to develop and validate the scores derived from the instrument which is designed to elicit CM knowledge itself. We employed the SOLO (Structure of the Learning Outcome) taxonomy to conceptualize our intial ideas about the structure of learning in the CM topic areas. The SOLO taxonomy (Biggs and Collins, 1982) is a general theoretical framework that may be used to construct an outcome space for a task related to cognition and we express it in terms of a hierarchy of observable outcome categories. Moreover, we adopted this particular cognitivist approach to mapping learning progressions in the CM framework in order to make explicit our hypotheses about how knowledge of and the practice of constructing measures progresses for the learner. That is to say, we developed the construct maps for this research, in part, on the basis of how we as educators have observed individuals progressively master the CM framework in our own classrooms and professional settings.

In Wilson's (2005) "building blocks" framework, experts and novices are expected to draw upon four aspects of the measurement process to demonstrate proficiency with constructing measures, and then to apply them to establish evidence for reliability and validity of the instrument. These aspects are first, four interconnected aspects of measuring: (a) construct mapping; (b) the items design; (c) the outcome space; and (d) the measurement model. While we suspect that some of the proficiencies across the four aspects may be strongly related, we nonetheless sought to carefully distinguish between each of the topics in the constuct definition phase. Hence, a total of six constructs were developed to represent each of the six domains shown in Figure 2 (i.e., the four building blocks plus reliability and validity).

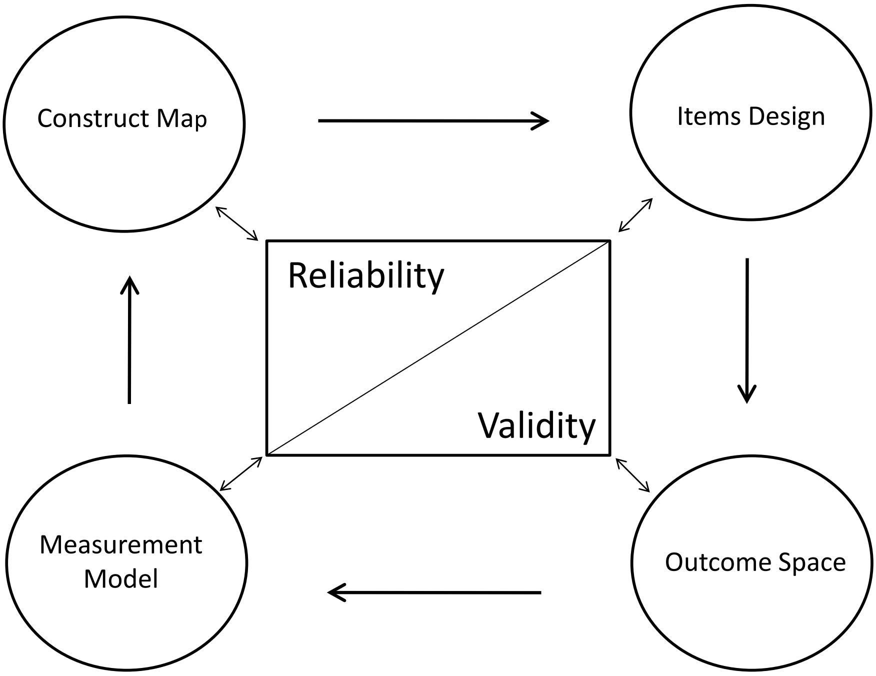


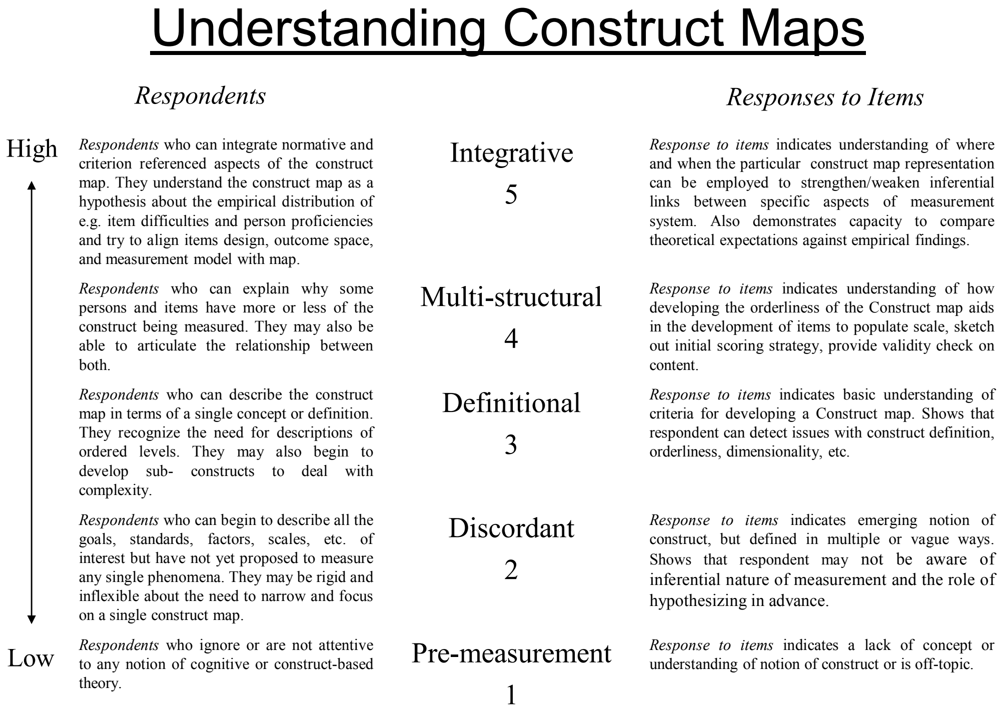
Figure 2. Relations among topics of CM knowledge.

In the first construct, there is the *Understanding Construct Maps* (UCM) variable, which focuses on the types and quality of the construct map representations the measurer uses to define a given variable. It covers the properties, affordances, and constraints of the construct mapping procedure as a tool for constructing measures by focusing on the distinct levels of the construct. For the second construct, there is the *Understanding the Items Design* (UID) variable. It focuses on the measurer's knowledge of traditional item formats and uses, but also encompasses the more sophsiticated notion of items as samples from a pool that may or may not bear a plausible relationship to the levels of a construct or cognitive theory under investigation. The third topic is the *Understanding the Outcome Space* (UOS) variable. It includes the measurer's knowledge and use of the properties of a good outcome space, which are drawn from prior research on progress variables and developmental assessment (Masters and Wilson, 1997). For the fourth topic, we have constructed the *Understanding the Wright Map* (UWM) variable, which focuses on the measurer's use of information derived from the output of a particular measurement model, in this case, a technically calibrated Wright map. We are interested in how individuals vary in the sophistication of their interpretations of these maps, and, in particular, how the Wright map is used to gather information about the structure and functioning of a hypothesized latent variable as represented by the construct map. The final two topics, both of which deal with the requirement of quality control for evidence-based interpretations derived from instruments such as surveys, tests, and assessments, focus on the measurer's knowledge of the types and uses of evidence for reliability and validity. We call these two variables *Understanding Quality Control-Evidence for Reliability* (UQC-RE) and *Understanding Quality Control-Evidence for Validity* (UQC-VE) respectively.

Drawing on previous research (Duckor, 2006), we carefully describe our hypotheses about the developmental trajectory or learning outcomes we expect for each construct. (See Appendix A for all 6 construct maps.) The reader will note that the structure of each of the six construct maps is similar: each construct map is characterized by variation with respect to both persons and items. Figure 3 provides an example of a construct map for the UCM variable.

As shown in Figure 3, at the lower end of the left column of our construct map, we posit the existence of novices ("pre-measurement" level) who are not yet aware of the basics for developing, analyzing and modifying different aspects of the building blocks, in this case, the construct map. Similarly, at the lower end of the right column, we expect novice responses to items to show evidence of the absence of the concept of the construct mapping, or vague and ill-defined notions about the properties of a construct map. These item responses demonstrate little or no use of schematic or strategic ways of thinking about measurement, such as those evidenced in the NRC (2001) report. These responses may even contain fundamental misconceptions (Braun and Mislevy, 2005) about the nature of measurement, assessment, and testing. Novice item responses typically show that they are not aware of the inferential nature of measurement in education and the social sciences, and they generally do not recognize the role of such basic fundamentals as hypothesizing in the development and validation of instruments.

At the upper end of the left column of our construct map, we hypothesize a group of experts (i.e., those at the "integrative" level) who can identify and use various mental models and schema for representing cognitive, observational and interpretive aspects of measurement. These individuals are able to flexibly use and adapt the building blocks while recognizing the potential affordances and constraints of a measurement situation. On the upper right column of our construct map, we hypothesize that these experts' responses to items will indicate that they know where and when the particular construct mapping procedure or representation can be employed. They strategically utilize different measurement frameworks, or special features of particular ones (e.g., phenomenography), to strengthen the infer-



Understanding Construct Maps

	Respondents		*Responses to Items*
High	*Respondents* who can integrate normative and criterion referenced aspects of the construct map. They understand the construct map as a hypothesis about the empirical distribution of e.g. item difficulties and person proficiencies and try to align items design, outcome space, and measurement model with map.	Integrative 5	*Response to items* indicates understanding of where and when the particular construct map representation can be employed to strengthen/weaken inferential links between specific aspects of measurement system. Also demonstrates capacity to compare theoretical expectations against empirical findings.
	Respondents who can explain why some persons and items have more or less of the construct being measured. They may also be able to articulate the relationship between both.	Multi-structural 4	*Response to items* indicates understanding of how developing the orderliness of the Construct map aids in the development of items to populate scale, sketch out initial scoring strategy, provide validity check on content.
	Respondents who can describe the construct map in terms of a single concept or definition. They recognize the need for descriptions of ordered levels. They may also begin to develop sub- constructs to deal with complexity.	Definitional 3	*Response to items* indicates basic understanding of criteria for developing a Construct map. Shows that respondent can detect issues with construct definition, orderliness, dimensionality, etc.
	Respondents who can begin to describe all the goals, standards, factors, scales, etc. of interest but have not yet proposed to measure any single phenomena. They may be rigid and inflexible about the need to narrow and focus on a single construct map.	Discordant 2	*Response to items* indicates emerging notion of construct, but defined in multiple or vague ways. Shows that respondent may not be aware of inferential nature of measurement and the role of hypothesizing in advance.
Low	*Respondents* who ignore or are not attentive to any notion of cognitive or construct-based theory.	Pre-measurement 1	*Response to items* indicates a lack of concept or understanding of notion of construct or is off-topic.

Figure 3. Construct map for UCM variable.

ential links between pieces of evidence in the process of instrument development and, ultimately, the validation of inferences drawn from scores. Item responses at this level typically demonstrate the experts' proficiency at comparing theoretical expectations about a construct against empirical findings related to multiple kinds of data.

While we are fairly confident that we have identified the extremes in our mapping of the UCM variable, we are intensely interested in the levels in between the "experts" and "novices," especially in our roles as measurement educators. We have observed that early encounters with the construct map often leave students of measurement in a "discordant" state (i.e., level 2 in Figure 3). These persons are able to describe multiple and complex goals, standards, factors, and so forth, that they wish to measure, but they have difficulty focusing on the definition of a single, latent variable. Their construct maps (i.e., responses to items at this level) are generally not well-defined, exhaustive or ordered in meaningful ways. Similarly, when asked to comment on others' construct maps, they tend to respond in superficial ways such as "add more levels" or miss fundamental flaws such as the presence of two or more variables in the construct definition.

The next level in the learning progression we call "definitional" (Figure 3) in part because it refers to individuals who are competent users of the construct mapping technique. Typically, these persons can describe their own construct maps and critique the construct maps of others in terms of basic requirements, such as the need for well defined qualitatively ordered categories that span the variable under investigation. In their own work, they are beginning to deal with variable complexity, in part, by breaking down or chunking phenomena into multiple or "sub-construct" maps. At this level of proficiency, we find they are typically able to identify the presence of potentially confounding aspects embedded in a variable definition.

In between the extremes of expert and novice, we have further hypothesized a "multi-structural" level of proficiency with construct mapping. At this level, persons are adept at explaining why some persons and items may have more or less of the construct. For instance, they typically have an implicit theory of person ability and item difficulty which guides the level of specificity they bring to category descriptions. Their responses to items (in particular when offering advice for improvement) shows that they understand how developing the orderliness of the construct map can aid in the development of items used to populate the scale, sketch out an initial scoring strategy, provide a potential validity check on instrument content, and so forth. We have noticed how these individuals have experienced and, hence, tend to value the importance of revising and reworking the construct map based on multiple iterations in the instrument development process.

Our aim in this study is to examine evidence for (or against) this theory of CM proficiency. To simplify our initial research in the Constructing Measurement framework, we have identified 5 levels of proficiency for each construct that we think are similar for each construct across all of the domains. (See Duckor, 2006, for a detailed discussion of the other five constructs.) Our primary interest in these developmental levels of proficiency is to better understand differences in performance so as to improve the learning outcomes for students and professionals who are interested in our field. The next section of this article addresses the methods and data sources employed to investigate our hypotheses.

Methods and Data Sources

Description of the respondents

A sample of 72 respondents was obtained from three pools of potential participants for this study (Duckor, 2006). The first pool consisted of students who had participated during the last several years in a graduate course, titled "Introduction to Measurement in Education and the Social Sciences," which is offered at the School of Education at the University of California at Berkeley, and is taught by the third author. The second sample pool consisted of individuals who participated in projects affiliated with the Center for Assessment and Evaluation of Student Learning, and which includes curriculum developers, teachers, university researchers, and K-12 science content specialists. The third sample pool consisted of individuals who participated in the 12th biennial International Objective Measurement Workshop; they include researchers, consultants, and university professors with a professional interest in Rasch item response modeling. Table 1 shows the demographic characteristics of participants.

E-mail communication was the primary method of recruitment for the sample. The main selection criterion apart from belonging to one of the pools of people likely to have at least some relevant knowledge, as described above, was a willingness to complete all of the 26 items on the CM instrument. Four exit interview items were also included with the instrument. Demographic data collected included variables related to educational background, measurement course and

Table 1

Selected Sample Demographic Characteristics

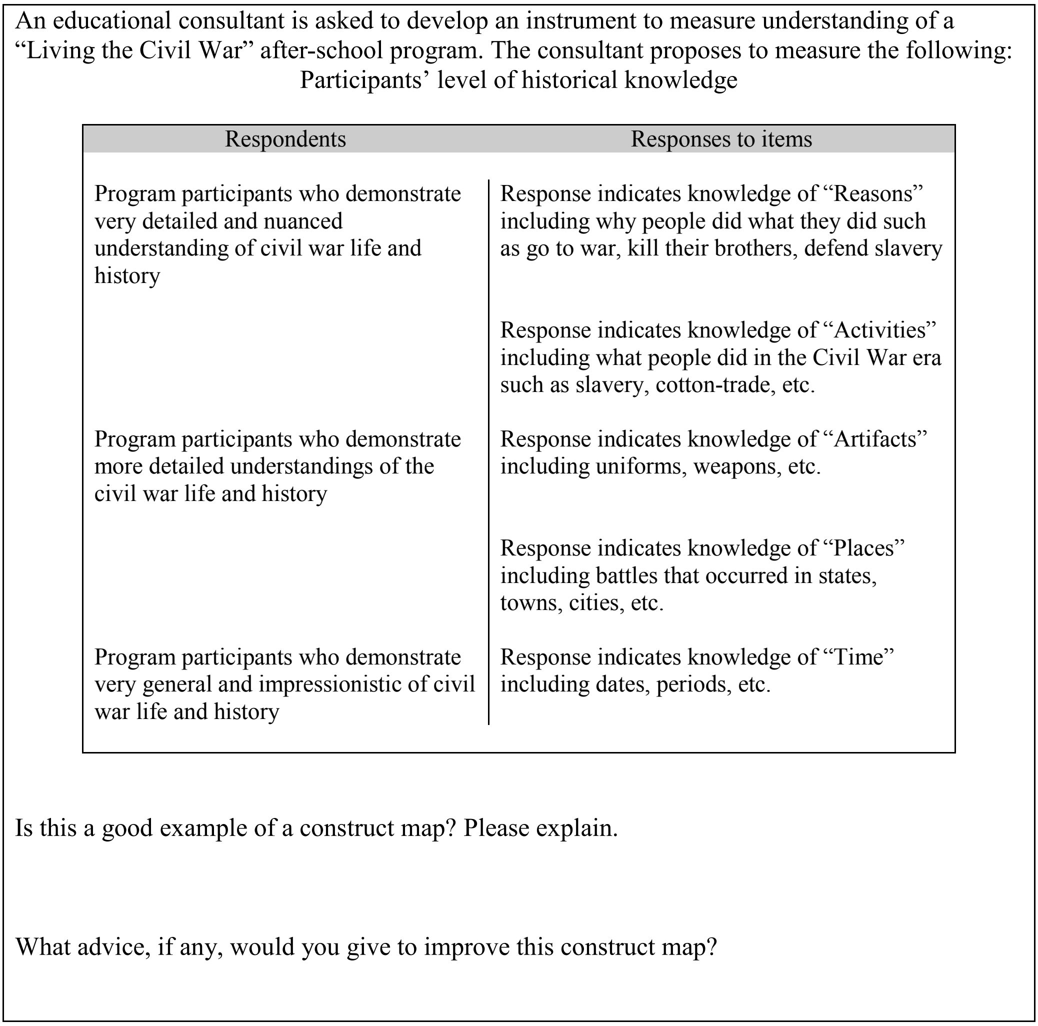
Characteristic	Frequency	Percentage
Female	42	58.3%
Under 40 years old	50	69.4%
Caucasian	46	63.9%
Graduate student	35	48.6%
Have obtained Masters degree	43	59.7%
Currently enrolled in PhD program	47	66.7%
PhD program in quantitative methods	19	40.4%

related experience, as well as teaching and professional experience with constructing measures.

Instrumentation

Items design. The *CM instrument* is a proficiency test designed to measure understanding and use of the CM framework. The test consists of 26 items: 8 fixed choice and 18 constructed response questions. Each item is targeted on a specific domain in the CM framework and is designed to span parts of a specific CM construct map. An example of an item from the UCM domain is shown in Figure 4.

This item is typical of the constructed response format used for this variable. It was designed to probe the understanding of directionality and ordering as they relate to the task of construct mapping. The item provides a written scenario, along with a representation of a construct map. There are two open-ended prompts, each requiring a short answer. The item tasks are intentionally designed to appear similar to the course-embedded assessments and research projects that use the CM framework. In both of these situations, the measurer engages in an item panel (Wilson, 2005) session that is used, among



An educational consultant is asked to develop an instrument to measure understanding of a "Living the Civil War" after-school program. The consultant proposes to measure the following:

Participants' level of historical knowledge

Respondents	Responses to items
Program participants who demonstrate very detailed and nuanced understanding of civil war life and history	Response indicates knowledge of "Reasons" including why people did what they did such as go to war, kill their brothers, defend slavery
	Response indicates knowledge of "Activities" including what people did in the Civil War era such as slavery, cotton-trade, etc.
Program participants who demonstrate more detailed understandings of the civil war life and history	Response indicates knowledge of "Artifacts" including uniforms, weapons, etc.
	Response indicates knowledge of "Places" including battles that occurred in states, towns, cities, etc.
Program participants who demonstrate very general and impressionistic of civil war life and history	Response indicates knowledge of "Time" including dates, periods, etc.

Is this a good example of a construct map? Please explain.

What advice, if any, would you give to improve this construct map?

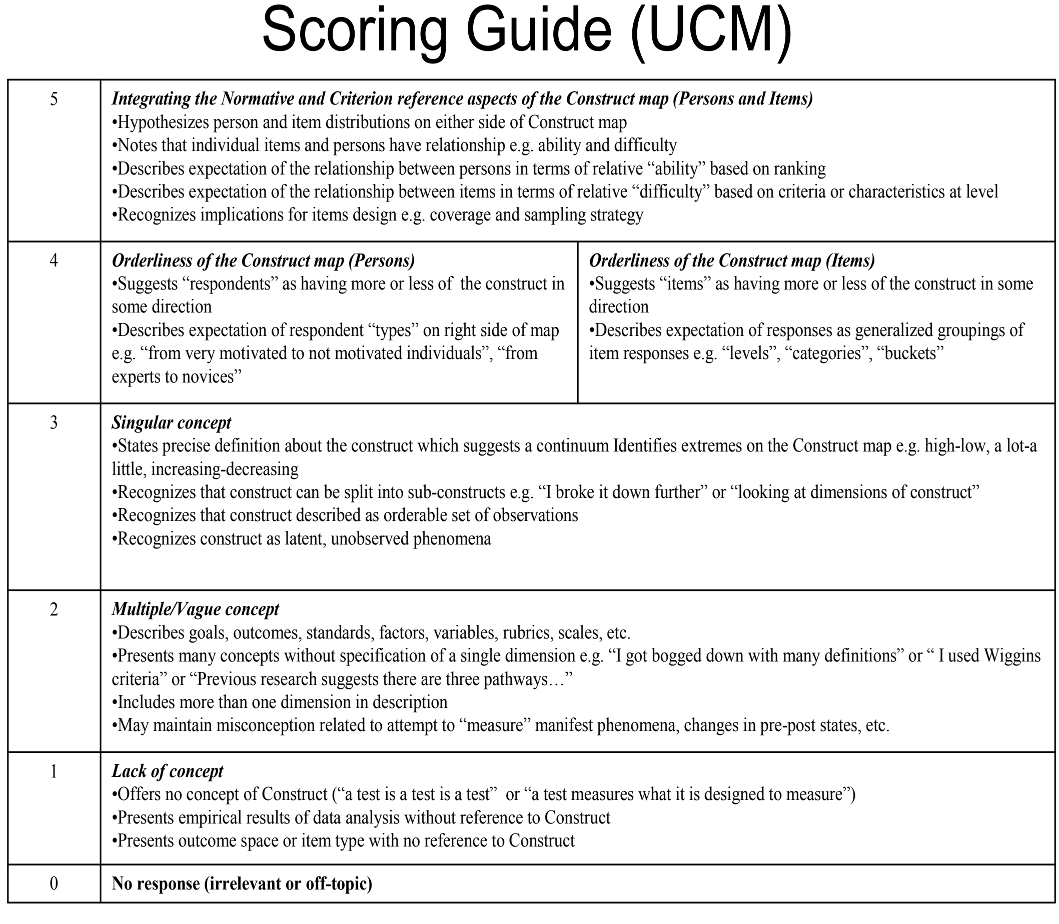
Figure 4. Item (UCM1) designed to probe the principle of order and directionality of variable.

other things, to examine the quality of a particular construct map. In these sessions, individuals are asked to give feedback, advice, and suggestions for improvement. With this short answer item format, respondents are allowed to provide a written explanation to support their responses.

We also used eight fixed-choice items which consisted of a stem and four answer choices. Respondents were instructed that some items may have more than one plausible option, and respondents were encouraged to select the one best option.

Scoring procedure. The constructed response items for the CM instrument were scored with the use of scoring guides directed at each of the CM domains (Duckor, 2006). A broader, more general scoring guide was first derived from the construct maps, which consists of a numeric performance level and a corresponding description of item response characteristics at that level. We refer to this generalized scoring guide as an "outcome space" shown in Figure 5 for the UCM variable.

As shown in Figure 5, this scoring guide is divided into partially ordered categories according to the way content experts and the course instructors see the underlying latent construct as a learning progression. That is, the theoretical expectation, based in part on research from the classroom learning experience, is that students tend to progress from less to more sophisticated understandings of the CM framework. This type of generalized scoring guide is designed to score student responses from both the CM instrument



Scoring Guide (UCM)

5	***Integrating the Normative and Criterion reference aspects of the Construct map (Persons and Items)*** •Hypothesizes person and item distributions on either side of Construct map •Notes that individual items and persons have relationship e.g. ability and difficulty •Describes expectation of the relationship between persons in terms of relative "ability" based on ranking •Describes expectation of the relationship between items in terms of relative "difficulty" based on criteria or characteristics at level •Recognizes implications for items design e.g. coverage and sampling strategy	
4	***Orderliness of the Construct map (Persons)*** •Suggests "respondents" as having more or less of the construct in some direction •Describes expectation of respondent "types" on right side of map e.g. "from very motivated to not motivated individuals", "from experts to novices"	***Orderliness of the Construct map (Items)*** •Suggests "items" as having more or less of the construct in some direction •Describes expectation of responses as generalized groupings of item responses e.g. "levels", "categories", "buckets"
3	***Singular concept*** •States precise definition about the construct which suggests a continuum Identifies extremes on the Construct map e.g. high-low, a lot-a little, increasing-decreasing •Recognizes that construct can be split into sub-constructs e.g. "I broke it down further" or "looking at dimensions of construct" •Recognizes that construct described as orderable set of observations •Recognizes construct as latent, unobserved phenomena	
2	***Multiple/Vague concept*** •Describes goals, outcomes, standards, factors, variables, rubrics, scales, etc. •Presents many concepts without specification of a single dimension e.g. "I got bogged down with many definitions" or " I used Wiggins criteria" or "Previous research suggests there are three pathways…" •Includes more than one dimension in description •May maintain misconception related to attempt to "measure" manifest phenomena, changes in pre-post states, etc.	
1	***Lack of concept*** •Offers no concept of Construct ("a test is a test is a test" or "a test measures what it is designed to measure") •Presents empirical results of data analysis without reference to Construct •Presents outcome space or item type with no reference to Construct	
0	**No response (irrelevant or off-topic)**	

Figure 5. General outcome space for UCM variable.

data and 274A course assignments, where improvement in understanding is expected over time.

In addition to the general outcome space for each CM variable, we also used item-specific scoring guides on the constructed response items. We refer to these as "exemplar documents," because they contain, for each level of the scoring guide (a) sample responses that can be assigned to that level and (b) brief annotations discussing how to interpret these responses. Responses are chosen both to exemplify clear cases that belong in each level, and difficult cases that need judgement and discussion. These exemplar documents have been designed to align with the generalized outcome space, so that the overall structure of the variable is preserved; that is, the categories used for describing levels of proficiency and item difficulty are consistent across scoring guides. These exemplar documents were primarily employed because they provide support for our rater training protocol, and they allow for more flexibility in measurement model specification. The fixed choice item responses for the variables related to understanding reliability and validity were scored polytomously.

Statistical procedures

Measurement model. The choice of any measurement model is always constrained by the affordances of data (e.g., sample size, item format, and dimensionality). In this study, we employed a partial credit Rasch model to calibrate items and measure persons (Wright and Masters, 1982). The parameters were estimated using ConQuest (Wu, Adams and Wilson, 1998). Item and step parameters were estimated with a Gauss-Hermite Quadrature with 15 nodes. The assumed population distribution was Gaussian. Statistical reports generated by ConQuest are used to describe estimates of these parameters and allow for the investigation of CM scale properties (including analyses of differential item functioning and general item analyses). We also employ standard analyses of item and person fit statistics to check model fit. In the next section, we report the results of validity and reliability studies conducted on the CM instrument.

Results

The results reported in this article are based on evidence for (and against) inferences about person proficiency based on global CM scale estimates. We found some evidence to support our hypothesis (resembling aspects of the SOLO Taxonomy) for a unidimensional structure of proficiency with the CM framework.[2] The 1999 "Testing Standards" (AERA, APA, NCME) for reliability and validity guide the two major lines of evidence reported, and the Standards are used to establish an argument for the CM instrument's potential interpretations and uses.

Four major pieces of validity evidence are presented here to support the meaningfulness of the scores derived from the CM instrument. First, our argument for *content validity* rests, first, on the development of the construct map that represents the intent to measure, and then, on the items that are designed to prompt student responses as well as the outcome space that is designed to value the responses according to the construct map (Wilson, 2005). The development of the UCM construct map is given as an example of this step, which occurred over a two-year period. First, interviews with content experts and course instructors were conducted, in which several hypotheses about the variable was advanced. This yielded a picture of possible learning progressions, which was later turned into an initial construct map. Then this map was revised after two item paneling sessions and a pilot testing phase. We then triangulated the item responses derived from the pilot CM instrument with an examination of student work from the course, which led to further improvements in the UCM construct map. Finally, we found additional supporting evidence of "the relationship between the test's content and the construct[s] it is intended to measure" (Standards, 1999, p. 11) from in-depth interviews conducted with novices and emerging intermediates over a semester dur-

[2] In his dissertation research, Duckor (2006) investigated the relationship among sub-scales to better define the association among variables and found evidence that these variables are moderate to strongly related. This study also identified several statistically significant predictors of CM proficiency, among which include relevant previous graduate course, research, and professional experience.

ing which the introductory measurement class was taught. In these semi-structured interviews, Duckor (2005) found that these students in the course experienced conceptual growth and changes in ways predicted by the construct (map) theory.

Secondly, we report on validity evidence based on *response processes* in order to establish evidence of "the fit between the construct and the detailed nature of performance or response engaged in by examinees" (Standards, 1999, p. 12). Nearly all of the 72 respondents completed the exit interview from the CM instrument. The overall findings from the exit interviews were positive. Sample responses are: "These are very good items, eliciting people's understanding of constructs, validity and reliability from a variety of aspects," "It is a clearly worded instrument," and "The items weren't confusing." Where there was confusion among respondents, it related to the fixed choice items. Comments focused mostly on the "poorly conceptualized alternatives" or distractors. (These same issues were also detected in our analysis of some of the Wright Maps where we found a lack of internal structure among fixed choice items to support the construct theory!) But, in general, the respondents reported that they understood the items on the CM instrument and that they were able to apply their knowledge of educational measurement to it. In addition, the respondents gave detailed feedback on the instrument, including suggestions for improvement. Most of these suggestions centered on the time required to complete the instrument. Finally, we were encouraged by the high response rate (over 80% in most cases) on the exit interview, which furthers our ability to draw substantive conclusions about examinee perceptions about the CM instrument and the construct it seeks to represent. All in all, the results from these exit interviews allow us to conclude that respondents were neither confused nor distracted by "noise" (e.g., reading load, language complexity and so forth) that might have adversely affected their ability to respond to the items in a construct-relevant manner.

Thirdly, we report on the validity evidence for the interpretation of CM scale scores based on the structure and functioning of the CM items as a whole. When applying an item response model to examine the validity evidence for the *internal structure* of the CM scale, it is important to report first on whether the items and person are well-fit by the model: In this case we focus on the results of the weighted mean square fit and *t* statistics. These model fit statistics are a necessary (but not sufficient) guide for evaluating the scaling evidence. As we shall see below, our analysis of both item and person fit statistics support the overall finding that the UCM instrument data fit the partial credit model well.

Generally, higher item fit values, indicating measurement model misfit, are most important to the internal structure validity argument, in part, because they signal that an item contributes less to the overall estimation of the latent variable or construct, whereas lower item fit values are less worrisome because they tend to indicate measurement complications, such as local dependence, rather than measurement fundamentals, such as measuring the wrong dimensions (Wilson, 2005). The results of the weighted mean square fit and t statistics support the overall finding that, at the item level, the CM instrument data fit the partial credit model well; in fact, only one item, UQCV14, appeared to misfit the model from a statistical standpoint.

The weighted mean square fit statistics for CM item step parameters indicate good overall fit using the interpretive framework (.75 < MNSQ < 1.33) developed by Wu, Adams, and Wilson (1998). Only two item steps (UQCV12 and UQCV14) appear to misfit the partial credit model. The weighted mean square value for both items is less than one (0.57), indicating that the observed variance is less than expected, which may be due to chance alone.

Investigation of respondent fit also plays a central role in evaluating the evidence for the internal structure of the CM instrument. ConQuest software was used to obtain statistics on person misfit. The results demonstrated that a total of six out of 72 (8%) respondents appear to have misfit the model. This is a bit more than expected (5%) but is quite close and well within acceptable levels. The results for the majority of CM instrument respondents indicate relatively good fit.

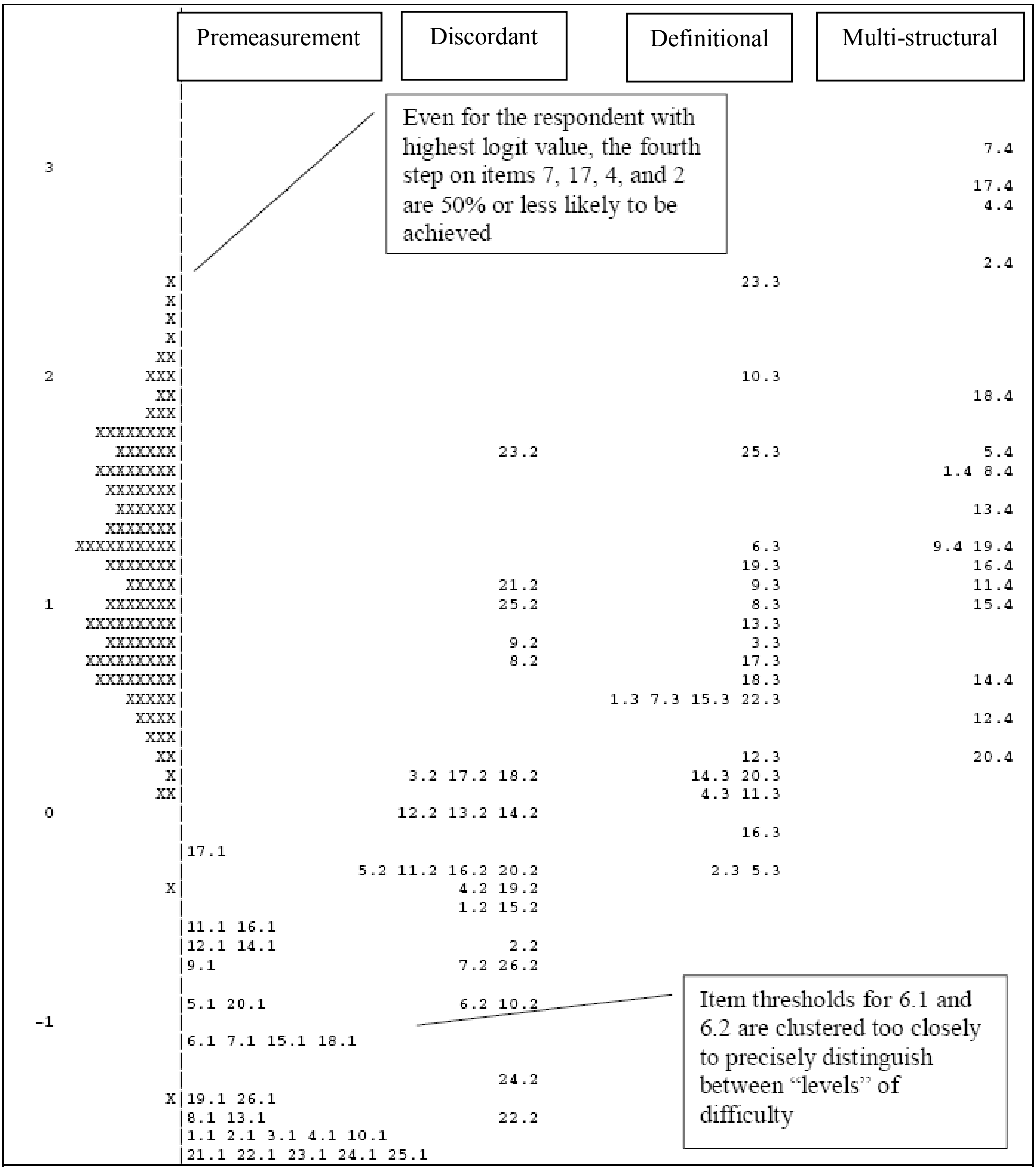
Usually, we worry first about the high misfitting respondents, rather than the low misfitting ones (Wright and Masters, 1982). Two of the most interesting misfitting respondents (a university professor and a curriculum developer) exhibited an infit mean square value of greater than 1.33, the upper boundary for what constitutes a good fit value (Adams and Khoo, 1996). Moreover, the weighted *t* values for both cases (–2.0 and –2.02) are statistically significant, that is, just outside the accepted range of –1.96 and 1.96. These results are not necessarily problematic, but do warrant

```
-----------------Level Responded-----------------Next Level----------------
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |OE7.4
                              OE4.4|   |
                                   |   |
                                   |   |OE2.4
                                   |   |
                                   |   |
                                   |   |OE18.4
                                   |   |
                             OE5.4 |   |OE1.4 FC5.2
                             OE8.4 |   |
                                   |   |
                        OE9.4 FC1.4|   |OE6.3
                              FC3.2|   |FC7.2
                              OE3.3|   |
                             OE18.3|   |
                  OE1.3 OE7.3 FC4.3|   |
                         ----------|---|
                                   |XXX|FC2.3
                                   |---|----------
                                   |   |OE17.1
                        OE2.3 FC2.2|   |OE11.1 OE16.1
                                   |   |
                                   |   |OE12.1 OE14.1 OE15.1
                              FC8.2|   |
                              OE6.2|   |
                                   |   |OE13.1
                                   |   |OE10.2
                                   |   |
                                   |   |
                              FC5.1|   |
                        FC6.2 FC7.1|   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                             OE10.1|   |
                                   |   |
---------------------------------------------------------------------------
```

Figure 6. Person fit map for high misfitting respondent examination of other validity evidence, including exit interviews from response processes.

The results for one of these misfitting persons is displayed on the person fit map shown in Figure 6. The left-hand side of this map shows which item responses were achieved by the respondent; the right-hand side shows which item responses were not achieved. The horizontal “XXX” represents the respondent’s estimated proficiency (0.54) on the CM scale. Each row represents 0.17 logits on the CM scale.

This respondent fit map exemplifies an unexpectedly poor fit (*infit ms* = 3.41; *t stat* = 4.179). This respondent provided responses that indicate the overall expected order predicted by the partial credit model was inappropriate for him or



Even for the respondent with highest logit value, the fourth step on items 7, 17, 4, and 2 are 50% or less likely to be achieved

Item thresholds for 6.1 and 6.2 are clustered too closely to precisely distinguish between “levels” of difficulty

Figure 7. Wright map of person proficiency and item thresholds for the CM scale

her. Upon closer examination of Figure 6, we notice that particular steps for items (UID4.4 and UOS5.4) were relatively easier for the respondent than was predicted by the partial credit measurement model. More importantly, it appears that the respondent found particular steps for all the items (e.g., UQCV12-17) from the understanding quality control domains relatively harder to reach than was predicted. Exit interview data on this group of items indicate that the respondent didn't "know the technical differences between these types of validities" and declined to hazard a guess. It is interesting to note that the other high misfitting respondent reported in the exit interview that while s/he considers her/himself a relative "expert" in the field of measurement: a "complete frustration at lack of meaningful context for the 'test'" may have interfered with her/his ability to give the best responses to the CM instrument. Further probing on this response is necessary, but one plausible interpretation is that the respondent felt that more relevant course work and professional experience with the CM framework may have improved her/his chances on performance on the instrument.

It is worth noting how the evidence in this case against the internal structure of the CM scale, as it pertains to a few respondents, could be triangulated with data collected from response processes. Thus, potential threats to the validity of score interpretations for particular individuals can be anticipated in the exit interviews and later weighed with other forms of quantitative evidence, such as those provided by Rasch item response modeling.

According to the recent *Standards* (APA, AERA, NCME, 1999), *internal structure* validity evidence refers to "the degree to which the relationships among test items and test components conform to the construct on which the proposed [instrument] score interpretations are based" (p. 13). We used a Wright map to examine the empirical ordering of persons and items in order to compare those to our theoretical expectations based on the CM construct. Figure 7 shows the distribution of respondent and item locations for the CM scale.

In comparing our construct theory to the empirical data analysis of the CM Wright map, we found moderate evidence for the banding of the item thresholds which would be consistent with the responses to items from the same levels having similar difficulties across most items. As shown in Figure 7, the *premeasurement* level of response to the items is represented by the first threshold, and spans the lower end of the scale (–.5 to –3 logits). The *discordant* and *definitional* response levels are represented by the second and third thresholds, which span the middle range of the scale (–3 to 2 logits). These levels represent transitional and emerging levels of proficiency in response to the CM items. Finally, we observe that the *multi-structural* level of response, which is represented by the fourth threshold, covers the upper end of the scale (2 to 5.5 logits). Thus, we can say that as the respondents improve in CM proficiency, they have a tendency to respond at a higher level to most of the items with the most sophisticated responses. On the whole, we observe that the relationship between the sides of the CM Wright map indicates that the respondents are covered across their entire range by the item thresholds; we did not detect either a ceiling or floor effect which suggests the CM instrument targets the respondents' proficiencies fairly well.

Nonetheless, since we are dealing with Thurstonian thresholds (which, by defnintion, cannot disorder), we must exercise caution in our interpretation of the evidence for internal structure. It does appear that the relative distances between these thresholds allow us to differentiate among respondents and may, in fact, warrant our interpretation of the qualitative difference between these "levels"—broadly speaking from pre-measurement to multi-structural levels. The item thresholds for UCM1.2, 1.3, and 1.4 are well-spaced across the scale but others, such as UID6.1 and 6.2, appear less well-spaced.

In fact, when we break out these item thresholds by "domain," a more complex picture of the construct theory to empirical data fit emerges—one that brings the issues of construct definition into sharper relief. While these items work well as a whole to define the unidimensional construct advanced in this study, it is clear that some item thresholds (and respective domains) lend themselves to our theory better than others. Here we provide two examples: one that suggests relatively good construct definition and another that does not. Figure 8 shows the distribution of respondent and item locations for the UCM sub-scale which represents the Understanding Construct Maps domain.

Although the data used to generate this particular sub-scale is comprised of responses from only three constructed response items, we found preliminary evidence for the four "levels" of the CM construct. With UCM1, we see a banding effect that follows our expectations of four qualitatively distinct levels of understanding construct maps. Similarly, we note the presence of bands or "levels" of item difficulty for UCM2 and UCM3, although the presence of measure-

```
WRIGHT MAP OF LATENT DISTRIBUTIONS AND THRESHOLDS (UCM SCALE)
======================================================================================================================
Generalised-Item Thresholds
                            |
             10             |
                            |
                            |
                            |
              9             |
                            |
                            |
                            |
              8             |
                            |
                           X|
              7             |
                            |
                            |
                        XXXX|
              6            X|
                          XX|
                         XXX|
                        XXXX|
              5           XX|
                         XXX|          UCM2.4
                         XXX|
                         XXX|
              4        XXXXX|
                        XXXX|
                       XXXXX|UCM1.4
              3       XXXXXX|
                  XXXXXXXXXX|
                  XXXXXXXXXX|
                      XXXXXX|
              2      XXXXXXX|
                       XXXXX|                   UCM3.3
                    XXXXXXXX|
                   XXXXXXXXX|UCM1.3
              1        XXXXX|
                      XXXXXX|                   UCM3.2
                          XX|
                        XXXX|
              0         XXXX|
                           X|          UCM2.3
                           X|UCM1.2
             -1            X|
                           X|          UCM2.2
                           X|
                           X|
             -2             |
                            |
                            |
                           X|          UCM2.1
             -3             |UCM1.1
                            |                   UCM3.1
                            |
                            |
             -4             |
                            |
                           X|
             -5             |
                            |
                            |
======================================================================================================================
Each 'X' represents   0.6 cases
The labels for thresholds show the levels of item, and step, respectively
Dummy case removed
======================================================================================================================
```

OPEN ENDED ITEMS

- UCM1: Historical knowledge
- UCM2: Motivation for college
- UCM3: Proficiency and confidence with math

Figure 8. Wright map of person proficiency and item threshold for the UCM scale.

ment error around the item threshold locations does not allow for definitive confirmation of the internal structure of the construct. Based on alternative scoring strategies, we also found that UCM3 appeared to mostly target the third "definitional" level of the CM construct map (Figure 3), which represents the difficulty in identifying and specifying a "singular concept" when measuring complex phenomena. While there is room for improvement in the calibration and evaluation the items design for this particular domain, we conclude that it is an example of relatively good validity evidence for the construct definition that corresponds with the general theory proposed by the CM construct map.

Figure 9 shows the distribution of respondent and item locations for the sub-scale UQC-RE, which represents the Understanding Quality

```
WRIGHT MAP OF LATENT DISTRIBUTIONS AND THRESHOLDS (UQC-RE SCALE)
========================================================================================================================
Generalised-Item Thresholds
------------------------------------------------------------------------------------------------------------------------
                            |
               6            |
                            |
                            |
                            |
                            |
                            |
                            |
               5            |
                            |
                            |
                           X|
                            |
                            |
                           X|
               4            |
                           X|
                          XX|
                          XX|                         RFC5.3
                          XX|
                          XX|
                         XXX|
               3        XXXX|ROE10.3
                        XXXX|
                        XXXX|                         RFC5.2
                   XXXXXXXXX|
                  XXXXXXXXXX|
                        XXXX|
                  XXXXXXXXXX|
                       XXXXX|
               2      XXXXXX|
                      XXXXXX|
                       XXXXX|
                    XXXXXXXX|
                    XXXXXXXX|        RFC4.3
                       XXXXX|
                       XXXXX|
               1          XX|
                        XXXX|
                         XXX|
                          XX|
                           X|
                          XX|
                           X|
               0           X|
                           X|                                             RFC8.2
                            |
                            |ROE10.2
                            |
                            |
                            |                                   RFC6.2
              -1            |                                             RFC8.1
                            |        RFC4.2           RFC5.1
                            |        RFC4.1
                            |
                            |                                   RFC6.1
                            |ROE10.1
                            |
                            |
              -2            |
========================================================================================================================
Each 'X' represents   0.6 cases
The labels for thresholds show the levels of item, and step, respectively
Dummy case removed
========================================================================================================================
```

FIXED CHOICE & OPEN ENDED ITEMS

- ROE10: "If a test has sufficient reliability, then it is a valid test"
- RFC4: The claim for an instrument's reliability is:
- RFC5: The most important part of the argument for the reliability of your instrument is:
- RFC6: The term 'reliability' usually implies evidence for an instrument's:
- RFC8: Measurement error can be the result of which of the following except:

Figure 9. Wright map of person proficiency and item threshold for the UQC-RE scale.

Control-Reliability Evidence domain. Here we draw a different conclusion based on the empirical results of the scaling procedure, as well as those from response processes evidence drawn from the exit interviews.[3] We do not find any compelling evidence for the expected internal structure of this sub-scale; there are no clear banding effects that correspond to our theory (spanning pre-measurement to integrative) understanding reliability evidence. Our scoring procedure, which presumes the existence of polytomous response categories, seems ill-suited to detect subtle variations in item difficulty. Instead, most of the fixed choice items (e.g., RFC4) function in a dichotomous fashion; more importantly, they do not cover or represent the construct in a way we can interpret. Our review of the available validity evidence based on both response processes data and IRT scaling suggests we need to go back to the drawing board for this domain.

The results from both a general item analysis and a differential item functioning analysis by gender using ConQuest add further weight to the internal structure aspect of our validity argument. There is neither evidence of gender DIF nor discordance of mean person locations in each of the item response categories (Duckor, 2006). Nonetheless, we note that our current items design does not provide an opportunity for respondents to demonstrate an *integrated* level of understanding on the CM instrument, which we address in the discussion section.

Our fourth and final source of validity evidence is based on the CM instrument's relations to *external variables*. Here we examined "the degree to which these relationships are consistent with the construct underlying the proposed [instrument] interpretations" (AERA, APA, NCME, 1999, p. 13). The principal source of evidence for the relationship between CM proficiencies and other variables is derived from the introduction to measurement course grades. In some cases, the instructors provided official grades; in others, the respondents provided self-reported grades. The results of the Pearson correlation coefficients based on both instructor and respondent self-report data allow us to conclude that the CM instrument and course grades assess a similar set of proficiencies: Despite our general concern with restriction of range in graduate course grade distributions, the scores for respondents on the CM instrument were positively and strongly correlated ($r = .89$) to the grades they received in the course (i.e., for the subset of respondents for whom we had course grades $N = 37$).

The results of the reliability analysis we conducted are presented in terms of evidence for reasonably small standard errors of measurement, in addition to indicators of internal consistency and rater reliability. We found that the reliability coefficient values for all of the CM scales were acceptable for research purposes. In fact, internal consistency indicators such as Cronbach's alpha (.89) and person separation reliability (.87) for the CM instrument as a whole were relatively high.

Given the nature of the CM instrument's design and its reliance on constructed response items, we also examined the inter-rater reliability evidence for the 18 non-objective items. The results indicated that the agreement was acceptably high: the Pearson correlation between the maximum likelihood estimates (MLE) values for both raters was very high ($r = .98$). Furthermore, we investigated the rater reliability using a multifaceted Rasch measurement model (Linacre, 1989) with ConQuest, which indicated that the model fit the data for the two raters well. Rater #1 scored approximately 0.036 logits higher than did rater #2 on the CM scale. This rater harshness parameter estimate (0.036) was just less than its standard error (0.038) and hence not statistically significant at the (.05) level.

[3] Duckor (2006) reported that more than 1 out of 5 respondents expressed concern about fixed choice items on the CM instrument. One respondent in the exit interview said: "The MC items were particularly snarly, since you've gone the route of 'mostly correct with one word sneakily changed to something else like substituting 'statement' for 'argument'. I hate that because I spend 10 minutes wondering if it's been done on purpose of if it's just a careless mistake." Other respondents stated their confusion about specific terminology: "I wasn't sure [what] the terms 'argument' and 'statement' meant" and "#5 of multiple choice—what are the coefficients referring to?" Another respondent reported difficulty in choosing an answer: "…the response options in the multiple choice section were too overlapping and similar."

Discussion

This article presents the results from empirically tested hypotheses about differences in knowledge of measurement in key content areas of the CM framework. It draws on a particular approach, the Constructing Measures framework, to better define what constitutes knowledge of the framework itself. Several constructs are identified in the specification of CM proficiency and items have been designed to represent them. A partial credit Rasch item response model was fit to the data generated by the mixed item format instrument to examine our theoretical expectations about the structure of CM proficiency. We found that there is evidence to warrant meaningful, consistent distinctions between levels of proficiency—ranging from pre-measurement to more integrated understanding of the CM framework—based on the CM scale locations.

Nonetheless, we are cautious in judging the results of the study. First and foremost, we hope to broaden the item formats currently available with the CM instrument. In particular, we see opportunities to gather more meaningful and consistent data from improved fixed choice items such as ordered multiple choice questions that target specific levels of proficiency. We also plan to expand the item bank to include more items for sub-constructs, such as Understanding the Outcome Space, that are not adequately covered by the current instrument. Secondly, there are limits to the study of the structure and function of the CM scale given our current measurement model specification. With a larger data set, we would like to fit a multidimensional item response model to examine our hypotheses about the nature of CM proficiency. Whether or not the use of such models will be feasible, we envision the need to study a more diverse population of graduate students, practitioners, and those who have not yet encountered formal training in this discipline. Thus, we plan to investigate how the constructs may vary across different populations and along different dimensions.

Our aim in this article has been to broaden the view of how expertise develops and how it can be differentiated from more novice-like ways of thinking about measurement. The study of differences in individual thinking about the role of "building blocks" or the assessment triangle in constructing measures in education and the social sciences is part of a long term research effort. There are no *a priori* reasons why we cannot approach the study of measurement with the same rigor and resources that we bring to bear on the study of other disciplines. In fact, if we are to progress in the communication and teaching of the principles of measurement among test developers, then we must engage in efforts to better understand the thinking and practices of those who do not yet know how to construct stable, meaningful measures, or more importantly, why they should.

References

Adams, R. J., and Khoo, S. T. (1996). Quest [Computer program]. Melbourne, Australia: ACER.

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

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

Braun, H. I., and Mislevy, R. (2005). Intuitive test theory. *Phi Delta Kappan*, *86*(7), 489-497.

Catley, K., Lehrer, R., and Reiser, B. (2004). *Tracing a prospective learning progression for developing understanding of evolution*. Washington, DC: National Academy Press.

Cronbach, L. J., Gleser, G. C. , Nanda, H., and Rajaratnam, N. (1972). *The dependability of behavioral measurement: Theory of generalizability for scores and profiles*. New York: Wiley.

Duckor, B. M. (2006). *Measuring measuring: An item response theory approach*. (Doctoral dissertation, University of California, Berkeley, 2006). 345 pp. Advisor: Wilson, Mark R. *UMI Dissertation Abstracts (ProQuest).*

Duckor, B. (2005, May). *Thinking about the act of measuring: The development of a theory of the construct*. Individual poster presented at the 2nd annual meeting of the Center for Assessment and Evaluation of Student Learning Conference, Santa Rosa, California. Available from Center for Assessment and Evaluation of Student Learning at http://www.caesl.org/conference2005/brent_sm.pdf

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York: Basic Books.

Fischer, G. H., and Molenaar, I. W. (Eds.) (1995). *Rasch models: Foundations, recent developments, and applications*. New York: Springer-Verlag.

Khatri, V., Vessey, I., Ramesh, V., Clay, P., and Park, S. (2006). Understanding conceptual schemas: Exploring the role of application and IS domain knowledge. *Information Systems Research*, *17*(1), 81.

Linacre, J. M. (1989). *Many-faceted Rasch measurement.* Chicago: MESA Press.

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

Masters, G. N., and Wilson, M. (1997). *Developmental assessment*. BEAR Research Report. Berkeley, CA: University of California, Berkeley.

Minstrell, J. (2000). Student thinking and related assessment: Creating a facet-based learning environment. In Committee on the Evaluation of National and State Assessments of Educational Progress. N. S. Raju, J. W. Pellegrino, M. W. Bertenthal, K. J. Mitchell and L. R. Jones (Eds.), *Grading the nation's report card: Research from the evaluation of NAEP* (pp. 44-73). Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Mislevy, R. J., Almond, R. G., and Lukas, J. F. (2003). *A brief introduction to evidence centered design*. CRESST Technical Paper Series. Los Angeles, CA: CRESST.

Popham, W. J. (2000). *Testing! Testing! What every parent should know about school tests*. Boston: Allyn and Bacon.

Popham, W. J. (2004). Why assessment illiteracy is professional suicide. *Educational Leadership, 62*(1), 82-83.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. J. W. Pellegrino, N. Chudowsky and R. Glaser, (Eds.). Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago: University of Chicago Press.)

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8,* 481-520.

Siegler, R. S. (1998) *Children's thinking* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72-101.

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

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16*, 433-451.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity.* New York: Cambridge University Press.

Wilson, M. R. (2005). *Constructing measures: An item response theory approach*. Mahwah, NJ: Lawrence Erlbaum.

Wright, B. D. (1968). Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing* (pp. 85-101). Princeton, NJ: Educational Testing Service.

Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wu, M. L., Adams, R. J., and Wilson, M. (1998). ACER ConQuest [computer program]. Hawthorn, Australia: ACER.

Appendix A
Construct Maps

Understanding Construct Maps

High	*Respondents* who can integrate normative and criterion referenced aspects of the construct map. They understand the construct map as a hypothesis about the empirical distribution of e.g. item difficulties and person proficiencies and try to align items design, outcome space, and measurement model with map.	Integrative 5	*Responses to items* indicate understanding of where and when the particular construct map representation can be employed to strengthen/weaken inferential links between specific aspects of measurement framework. Also demonstrates capacity to compare theoretical expectations against empirical findings.
	Respondents who can explain why some persons and items have more or less of the construct being measured. They may also be able to articulate the relationship between both.	Multi-structural 4	*Responses to items* indicate understanding of how developing the orderliness of the Construct map aids in the development of items to populate scale, sketch out initial scoring strategy, provide validity check on content.
	Respondents who can describe the construct map in terms of a single concept or definition. They recognize the need for descriptions of ordered levels. They may also begin to develop sub- constructs to deal with complexity.	Definitional 3	*Responses to items* indicate basic understanding of criteria for developing a Construct map. Shows that respondent can detect issues with construct definition, orderliness, dimensionality, etc.
	Respondents who can begin to describe all the goals, standards, factors, scales, etc. of interest but have not yet proposed to measure any single phenomena. They may be rigid and inflexible about the need to narrow and focus on a single construct map.	Discordant 2	*Responses to items* indicate emerging notion of construct, but defined in multiple or vague ways. Shows that respondent may not be aware of inferential nature of measurement and the role of hypothesizing in advance.
Low	*Respondents* who ignore or are not attentive to any notion of cognitive or construct-based theory.	Pre-measurement 1	*Responses to items* indicate a lack of concept or understanding of notion of construct or is off-topic.

Understanding Items Design

High	*Respondents* who can explain why items may not be the realization of a construct. Each items design had empirical properties with implications for construct theory. They are flexible and adaptive, knowing when and where to use a specific items design to yield appropriate information about construct.	Integrative 5	*Responses to items* indicate where and when the particular items design is likely to strengthen/weaken inferential links between specific aspects of measurement framework.
	Respondents who can explain why some items in sample are better than others and that mixing design specifications can yield better results. They may also invoke validity and reliability criteria.	Multi-structural 4	*Responses to items* explain why items design is appropriate and how choices generally relate to other aspects of instrument e.g. building blocks. Argues that items represent cognitive theory, domain or construct.
	Respondents who can describe generic qualities of items in terms of format, type, etc. The notion of any item representing a construct is not of primary concern to them.	Definitional 3	*Responses to items* indicate how items design e.g. levels of pre-specification has implications for observations and inferences. May note generic and conventional indicators of item quality.
	Respondents who do not distinguish between item and construct and often equate the two. They may suggest that specific types of items are necessary to measure type of construct e.g. essays only measure higher order thinking.	Discordant 2	*Responses to items* define which "items" e.g. Likert, multiple-choice, essay, etc. are used. No notion of choice of sample of items that can be designed to measure construct.
Low	*Respondents* who talk naively or rigidly about "questions", "essays", etc. without any notion of these instances as "items" that may or may not stimulate or structure opportunities for observation.	Pre-measurement 1	*Responses to items* does not plausibly relate to notion of "item" or is off-topic.

Appendix A—Construct Maps continued on next page.

Appendix A—Construct Maps continued from previous page.

Understanding Outcome Space

High	*Respondents* who can explain why a particular scoring scheme was applied to data. They are flexible and adaptive e.g. can imagine rescoring data with different models to check on the construct theory. For them, the choice of outcome space is tied to both the items design and the nature of the information sought about the construct.	Integrative 4	*Responses to items* indicate where and when the particular outcome space is likely to strengthen/weaken inferential links between specific aspects of measurement framework.
	Respondents who can explain why some scoring strategies are better than others and that mixing outcome space design specifications can yield better results. They may also invoke criteria for a good outcome space e.g. exhaustiveness, well-defined, ordered, research based.	Multi-structural 3	*Responses to items* explain why outcome space design is appropriate and how choices generally relate to other aspects of instrument e.g. building blocks. May also indicate how outcome space design has implications for what counts as response, what can be scored, what will be ordered as more-less, higher-lower, etc.
	Respondents who do not distinguish between outcome space and construct and often equate the two. They may suggest that specific types of scoring schemes are necessary to measure type of construct or domains e.g. rubrics are best for scoring data in English and subjects in the humanities.	Definitional 2	*Responses to items* define what or which "responses" look like in a outcome space design, e.g. Likert, Guttman, SOLO Taxonomy. Response begins to define what "response" looks like using conventional types without clear justification or rationale.
Low	*Respondents* who talk naively or rigidly about rubrics or stems/distractors without any notion of these instances as "outcomes" that have implications for construct representation and instrument design. They tend to reflexively favor or oppose a particular scoring approach e.g. use Likert scales or rubrics.	Pre-measurement 1	*Responses to items* does not plausibly relate to notion of "outcome space" or is off-topic.

Understanding Wright Maps

High	*Respondents* who can use data analysis to interpret theoretical expectations or hypotheses about construct. They can explain the relations between Wright Map (WM) and Construct Map. They use the WM as a tool to rethink the construct, items design and outcome space, even choice of measurement model.	Integrative 5	*Responses to items* indicate extended understanding of role of data analysis for revision of construct maps, items design, outcome space and evidence for quality control. Shows how the respondent gives advice that is conditional on construct theory. May also raise issues related to reliability and validity.
	Respondents who compare construct theory e.g. item difficulty ordering to estimations provided by Wight Map and use information to guide appropriate revision of construct theory, items design, even outcome space or scoring strategy.	Multi-structural 4	*Responses to items* indicate understanding of how to compare Wright Map with construct map. Respondents can explain need for revision based on specific analysis e.g. item or step locations, relative ordering of items, etc. Considers possibility of revising either construct map theory or rescoring data to reexamine Wright Map.
	Respondents who can evaluate the meaning of item and person locations on Wright Map. They may generate generic advice e.g. delete items that do not fit or target more items at specific level/band.	Definitional 3	*Responses to items* indicate recognition of the meaning of either person or item locations on Wright Map. Shows that respondents can generate general interpretation of scale.
	Respondents who can describe item and person distributions but lack any depth of analysis. They tend to focus on floor and ceiling effects or the type of distribution.	Discordant 2	*Responses to items* indicate respondent can identify distributions or basic patterns e.g. distributions on the item or person side of the Wright map. Includes comments about distributions e.g. skewed, bunching, gaps, etc. on Wright Map.
Low	*Respondents* who attach results from data analysis with no construct-based interpretation or who do not connect those results to other aspects of measurement framework. They may have no or very limited experience with Wright Map.	Pre-measurement 1	*Responses to items* indicate a lack of any notion of measurement model or is off-topic.

Appendix A—Construct Maps continued on next page.

Appendix A—Construct Maps continued from previous page.

Understanding Quality Control: Evidence for Validity

High	*Respondents* who can make an coherent argument for validity of an instrument based on appropriate pieces of evidence. They understand the affordances/constraints of different types of validity evidence. They can suggest specific strategies for increasing the quality of evidence based on the instrument's particular purpose or uses. They marshal evidence of validity and reliability together to support overall construct validity argument.	Integrative 5	*Responses to items* indicate extended understanding of role of validity evidence within the quality control domain or measurement framework. Allows for articulation of multiple criteria/standards by which to judge the quality of findings. Advice e.g. improve link between content and internal structure validity is appropriate to instrument's uses or purposes.
	Respondents who can explain why some validity evidence is more appropriate than others based on e.g. their instrument, item and outcome space design, and uses. Several forms of evidence are examined and marshaled in support of inferences. They may have a schematic approach to validity based in a measurement model e.g. IRT.	Multi-structural 4	*Responses to items* indicate understanding of validity argument that uses all of the available forms of evidence for validity in an integrated way. Allows for explanation of score as an inference to an underlying construct or trait. Provides explanation and evaluation of evidence presented, including its limitations.
	Respondents who can describe all procedural aspects of validity e.g. data collected, operations performed, types reported. Nonetheless, there is no clear, coherent integration of this with e.g. validity evidence. They are aware of some misconceptions and can state e.g. validation is a process, on-going, multi-dimensional, etc.	Definitional 3	*Responses to items* indicate understanding of validity argument with one or more of the forms of evidence for validity. Includes description of types of validity e.g. content, response processes, internal structure, relations to other variables, consequences.
Low	*Respondents* who can identify basic notions about validity in declarative terms. They know the basic definition of concepts and terms. They may not yet see the relationship between validity and validity.	Discordant 2	*Responses to items* indicate understanding of conventional validity indices without reference to argument or evidentiary base. Shows grasp of notions or definitions e.g. meaningfulness.
	Respondents who offer no notion or data analysis related to the concept of quality control.	Pre-measurement 1	*Responses to items* indicate naïve or subjective notion of validity or off-topic.

Understanding Quality Control: Evidence for Reliability

High	*Respondents* who can make an coherent argument for reliability of an instrument based on appropriate pieces of evidence. They understand the affordances/constraints of different reliability indicators and suggest specific strategies for increasing the quality of evidence based on the instrument's particular purpose or uses. They understand reliability evidence as inextricably related to their validity argument.	Integrative 5	*Responses to items* indicate extended understanding of role of reliability evidence within the quality control domain or measurement framework. Allows for articulation of multiple criteria/standards by which to judge the quality of findings. Advice e.g. increasing value of reliability indicator is appropriate to instrument's uses or purposes.
	Respondents who can explain why some reliability evidence is more appropriate than others based on e.g. their instrument, specific item design or choice of outcome space. They may have a schematic approach to reliability based in a measurement model e.g. IRT.	Multi-structural 4	*Responses to items* indicate understanding of reliability argument that uses all of the available forms of evidence for reliability in an integrated way. Provides explanation and evaluation of any evidence presented, including its limitations.
	Respondents who can describe all procedural aspects of reliability e.g. data collected, operations performed, coefficients calculated and reported. Nonetheless, there is no clear, coherent integration of this with e.g. validity evidence. They have begun to connect reliability and validity concepts.	Definitional 3	*Responses to items* indicate understanding of reliability argument with one or more of the forms of evidence for reliability. Includes description of types/indicators of reliability e.g. internal consistency, test-retest, inter-rater.
	Respondents who can identify basic notions about reliability in declarative terms. They know the basic definition of concepts and terms. They may not yet see the relationship between reliability and validity.	Discordant 2	*Responses to items* indicate understanding of conventional reliability indices without reference to argument or evidentiary base. Shows grasp of notions or definitions e.g. consistency, measurement error.
Low	*Respondents* who offer no notion or data analysis related to the concept of quality control.	Pre-measurement 1	*Responses to items* indicate naïve or subjective notion of reliability or off –topic.