

Measurement Models as Narrative Structures

CSE Report 680

Robert J. Mislevy
CRESST/University of Maryland

Chun-Wei Huang
WestEd

May, 2006

Correction in Authorship: Versions of this report prior to September 30, 2006 contained a cover page listing Robert J. Misley as author, with a full list of correct authors on a subsequent page. This error was corrected on September 29th. The complete list of contributing authors is: Robert J. Mislevy and Chun-Wei Huang.

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.6 Study Group Activity on Cognitive Validity
Robert J. Mislevy, Project Director, CRESST/University of Maryland

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

This report will appear as a chapter in *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications*, M. von Davier and C. H. Carstensen (Eds.), Springer.

MEASUREMENT MODELS AS NARRATIVE STRUCTURES

Robert J. Mislevy

CRESST/University of Maryland

Chun-Wei Huang

WestEd

Abstract

Advances in cognitive research increase the need for assessment that can address the processes and the strategies by which persons solve problems. Several psychometric models have been introduced to handle claims cast in information-processing terms, explicitly modeling performance in terms of theory-based predictions of performance. Cognitively based item response theory IRT models incorporate features of items that influence persons' responses, and relate these features to response probabilities. A number of these models are discussed in terms of the interplay between the formal characteristics of the probability models and the substantive narratives they support. Examples include the Linear Logistic Test Model, mixtures of IRT models, and multivariate structured IRT models.

1.0 Introduction

An active area in psychometric research has been developing models that address strategies by which examinees respond to tasks. One purpose of this presentation is to describe some of these models and the relationships among them. The other is to lay out a framework for discussing topics of this sort, foregrounding the interplay between the technical affordances of the probability-based psychometric models and the substantive arguments they are meant to support.

The framework consists of a narrative structure overlaid by a model that supports probability-based reasoning. The narrative component is a specialization of Toulmin's (1958) general argument structure to assessment arguments (Mislevy, 2003). The component that supports probability-based reasoning is the measurement model (Mislevy, 1994; Mislevy & Gitomer, 1996). It is through the narrative structure

connecting them to real-world phenomena that the formal variables and conditional distributions in measurement models acquire situated meanings. The key feature of the measurement models we discuss is that the narrative space has been extended beyond the measurement theme that characterizes trait/differential psychology, to include relationships among the ways people process information and features of tasks that interact with their problem solving—themes that have emerged from the information-processing research in cognitive psychology in the tradition of Newell and Simon’s (1972) classic *Human Problem Solving*.

Section 2 reviews Toulmin’s structure for arguments and its application to educational and psychological testing. Section 3 describes the extension to probability-based models and illustrates it with the basic Rasch model for dichotomous items (Rasch, 1960). Section 4 describes themes that are the basis of extensions from basic measurement models such as the Rasch model, including mixtures, differential item functioning (DIF), multiple groups of examinees, and covariates for tasks based on theories of problem-solving. Section 5 describes a number of these models and relates the structures of the probability models to the structures of the narratives they embody. Section 6 concludes with a comment about contrasting aims of psychometric modeling.

2.0 Assessment Arguments and Measurement Models

Toulmin (1958) proposed a schema for how we use substantive theories and accumulated experience to reason from particular data to particular claims. Figure 1 outlines the structure of a simple argument. The *claim* (C) is a proposition we wish to support with *data* (D). The arrow represents inference, which is justified by a *warrant* (W), a generalization that justifies the inference from the particular data to the particular claim. Theory and experience—both personal and formal, such as empirical studies and prior research findings—provide *backing* (B) for the warrant. In any particular case we reason back through the warrant, so we may need to qualify our conclusions because there may be *alternative explanations* (A) for the data, which may in turn be supported or uncut by *rebuttal data* (R).

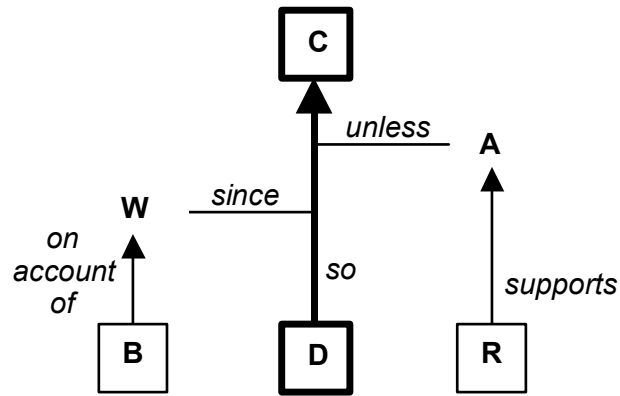


Figure 1. Toulmin's structure for arguments. Reasoning flows from data (D) to claim (C) by justification of a warrant (W), which in turn is supported by backing (B). The inference may need to be qualified by alternative explanations (A), which may have rebuttal evidence (R) to support them.

The foundation of an educational or psychological assessment argument is a concept of the nature of proficiency (e.g., knowledge, ability, propensity to act in certain ways in certain situations). It determines the nature of every element in the argument structure, and the rationale that orchestrates them as a coherent argument. As Messick (1994) asks: What kinds of things might one wish to say about persons? What kinds of things does one need to see an examinee say or do, in what kinds of situations? How are they related? The answers to these questions become claims, data, and warrants respectively in assessment arguments.

In particular, there are myriad aspects of persons, situations, and persons' actions within situations to which we might attend, and ways in which we might characterize them. The conception of proficiency shapes which of these will constitute data. An assessment argument generally includes three classes of data: aspects of the circumstances in which the person is acting, over which an assessment designer generally has principal influence; aspects of the person's behavior in the situations, over which the person has principal influence; and additional knowledge about the person's history or relationship to the observational situation as may be further required. These latter factors are essential to assessment in practice, even though they are often tacit, embedded in forms and practices.

To illustrate, consider a kind of task often used to assess spatial rotation (Cooper & Shepard, 1973). A subject is first shown a target figure, in this case a nonisosceles right triangle with a certain angle, then a second version of the target, rotated by a specified

number of degrees from the target (Figure 2). The subject must indicate whether the stimulus is identical to the target or a mirror image of it. Lower response latencies are usually taken as evidence of higher proficiency for tasks like these, but we will use correctness: More proficient subjects are posited to be more likely to make correct responses than less proficient subjects. The Toulmin diagram for an assessment argument based on Sue's correct response to an item of this type is shown in Figure 3. Data about the item are the triangles' acute angle, the rotation from the target, and whether the stimulus is the same or different. Data about the subject, such as gender and ethnicity, may not be available to the analyst but are not relevant in the basic measurement models illustrated here. In this example correctness will constitute the data about the performance. Reasoning back through the warrant, the claim is that Sue has a high level of proficiency in spatial rotation, based on the observation of this item response. We will address alternative explanations in the next section.

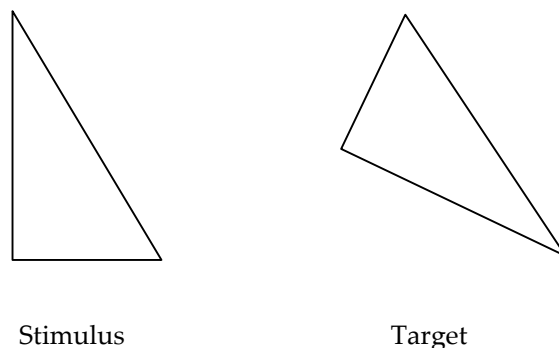


Figure 2. A spatial rotation item.

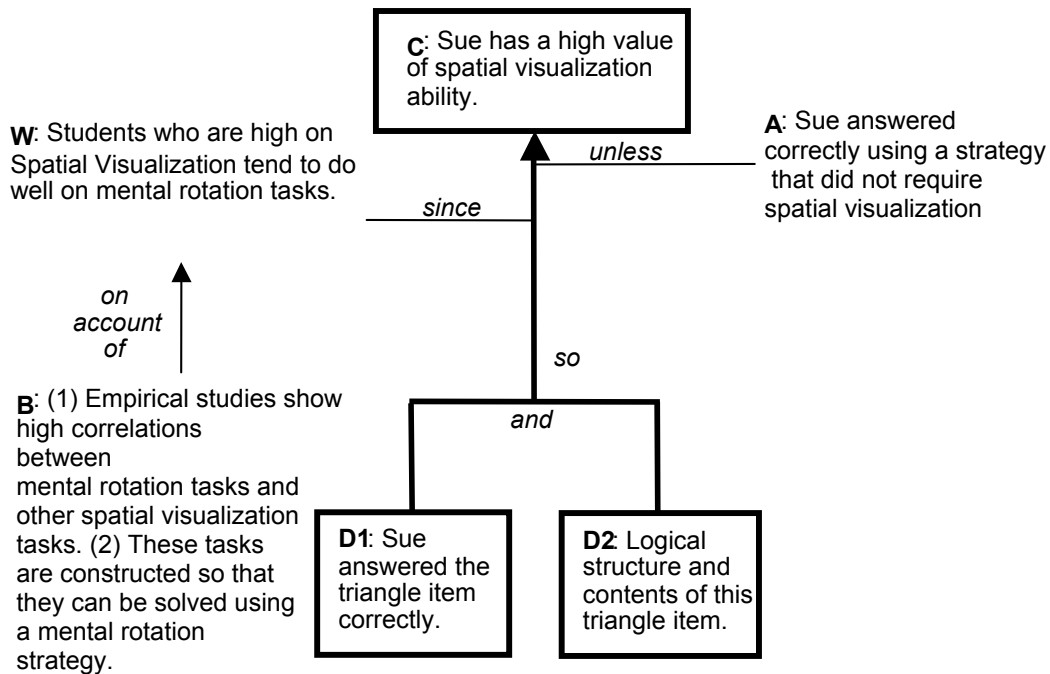


Figure 3. Toulmin diagram for one assessment task. Note that the warrant requires a conjunction of data about the nature of Sue's performance and the nature of the performance situation.

Of course a single item provides meager information, so spatial rotation tests generally consist of many items of the same kind. As such the same warrant applies in each case, as shown in Figure 4. Reasoning back through the same substantive warrant is called for, of course, and it is clear that more information is available. How is the evidence contained in the data across tasks to be synthesized, and how might we use this richer body of evidence to phrase more refined claims? This is where probability-based models come in.

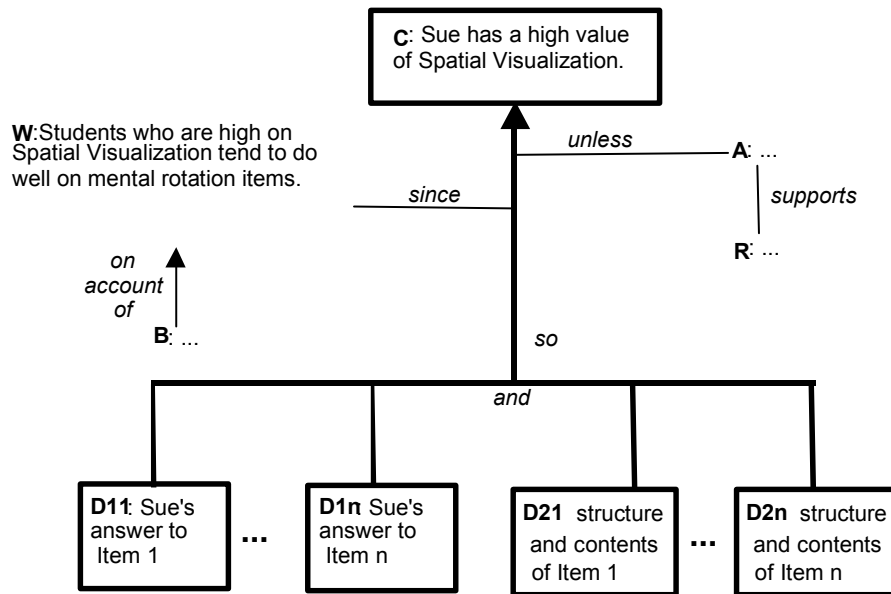


Figure 4: Toulmin diagram for several tasks of the same kind. The same general warrant is employed, as adapted to the particulars of each piece of data as they fit into the same scheme.

3.0 The Role of Probability-Models

3.1 From Arguments to Probability Models

Recognizing assessment as an evidentiary argument, one would like a mechanism to reason back up through the warrant, from data about what examinees say, do, or make, to claims about their knowledge and proficiencies more broadly conceived. Probability-based reasoning supports coherent reverse reasoning, specifically through Bayes' theorem. We construct a probability model that approximates key features of the situation in terms of variables and their interrelationships. Figure 5 shows the structure of an IRT model for the similar-tasks example described earlier. Details appear in the following section; we first address features illustrated here that are common to, and characteristic of, psychometric models more generally.

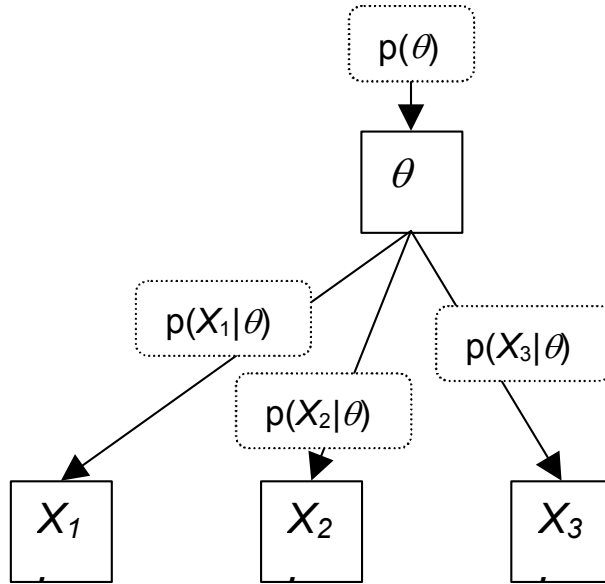


Figure 5. Graph for an item response theory (IRT) model.

There is an important difference between the variables in a probability model and the corresponding entities, claims and data, in a Toulmin diagram. A claim in a Toulmin diagram is a particular proposition that one seeks to support; a datum is a particular proposition about an aspect of an observation. A variable addresses not only the particular claim or observation, but also other claims or observations that could be entertained. As a datum in an argument, one might say that the response to Item j is correct. As a value of the item response variable X_j , we would say that the value of X_j is “correct” or 1 *as opposed to* “incorrect” or 0. If you know what the value of the variable is, you also know what it is not.

Whereas a claim in a Toulmin diagram is a particular proposition, a proficiency variable θ in a psychometric model characterizes ranges or potential values for selected aspects of proficiency. The possible values of the unobservable, possibly vector-valued, θ correspond to different states, levels, or configurations of proficiency. In Figure 1, the generic θ takes the particular form of a real-valued scalar θ that characterizes an examinee's propensity to make correct rather than incorrect responses. As formal entities, these variables can correspond to aspects of proficiency cast in trait, behavioral, information-processing, developmental, sociocultural, or any psychological perspective; that same perspective will drive the nature of observations and the relationships between them (Mislevy, 2003)—that is, the view of proficiency in the space of narratives a given probability model is constructed to support. A probability distribution over θ indicates knowledge at a given point in time about what the value of θ might be. The

prior probability distribution $p(\gamma)$ expresses what is known about a person's value of γ before responses or values of covariates are known.

A possibly vector-valued observable variable X characterizes selected aspects of a person's response. X 's are modeled as depending in probability on the person variables through conditional probabilities $p(x|\gamma)$. In this formulation the direction of reasoning flows, like the substantive warrant, in a deductive direction; that is, expectations for what observables might be if person variables were known.

The support for a substantive claim is expressed in terms of a probability distribution that represents degree-of-belief about corresponding values of γ . The situated meaning of such a claim arises from the nature of observations it is posited to affect and the substantive grounding of the model. Once such a model is fit and parameters have been estimated from initial data (pretest data, or 'calibration' data), Bayes' theorem can be used to update belief about person variables in light of task performances:

$$p(\gamma|x) = \frac{p(x|\gamma)p(\gamma)}{p(x)}.$$

The probability model becomes an additional component of a warrant that permits a quantitative expression of support for claims, and affords the calculus of probability to synthesize multiple, possibly conflicting, possibly overlapping, pieces of evidence. These advantages do not come for free. Additional backing is required for the probability-based aspects of the warrant, in the form of the pretest data. Additional alternative explanations for good or poor performance are introduced in connection with model misspecification and data errors.

Shafer (1976) defines a "frame of discernment" as all of the possible subsets of combinations of values that the variables in an inferential problem at a given point in time might take. The term "frame" emphasizes how a frame of discernment circumscribes the universe in which inference will take place. The term "discernment" emphasizes how a frame of discernment reflects purposive choices about what is important to recognize in the inferential situation, how to categorize observations, and from what perspective and at what level of detail variables should be defined.

Powerful methods are available for reasoning in probability models. For example, coherent updating of belief about any subset of variables, given new information about any other subset; clear expression of degree of support for claims expressed in terms of

values of ν 's or X 's; and the capability to express relationships of considerable subtlety and complexity (Schum, 1994), as might arise in simulation-based assessments tapping many aspects of knowledge and producing complex performances with sequential dependencies. These advantages obtain only for inferences that can be expressed in terms of a model's frame of discernment, however. The structure of the relationships embodied in a model may be quite flexible, but they effectively lay out the narrative space of stories that can be told, in terms of all the possible values that the variables might take. Questions concerning features of situations or patterns of interactions outside this universe of discourse cannot be asked, let alone answered. And to the extent that unmodeled patterns do exist in the real-world setting, they can distort inferences made through the model. We return to this issue in Section 5 in connection with multiple problem-solving strategies.

Model criticism tools do help deal with these problems. Tests of overall model fit are available. Even more useful are tests for particular suspected departures, such as adequacy across subsets of the data partitioned by features not in the model. Some patterns of observables—for example, a given person's pattern of responses—may be so improbable under the model as to cast doubt on using the model for that individual, even if the model fits well in general.

3.2 Example: The Rasch Model for Dichotomous Items

The Rasch IRT model for dichotomous items (RM: Rasch, 1960) posits that a probability of response to Item j given ν takes the following form:

$$P(X_j = 1 | \theta, b_j) = \Psi(\theta - b_j), \quad (1)$$

where $\Psi(\cdot) \equiv \exp(\cdot) / [1 + \exp(\cdot)]$ is the cumulative logistic probability distribution, ν is a one-dimensional measure of proficiency, b_j is a difficulty parameter for Item j , and x_j is 1 if right and 0 if wrong. Under the usual IRT assumption of conditional independence, the probability of a vector of responses to n items is

$$P(x_1, \dots, x_n | \theta, \beta_1, \dots, \beta_n) = \prod_{j=1}^n P(x_j | \theta, \beta_j) \quad (2)$$

The RM corresponds to a narrative space in which persons may differ as to their probability of answering items correctly, specified by ν , items may differ as to their probabilities of being answered correctly, specified by β_j , and the probability of the outcome when a person with proficiency ν attempts Item j is given by (1). Exactly the same difference in log odds (i.e., $\ln(p/(1-p))$) is obtained when we compute differences

between two given persons for any item across the collection for which the model is presumed to hold. The only differences among persons that can be expressed in the model is as to their overall propensity; all persons with the same evidence about their θ 's (in the case of a test, all persons with the same total score) are indistinguishable through the lens of the model.

These main effects patterns for comparing persons and similarly for comparing items render the RM a probabilistic version of a fundamental measurement model (Campbell, 1920), specifically, a conjoint measurement model as described by Luce and Tukey (1964) (see Fischer, 1968; Keats, 1971; Michell, 1997, 1999; Perline et al., 1979; and Roskam & Jansen, 1984). Although high θ persons sometimes answer low θ items incorrectly and vice versa, patterns in which some items are systematically easier for some people than others lie outside the narrative space of the RM. The narrative theme of fundamental measurement accords well with the trait or differential psychological perspective.

Model criticism tools such as item fit and person fit indices (e.g., Meijer & Sijtsma, 2001) allow the analyst to detect situations in which items do not appear to be equally difficult given overall proficiency for different groups of persons, or a person's response patterns is so unlike those of most people with similar overall proficiencies that the same substantive interpretation is not supported. In this way the item-level probabilistic framework grounds much stronger inference than the still-widespread practice of treating all examinees with the same total score as equivalent, without regard for systematic patterns within the data that would argue otherwise.

The basic RM does not encompass covariates q_j for items or w_i for persons. It is not an explanatory model, in the sense of De Boeck and Wilson (2004): Considerations of correlates of person proficiency and item difficulty, and thus substantive explanations of its character and probabilistic tests of conjectures to this effect, lie outside the model. Tests of whether θ 's are invariant across distinguishable groups of students and examinations of the relationships between item difficulties and item features are starting points for some of the extensions we discuss in Section 5. Such issues are there incorporated into measurement models, and the larger universe of inference that can be addressed in them supports a correspondingly larger narrative space.

4.0 Modeling Solution Processes

The “cognitive revolution” of the 1960s and 1970s, exemplified by Newell and Simon’s (1972) *Human Information Processing*, called attention to the nature of knowledge, and how people acquire and use it. How do people represent the information in a situation? What operations and strategies do they use to solve problems? What aspects of problems make them difficult, or call for various knowledge or processes? Strong parallels to computation and artificial intelligence appear in the use of rules, production systems, task decompositions, and means-ends analyses. The key insight is modeling problem-solving in these terms in light of the capabilities and the limitations of human thought and memory that are revealed by psychological experiments.

Among the tools developed to study cognitive processes is cognitive task analysis (CTA). CTA is a disciplined process of investigating the knowledge structures and strategies that individuals at targeted levels of ability use to solve specific types of tasks, through observable evidence of those structures and strategies. A CTA seeks to expose (a) essential features of task situations for eliciting certain behaviors, (b) internal representations of task situations, (c) the relationship between problem-solving behavior and internal representation, (d) processes used to solve problems, and (e) task characteristics that impact problem-solving processes and task difficulty (Newell & Simon, 1972).

In the 1970s researchers such as Carroll (1976) and Sternberg (1977) studied test items in these terms as psychological tasks. Others, including Whitely (1976) and Tatsuoka (Klein et al., 1981), designed aptitude and achievement test items around features motivated by theories of knowledge and performance in a given domain. For example, Mosenthal and Kirsch’s (1991) cognitive model for processing documents indicates that the difficulty of a task will be driven by (a) features of the document in question, such as the number of organizing categories; (b) features of the directive, such as the number of features that must be matched; and (c) the correspondence between the two, as determined by the degree to which the document has been designed to facilitate the inference that must be drawn. In the Third Edition of the influential volume *Educational Measurement* (Linn, 1989), Snow and Lohman (1989) asserted that

Summary test scores, and factors based on them, have often been thought of as “signs” indicating the presence of underlying, latent traits. ... An alternative interpretation of test scores as samples of cognitive processes and contents, and of correlations as indicating the

similarity or overlap of this sampling, is equally justifiable and could be theoretically more useful. The evidence from cognitive psychology suggests that test performances are comprised of complex assemblies of component information-processing actions that are adapted to task requirements during performance (p. 317).

Even when considering performances on familiar tasks, the cognitive perspective entails a new narrative space, in order to cast claims about persons, to characterize relevant features of tasks, and to express conjectures about relationships between task features and person performances. A new narrative space in turn entails a new frame of discernment, to bring into the probabilistic model those features and relationships that are central to the discourse of cognitive explanation but were irrelevant for the strict purpose of measuring traits.

We may distinguish five cases for modeling strategy use (extending a list given by Junker, 1999):

Case 0: No explicit modeling of strategies (basic IRT models)

Case 1: Common strategy presumed across persons.

Case 2: Strategy may differ between persons

Case 3: Strategy may differ between tasks, within persons

Case 4: Strategy may change within task, within persons

The Rasch model discussed in Section 3.2 is an example of Case 0. Models below include Case 1 (the linear logistic test model or LLTM), Case 2 (mixtures of Rasch models, mixtures of LLTMs), and Case 3 (the Andersen/Rasch model).

5.0 A Space of Models

This section describes a number of cognitively motivated extensions of IRT. Enough research has been done along these lines that a comprehensive review is beyond the scope of this presentation. We confine attention to tasks with single right/wrong responses and to extensions of the RM. We will illustrate three notable extensions of the narrative space: Story lines that reflect aspects of how persons solve tasks, how features of tasks influence their difficulty under a given approach, and that an observer may or may not know about the approach a person is taking for a given task. Interest lies in how these narrative themes are incorporated into parameters and structures of conditional probability distributions. The models described below are ordered approximately in terms of increasing complexity. A strict linear order does not

exist, but cases in which one model can be viewed as an extension of models discussed previously will be noted.

5.1 Differential Item Functioning (DIF)

As noted in Section 3.2, it is a common practice in educational and psychological testing to sum over item responses and treat all examinees with the same total score as exchangeable with respect to whatever the test is purported to “measure.” It is a matter of some importance that similar scores for students of different demographic groups based on, say, gender and race/ethnicity reflect similar performances on the items that comprise a test. Also as noted in Section 3.2, probability-based IRT models such as the Rasch model make it possible to test whether the patterns in a given data set support this interpretation (Thissen, Steinberg, & Wainer, 1993). Differential item functioning (DIF) means that typical performance on certain items varies substantially across groups among students with similar overall proficiency. That is, the difficulties of items vary across known groups of students.

A model that incorporates DIF with respect to manifest groups of students thus incorporates an observed student covariate w into the probability model:

$$P(X_j = 1 | \theta, w, b_j) = \Psi(\theta - b_{wj}), \quad (3)$$

where the item difficulty now depends on group membership, as indicated by the group index w on group-specific item parameters b_{wj} . It may be the case that only certain items exhibit DIF across groups. Nevertheless, substantive interpretations of examinees’ performances, and by extension their proficiencies, are incomplete without taking their group membership into account. Equation 3 affords no substantive explanation for these differences. It is compelling to examine the items that differentiate the groups. Does background knowledge differ among different groups of people? Are different people using different strategies to solve items? Conjectures about patterns suggested by substantive knowledge about items can be incorporated by using the approach discussed in Section 5.3.

It may be found that the RM fits well within the classes determined by partitioning persons and responses on the basis of w . In these circumstances one again obtains measurement models in the sense of probabilistic versions of conjoint measurement.

5.2 Mixtures of Rasch Models

The not uncommon finding of DIF among manifest groups raises the possibility that this phenomenon may be occurring even when the analyst does not happen to

know persons' values on the appropriate grouping variable. Mixture distribution Rasch models (e.g., Kelderman & Macready, 1990; Rost, 1990) incorporate an *unobserved* student covariate ν into the probability model:

$$P(X_j = 1 | \theta, \phi, b_j) = \Psi(\theta - b_{\phi_j}), \quad (4)$$

where the interpretation is the same as in the DIF model (3) except that now which group a given student belongs to is not known with certainty. Equation 4 can be described in terms of latent trait models within a latent class model. Given a student's pattern of observed responses and estimates of the group-specific item parameters, one uses Bayes' theorem to compute the posterior probability that the student belongs to each of the possible latent groups. A mixture IRT model obtains when patterns of relative difficulty for certain items appear to differ in a consistent manner in subsets of a dataset. As with DIF models, it is compelling to examine the items that differentiate the groups discovered in an application of a mixture model.

Although the probabilistic version of conjoint measurement would hold within groups, and this narrative theme could be used in discussing results, the mixture model of (4) is not itself a measurement model in this strict sense. Whether such models ought to be called measurement models is an open question.

Glück et al. (2002) provide an interesting example of a mixture Rasch model to study the effects of strategy training for spatial rotation tasks. Pre-test and post-test subsets of data are distinguished, and within time points proportions of students using a true rotational strategy and a less effective pattern matching strategy are estimated. The efficacy of each strategy as applied to three kinds of tasks could be predicted, so that when unrestricted mixture Rasch models were fit it was possible to identify resulting classes with strategies. They found that almost all of the students who used the pattern strategy at the pretest had switched to a spatial strategy at the posttest, after receiving training to that effect.

5.3 The LLTM

In the Linear Logistic Test Model (LLTM; Fischer, 1973; Scheiblechner, 1972), cognitively-based features of items and persons' probabilities of response are related through a so-called Q -matrix (Tatsuoka, 1983): q_{jk} indicates the degree to which feature k applies to item j . In simple cases, q_{jk} is 1 if feature k is present in item j and 0 if not. The LLTM extends the Rasch model by positing a linear structure for the ν_j s:

$$\beta_j = \sum_k q_{jk} \eta_k = q_j' \eta, \quad (5)$$

where η_k is a contribution to item difficulty entailed by feature k . Features can refer to a requirement for applying a particular skill, using a particular piece of information, carrying out a procedure, or some surface feature of an item—exactly the kinds of elements that Newell and Simon (1972) sought to uncover in cognitive task analysis as correlates of task difficulty.

The LLTM supports probability-based reasoning for a narrative space that addresses conjectures about the reasons that items are difficult and the nature of proficiency. In particular, any given value of β_j can now be interpreted in terms of expected performance in situations described by their theoretically relevant features. The LLTM is a measurement model in the sense of probabilistic conjoint measurement, so it also supports the narrative theme of fundamental measurement in the comparison of persons.

As an example, Fischer (1973) used the LLTM to model the difficulty of multi-step calculus items, as a function of how many times each of seven differentiation formulas had to be applied. He used statistical tests to determine how well the smaller set of features accounted for empirical patterns of difficulty, and whether repeated applications of a rule contributed additional increments to difficulty (they didn't). A relaxed version of (4), the random-weights LLTM (RW-LLTM; Rimjin & De Boeck, 2002), allows items with the same features to differ in difficulty, presumably due to non-modeled item features. Sheehan and Mislevy (1990) fit the RW-LLTM to a data set using features based on the Mosenthal and Kirsch (1991) cognitive analysis of document literacy tasks described in Section 4.

One can argue that models such as the LLTM marked a realization of Cronbach's (1957) call for the synthesis of the "two disciplines" of psychology, the experimental and the correlational, for they bring substantive theory, task design, and measurement modeling into a unified framework (Embretson, 1985a, 1998). When items are generated in accordance with theory and patterns among responses that can be predicted and tested against that theory, every item provides a new test of the theory (Bejar, 2002). Note that the reach of the basic LLTM extends only to final responses, not the identities or the sequences of processes that persons may carry out during the course of solution. Behavior at this more detailed level is central to cognitive task analysis, and inferences at this level are required in many intelligent tutoring systems to provide feedback or select instruction (e.g., Martin & VanLehn, 1995). This so-called model tracing lies below

the level that can be addressed in the narrative space supported by the LLTM, but in favorable cases the patterns LLTM can address will appear as emergent phenomena in overall performance. Steps in the direction of model tracing are seen for example in Embretson’s (1985b) model for multi-step problems: Each step produces a result to be modeled in terms of Rasch model or LLTM-like structures, and the final product is a stochastic outcome of step-level outcomes.

5.4 Multiple-Group LLTM

By combining elements of the DIF model (3) and the LLTM (5), one obtains a model that supports narratives about how items with different features are differentially difficult to members of different manifest groups:

$$P(X_j = 1 | \theta, w, b_j) = \Psi(\theta - b_{wj}),$$

where

$$\beta_{wj} = \sum_k q_{wjk} \eta_{wk} = q_j' \eta_w. \tag{6}$$

Equation 6 shows that the both codings of item covariates, q_{wjk} , and contributions to item difficulty, η_{wk} , can differ across groups. That is, both which features of items are relevant and how they are relevant can differ, presumably in accordance with a substantive theory that underlies the intended narrative space. Spada and McGaw (1985), for example, define groups in terms of educational treatments and pre- and post-test occasions, and item features in terms of curricular elements.

5.5 Multivariate Structured Models

Providing theoretically derived multidimensional characterizations of persons’ knowledge and skills is called cognitive diagnosis (Nichols, Chipman, & Brennan, 1995). Three features of cognitive diagnostic models are of interest here: First, the model space, hence the supported narrative space, supports qualified claims about persons, i.e., claims that have the word “but” in them; for example, “Steven is familiar with the strategy of space-splitting, but he is not sufficiently familiar with the canopy system to apply it there.” Rather than claims about overall proficiency in a domain, a multidimensional model accounts for persons’ performance in terms of profiles of knowledge and skill. Second, the tasks used to provide observations can be complex in that each depends on one or more of these dimensions of proficiency, and different tasks may impose different profiles of demand on them. Third, as with task design under the LLTM, substantive theory guides the construction of both the tasks

themselves and the probability model for analyzing the ensuing performances. The narrative depends on a conception of how persons with different levels or configurations of proficiency are likely to act in different ways in settings with different cognitively relevant features.

Most cognitively-based multidimensional IRT models posit compensatory or conjunctive combinations of proficiencies to determine response probabilities. The reader interested in conjunctive models is referred to Junker and Sijtsma (2001). In compensatory models, proficiencies combine so that a lack in one proficiency can be made up with an excess in another proficiency; that is, $a_j' \theta = a_{j1} \theta_1 + \dots + a_{jD} \theta_D$, where θ is a D -dimensional vector. The a_{jd} s indicate the extent to which proficiency d is required to succeed on item j . The A -matrix indicating examinee proficiency requirements is analogous to a Q -matrix specification of task features. A is estimated in some models (Ackerman, 1994), but in applications more strongly grounded in cognitive theory they are treated as known, their values depending on the knowledge and skill requirements that have been designed into each item. As an example, Adams, Wilson, and Wang's (1997) multidimensional random coefficients multinomial logit model (MRCMLM) is a multivariate generalization of the LLTM. Under the MRCMLM, the probability of a correct response is modeled as

$$\Pr(X_j = 1 | \theta, \eta, a_j, q_j) = \Psi(a_j' \theta + q_j' \eta). \quad (7)$$

De Boeck and his colleagues (e.g., Hoskens & De Boeck, 1995, Janssen & De Boeck, 1997, Rijmen & De Boeck, 2002) have carried out an active program of research using such models to investigate hypotheses about the psychological processes underlying item performance. In the extension of the MRCMLM to polytomous responses, each response category has its own a and q vectors to indicate which aspects of proficiency are evidenced in that response and which features of the category context contribute to its occurrence. Different aspects of proficiency may be involved in different responses to a given item, and different item features can be associated with different combinations of proficiencies.

Structured multivariate models such as the MRCMLM provide the means to solve a thorny problem in task-based language assessment (Mislevy, Steinberg, & Almond, 2002). Real-world language use draws simultaneously on several aspects of competence: from phonemic and morphological, through lexical and syntactic, to pragmatic, substantive, and cultural (Bachman, 1990). Research has shed light on factors that increase challenge in each aspect, such as the complexity of sentence structures and

the familiarity of the content. The challenge a task presents to a given person depends on both of what Robinson (2001) calls “complexity factors” and “difficulty factors.” The former are associated with features such as syntactic complexity and time pressure, which through information-processing arguments are expected to increase the cognitive load for most people. The latter he associates with features by which a task is rendered differentially hard or easy for particular persons, such as familiarity with content and prior experience with the genre.

Tasks differ from one to another with regard to the mix of demands they offer and the degree to which these demands interact with persons. How might one make sense of such complex data? The frame of discernment of multivariate structured models can support claims of the desired structure: Within the probability model, one can (a) characterize task demands with respect to complexity factors via a Q -matrix, (b) define a multivariate Ψ in terms of aspects of proficiency along which persons may differ with respect to planned variations in tasks tapping Robinson’s “difficulty factors,” and (c) indicate through an A -matrix which dimensions of Ψ are involved to what degrees for each observable variable. A model so constructed embodies generalizations about how persons with different profiles of language proficiencies are likely to act in situations with specified features—again, the direction of reasoning that accords with a warrant. From patterns of performance across tasks with different profiles of demand, the analyst applies Bayes’ theorem to infer a person’s profile of proficiency. We have argued elsewhere that the key to applying such models is defining from the beginning a joint narrative space for desired inferences, a design space for tasks to support claims so framed, and a structured probability model for responses that accords with the narrative and design spaces (Mislevy et al., 2002).

5.6 Structured Mixture Models

Cognitive task analyses reveal that different subjects may apply different strategies to the same problems (Simon, 1975), including familiar item types from educational and psychological testing (Kyllonen, Lohman, & Snow, 1984). Further, comparisons of experts’ and novices’ problem solving suggest that the sophistication with which one chooses and monitors strategy use develops as expertise grows. Strategy use is therefore a potential target for inference in assessment. This section considers Case 2 models for solution strategies, or structured mixture models.

The narrative themes embodied in mixed strategy models (e.g., Wilson, 1989, Mislevy & Verhelst, 1990) are these: Different persons may use different strategies but

are presumed to use the same strategy for all items. It is not known which strategy a person is using. Features of tasks that render them difficult are posited for each strategy.

Structured mixture models incorporate multiple Q -matrices to differentiate the strategies that may be used to solve the test items. Consider the case of M strategies; each person applies one of them to all items, and item difficulty under strategy m depends on features of the task that are relevant under this strategy in accordance with an LLTM structure. Specifically, the difficulty of item j under strategy m is $b_{jm} = \sum_k q_{jmk} \eta_{mk}$. Define for each person the vector parameter $\boldsymbol{\nu}^i = (\nu_{i1}, \dots, \nu_{iM})$, where $\nu_{im} = 1$ if person i uses strategy m and 0 if not, and denote the proficiency of person i under strategy m as θ_{im} . The probability of a correct response under such a model takes the form

$$\Pr(X_{ij} = 1 | \theta_i, \phi_i, q_j, \boldsymbol{\eta}) = \prod_m \left[\Psi \left(\theta_{im} - \sum_k q_{jmk} \eta_{mk} \right) \right]^{\nu_{im}} \quad (8)$$

As a first example, Wilson's (1989) Saltus model addresses developing proficiency that occurs in Piagetian stages. Balance beam tasks are a familiar example (Siegler, 1981). Movement to a new stage involves the acquisition of a new rule, so that certain classes of tasks become relatively easier. The Saltus model posits that a Rasch model holds across all items for persons at a given stage, but that these models may differ from one stage to another by shift parameters that depend on stage membership and its effect on items in each item class. Figure 6 illustrates a Saltus model for three stages, and Table 1 shows one way of parameterizing the Q -matrices and $\boldsymbol{\nu}$'s. (The illustrated model is a special case of the Saltus model in which each group of tasks becomes easier by a given amount once an student reaches a particular stage, and the shifts for that task group are zero before that stage and constrained to equality across groups thereafter. In an unrestricted Saltus model, the shift parameters for a group of tasks may vary from stage to stage.)

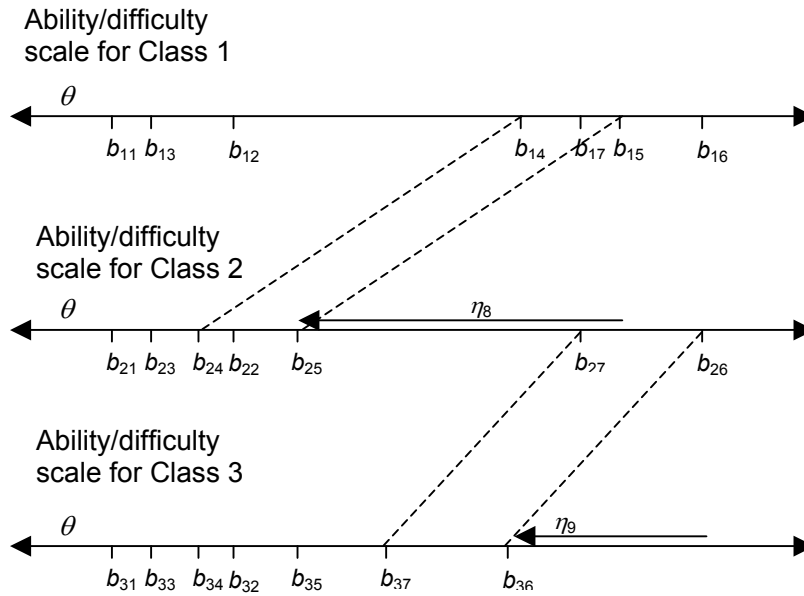


Figure 6. Three Saltus Rasch models.

As a second example, consider the finding that subjects may solve putative mental rotation items such as the one shown as Figure 2 either by the anticipated rotation strategy or by a feature-based analytic strategies (Hochberg & Gellman, 1977). Mislevy et al. (1991) modeled response times in this situation in terms of a mixture of the two strategies. The analytic strategy checks which direction, clockwise or counterclockwise from the right angle, one finds the sharper angle in the stimulus and target triangles. Difficulty is posited to increase linearly with degree of rotation under the rotational strategy (Shepard & Meltzer, 1971), but depend mainly on the acuteness of the angle under the analytic strategy. The task in Figure 2 would thus be relatively difficult under the rotational strategy but easy under the analytic strategy.

Table 1

Saltus Q-matrices and θ 's for an example with seven items and three stages.

Q-matrix for Class 1	Item	θ_{11}	θ_{12}	θ_{13}	θ_{14}	θ_{15}	θ_{16}	θ_{17}	θ_{18}	θ_{19}
	1	1	0	0	0	0	0	0	0	0
	2	0	1	0	0	0	0	0	0	0
	3	0	0	1	0	0	0	0	0	0
	4	0	0	0	1	0	0	0	0	0
	5	0	0	0	0	1	0	0	0	0
	6	0	0	0	0	0	1	0	0	0
	7	0	0	0	0	0	0	1	0	0

Q-matrix for Class 2	Item	θ_{21}	θ_{22}	θ_{23}	θ_{24}	θ_{25}	θ_{26}	θ_{27}	θ_{28}	θ_{29}
	1	1	0	0	0	0	0	0	0	0
	2	0	1	0	0	0	0	0	0	0
	3	0	0	1	0	0	0	0	0	0
	4	0	0	0	1	0	0	0	1	0
	5	0	0	0	0	1	0	0	1	0
	6	0	0	0	0	0	1	0	0	0
	7	0	0	0	0	0	0	1	0	0

Q-matrix for Class 3	Item	θ_{31}	θ_{32}	θ_{33}	θ_{34}	θ_{35}	θ_{36}	θ_{37}	θ_{38}	θ_{39}
	1	1	0	0	0	0	0	0	0	0
	2	0	1	0	0	0	0	0	0	0
	3	0	0	1	0	0	0	0	0	0
	4	0	0	0	1	0	0	0	1	0
	5	0	0	0	0	1	0	0	1	0
	6	0	0	0	0	0	1	0	0	1
	7	0	0	0	0	0	0	1	0	1

θ_{11} : Difficulty parameter for Item 1 in Class 1

θ_{12} : Difficulty parameter for Item 2 in Class 1

θ_{13} : Difficulty parameter for Item 3 in Class 1

θ_{14} : Difficulty parameter for Item 4 in Class 1

θ_{15} : Difficulty parameter for Item 5 in Class 1

θ_{16} : Difficulty parameter for Item 6 in Class 1

θ_{17} : Difficulty parameter for Item 6 in Class 1

θ_{18} : Shift for Items 4 and 5 for examinees in Classes 2 and 3

θ_{19} : Shift for Items 6 and 7 for examinees in Class 3

As a third and final example, cognitive analysis by Tatsuoka and her colleagues (Klein et al., 1981) found that the 530 middle-school students she studied characteristically solved mixed number subtraction problems using one of two strategies:

Method A: Convert mixed numbers to improper fractions, subtract, then reduce if necessary.

Method B: Separate mixed numbers into whole number and fractional parts, subtract as two sub-problems, borrowing one from minuend whole number if necessary, then reduce if necessary.

Tatsuoka (1983) further detailed the sub-processes required for solution under each method, and identified the attributes of items that called for the use of sub-processes under each strategy. An item like $7\frac{2}{3} - 5\frac{1}{3}$ is hard under Method A but easy under Method B; an item like $2\frac{1}{3} - 1\frac{2}{3}$ is the opposite. A response vector with most of the first kind of items right and the second kind wrong shifts belief toward Method B. The opposite pattern shifts belief toward the use of Method A. Note that these response patterns constitute noise, in the form of conflicting evidence, in an overall proficiency model, but constitute evidence about strategy usage under the mixture model. The narrative space of how students might be solving problems differently, and how an observer might see patterns that suggest which strategy is being used, supports a conjecture that cannot be framed within the overall proficiency model.

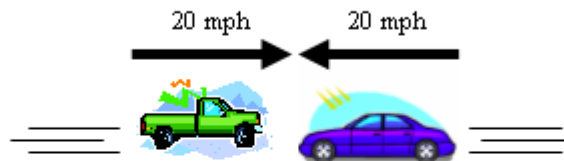
5.7 A Model for With-Person Mixtures of Strategy Use

The final model we discuss concerns a Case 3 instance of modeling strategy use. The narrative themes are these: A known fixed set of strategies exists to solve tasks in a given domain. A person may be using any of the strategies to solve a given task, although persons differ in their propensities to use different strategies, and tasks differ, by virtue of their features, in eliciting the use of different strategies. It is observed which strategy a person uses to solve each task. The inference of interest is, for each person, their propensities to use each of the strategies.

Examples of domain of tasks in which this narrative space applies can be found in science, where strategies correspond to conceptions and misconceptions in the domain that have been revealed by cognitive research (McCloskey, 1983). Researchers have

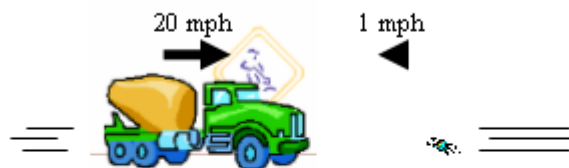
developed assessments in which tasks present situations, and multiple-choice options for predictions or explanations correspond to particular misconceptions. The Hestenes, Wells, and Swackhamer (1992) Force Concept Inventory (FCI) is an example. Figure 7 gives two examples of the kind of items found on the FCI, both based on Newton's Third Law, "For every action, there is an equal and opposite reaction." The first tends to evoke the Newtonian response because it is a paradigmatic Third-Law situation. The second is equivalent to an expert, but tends to evoke the response based on a common misconception, namely that the truck exerts more force than the fly because it has a greater mass.

What are the forces at the instant of impact?



- A. The truck exerts the same amount of force on the car as the car exerts on the truck.**
- B. The car exerts more force on the truck than the truck exerts on the car.**
- C. The truck exerts more force on the car than the car exerts on the truck.**
- D. There's no force because they both stop.**

What are the forces at the instant of impact?



- A. The truck exerts the same amount of force on the fly as the fly exerts on the truck.**
- B. The fly exerts more force on the truck than the truck exerts on the fly.**
- C. The truck exerts more force on the fly than the fly exerts on the truck.**
- D. There's no force because they both stop.**

Figure 7. Two items testing misconceptions about Newton's Third Law.

Huang (2003) used a Rasch model studied by Andersen (1973) to analyze responses to the FCI, for which responses could all be classified into three approaches to

force and motion problems: Newtonian, impetus theory, and nonscientific response. The response of Examinee i to Item j is coded as 1, 2, or 3, for the approach used. Each examinee is characterized by three parameters θ_{ik} indicating propensities to use each approach, and each item is characterized by three parameters indicating propensities to evoke each approach. Strategy choice is modeled as

$$\Pr(X_{ij} = k | \theta_i, \beta_j) = \frac{\exp(\theta_{ik} - \beta_{jk})}{\sum_{m=1}^3 \exp(\theta_{im} - \beta_{jm})}. \quad (9)$$

We may note that this model presumes that strategy use can be ascertained as an observable for each task, and that the categorization of strategies is exhaustive. Note also that the model addresses strategy approach only, not proficiency in using a given strategy. Were proficiency within strategy also a target of inference, then data concerning strategy application, such as correctness, would additionally be required. It would be modeled jointly with strategy choice, through a model such as the Rasch model where the proficiency for a given strategy applied only to tasks where the student was observed to have used that strategy. Further, if task features related to difficulty under the various strategies were available, then the within-strategy proficiency models could feature LLTM-like structures for task difficulties, or MRCMLM-like structures if the models were multivariate.

6.0 Closing Comment

Before the advent of item response theory, total scores on putatively similar tasks were taken a fortiori to be operationally defined measures—of what, and in what sense of measurement, to be determined partly through the thinking that led to the construction of the items and partly through correlations of these scores with other scores. Both issues lie outside the scope of the classical test theory model generally used to model uncertainty associated with persons' scores. Michell (2000) argues that this practice constitutes an abrogation of responsibility on the part of those who wish to contend that test scores are measures of quantitative psychological traits, where “measurement” is meant in Campbell’s (1920) classical sense of the term.

Developments in psychometric models and in cognitive research have moved the debate forward in ways both anticipated and unanticipated. A case can be made that the family of Rasch models does embody the axioms of fundamental measurement in a falsifiable probabilistic framework. Hence the claim of a quantitative measured trait can

be put to the test in any given data set. The question is becoming not so much whether scores reflect fundamentally measured attributes as whether the measurement narrative is sufficiently well approximated to ground applied work in particular applications. As research provides insight into the nature of human capabilities, extensions of psychometric models bring hypothesized data patterns into the probabilistic models where they too can be put to the test. This establishes methodological tools to address the questions of whether psychological attributes are quantitative, and to explore their nature with experiments and statistical tests. But the same research reveals that much that is important in the acquisition and use of knowledge is better expressed in terms other than common measured attributes. Now that tools have at last been developed to address the fundamental questions of trait psychology, a future may lie in using those tools for inferences that lie beyond its narrative space.

References

- Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7*, 255-278.
- Adams, R., Wilson, M.R., & Wang, W.C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.
- Andersen, E.B. (1973). *Conditional inference and models for measuring*. Copenhagen: Danish Institute for Mental Health.
- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bejar, I.I. (2002). Generative testing: From conception to implementation. In S.H. Irvine & P.C. Kyllonen (Eds.), *Item generation for test development* (pp. 199-217). Hillsdale, NJ: Erlbaum.
- Campbell N.R. (1920). *Physics: the elements*. Cambridge: Cambridge University Press
- Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new structure of intellect. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 27-57). Hillsdale, NJ: Erlbaum.
- Cooper, L. A., & Shepard, R. N. (1973). Chronometric studies of the rotation of mental images. In W. G. Chase (Ed.), *Visual Information Processing* (pp. 75-176). New York: Academic Press.
- Cronbach, L.J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671-684.
- De Boeck, P., & Wilson, M. (Eds.) (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.
- Embretson, S. E. (1985a). A general latent trait model for response processes. *Psychometrika, 49*, 175-186.
- Embretson, S.E. (1985b). Multicomponent latent trait models for test design. In S.E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 195-218). Orlando: Academic Press.
- Embretson, S.E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 380-396.

- Fischer, G. H. (1968). *Psychologische testtheorie*. Bern: Huber.
- Fischer, G. H. (1973). Logistic latent trait models with linear constraints. *Psychometrika*, 48, 3-26.
- Glück, J., Machat, R., Jirasko, M., & Rollett, B. (2002). Training-related changes in solution strategy in a spatial test: An application of item response models. *Learning and Individual Differences*, 13, 1-22
- Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30, 141-151.
- Hochberg, J., & Gellman, L. (1977). The effect of landmark features on mental rotation times. *Memory & Cognition*, 5, 23-26.
- Hoskens, M. & De Boeck, P. (1995). Componential IRT models for polytomous items. *Journal of Educational Measurement*, 32, 364-384.
- Huang, C.W. (2003). *Psychometric analyses based on evidence-centered design and cognitive science of learning to explore students' problem-solving in physics*. Doctoral dissertation, University of Maryland, College Park.
- Janssen, R., & De Boeck, P. (1997). Psychometric modeling of componentially designed synonym tasks. *Applied Psychological Measurement*, 21, 37-50.
- Junker, B. W. (1999). Some statistical models and computational methods that may be useful for cognitively-relevant assessment. Paper prepared for the Committee on the Foundations of Assessment, National Research Council.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25, 258-272.
- Keats, J.A. (1971). *An Introduction to Quantitative Psychology*. Sydney: John Wiley.
- Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement*, 27, 307-327.
- Klein, M. F., Birenbaum, M., Standiford, S. N., & Tatsuoka, K. K. (1981). *Logical error analysis and construction of tests to diagnose student "bugs" in addition and subtraction of fractions* (Research Report 81-6). Urbana, IL: Computer-based Education Research Laboratory, University of Illinois.

- Linn, R.L. (Ed.) (1989) *Educational measurement* (3rd Ed.) New York: American Council on Education/Macmillan.
- Luce, R.D., & Tukey, J.W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, 1, 1-27.
- Kyllonen, P.C., Lohman, D.F., & Snow, R.E. (1984). Effects of aptitudes, strategy training, and test facets on spatial task performance. *Journal of Educational Psychology*, 76, 130-145.
- Martin, J.D., & VanLehn, K. (1995). A Bayesian approach to cognitive assessment. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 141-165). Hillsdale, NJ: Erlbaum.
- McCloskey, M. (1983). Intuitive physics. *Scientific American*, 248(4), 122-130.
- Meijer R.R., & Sijtsma K. (2001) Methodology review: evaluating person fit. *Applied Psychological Measurement* 25, 107-135
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23.
- Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88, 355-383.
- Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. New York: Cambridge University Press.
- Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology*, 10, 639-667.
- Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika*, 59, 439-483.
- Mislevy, R.J. (2003). Substance and structure in assessment arguments. *Law, Probability, and Risk*, 2, 237-258.
- Mislevy, R.J., & Gitomer, D.H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction*, 5, 253-282.
- Mislevy, R.J., Steinberg, L.S., & Almond, R.A. (2002). Design and analysis in task-based language assessment. *Language Assessment*, 19, 477-496.

- Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15*, 363-378.
- Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195-215.
- Mislevy, R.J., Wingersky, M.S., Irvine, S.H., & Dann, P.L. (1991). Resolving mixtures of strategies in spatial visualization tasks. *British Journal of Mathematical and Statistical Psychology, 44*, 265-288.
- Mosenthal, P. B., & Kirsch, I. S. (1991). Toward an explanatory model of document process. *Discourse Processes, 14*, 147-180.
- Newell, A., & Simon, H.A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
- Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (1995). *Cognitively diagnostic assessment*. Hillsdale, NJ: Erlbaum.
- Perline, R., Wright, B.D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement, 3*, 237-255.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
- Rijmen, F., & De Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement, 26*, 271-285.
- Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design. In P. Robinson (Ed.), *Cognition and Second Language Acquisition* (287-318). Cambridge: Cambridge University Press.
- Roskam, E.E., & Jansen, P.G.W. (1984). A new derivation of the Rasch model. In E. Degreef & J. van Brugghenaut (Eds.), *Trends in mathematical psychology* (pp. 293-307). Amsterdam: North-Holland .
- Rost, J. (1990). Rasch models in latent classes - an integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271-282
- Schum, D.A. (1994). *The evidential foundations of probabilistic reasoning*. New York: Wiley.

- Scheiblechner, H. (1972). Das Lernen und Lösen komplexer Denkaufgaben. (The learning and solution of complex cognitive tasks.) *Zeitschrift für experimentelle und Angewandte Psychologie*, 19, 476-506.
- Shafer, G. (1976). *A mathematical theory of evidence*. Princeton: Princeton University Press.
- Sheehan, K. M., & Mislevy, R. J. (1990). Integrating cognitive and psychometric models in a measure of document literacy. *Journal of Educational Measurement*, 27, 255-272.
- Shepard, R. N. & Meltzer, J (1971) Mental rotation of three-dimensional objects. *Science*, 171, 701-703.
- Siegler, R.S. (1981). Developmental sequences within and between concepts. *Monograph of the Society for Research in Child Development*, Serial No. 189, 46.
- Simon, H. (1975). The functional equivalence of problem solving skills. *Cognitive Psychology*, 7, 268-288.
- Spada, H. & McGaw, B. (1985). The assessment of learning effects with linear logistic test models. In S. Embretson (Ed.), *Test design: New directions in psychology and psychometrics* (pp. 169-193). New York: Academic Press
- Snow, R.E., & Lohman, D.F. (1989). Implications of cognitive psychology for educational measurement. In R.L. Linn (Ed.), *Educational measurement* (3rd Ed.) (pp. 263-331). New York: American Council on Education/Macmillan.
- Sternberg, R. J. (1977). *Intelligence, information-processing, and analogical reasoning: The componential analysis of human abilities*. Hillsdale, NJ: Erlbaum
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20, 345-354.
- Thissen, D., Steinberg, L. & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, 67-113.
- Toulmin, S.E. (1958). *The uses of argument*. Cambridge: Cambridge University Press.
- Whitely, S. E. (1976). Solving verbal analogies: Some cognitive components of intelligence test items. *Journal of Educational Psychology*, 68, 234-242.

Wilson, M.R. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin*, 105, 276-289.