On the interplay between nonparametric and parametric IRT, with some thoughts about the future

Brian Junker [1]

ABSTRACT In this chapter I review some of the important research in nonparametric and parametric item response theory (IRT) today, and consider some current measurement challenges in education and cognitive psychology. This leads to assessment models that do not look very much like today's IRT models, but for which the tools and conceptual framework of nonparametric and parametric IRT are still quite well suited.

1 Introduction

In introducing Susan Embretson's 1999 Presidential Address at the European Meeting of the Psychometric Society in Lüneburg Germany, Ivo Molenaar defined psychometrics as "mathematical statistics in the service of substantive psychology": that definition cuts a pretty wide swath, and indicates just how general the psychometric enterprise can and should be. Item response theory (IRT; e.g., Fischer & Molenaar, 1995; Van der Linden & Hambleton, 1997) is a psychometric approach to modeling data from social surveys and educational and psychological tests, dating back at least to Lord (1952) and Rasch (1960), and to the work of Loevinger and Guttman before them. IRT enables us to study the characteristics of test or survey items across multiple respondent populations, and to study respondents' propensities to answer positively across various items. IRT has arguably been one of the most successful and widely used techniques in psychometrics, with applications in developmental, social, educational and cognitive psychology for example, as well as in medical research, demography and other social science settings.

In this chapter I will try to briefly summarize some of the important research in nonparametric and parametric IRT today, emphasizing the interplay between parametric and nonparametric models that is the hallmark of the approach initiated in the Netherlands by Mokken and pursued by Molenaar, Sijtsma, and their colleagues, and re-ignited in the U.S. by Holland, Rosenbaum, and their colleagues. I will try to show that a broad understanding of IRT as an instance of "mathematical statistics in the service of substantive psychology", together with an appreciation

[1] Department of Statistics, Carnegie Mellon University, 232 Baker Hall, Pittsburgh PA 15213, brian@stat.cmu.edu

of some of the current measurement challenges in education and cognitive psychology, lead us to assessment models that do not look very much like today's IRT models, but for which the tools and conceptual framework of nonparametric and parametric IRT are particularly well suited.

2 Nonparametric IRT: Scale Construction

To save space and preserve focus, my summary of nonparametric IRT will concentrate on the scaling theory techniques introduced by Mokken and pursued by Molenaar, Sijtsma, and their colleagues. Important related work on nonparametric essential unidimensionality (e.g., Stout 1987, 1990; Zhang & Stout, 1996; and Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996), nonparametric regression estimates of item response functions and test response surfaces (especially Ramsay, 1991, 1995, 1996), and related parametric and nonparametric work (e.g., Meijer, 1996; Cliff & Donoghue, 1992; Drasgow, Levine, Tsien, Williams, & Mead, 1995; Samejima, 1997) will not be considered in detail.

2.1 *Monotone Homogeneity / Strict Unidimensionality*

The Mokken (1971) model of *monotone homogeneity* starts with very few assumptions. A collection of item response variables $\boldsymbol{X} = (X_1, \ldots, X_k)$, which may include dichotomous, ordered polytomous, or even continuous responses satisfies this model, if θ, the latent trait, is a real-valued random variable (unidimensionality); if each item step response function (ISRF) $P[X_i > c|\theta]$ is nondecreasing in θ for each item response variable X_i and each real threshold c (monotonicity); and if

$$P[X_1 > c_1, \ldots, X_k > c_k|\theta] = \prod_{i=1}^{k} P[X_i > c_i|\theta] \qquad (1.1)$$

for all possible cutoffs c_i (local independence). When the response is dichotomous ($X_i = 0$ or 1) I refer to $P_i(\theta) \equiv P[X_i = 1|\theta]$ as the item response function (IRF). In that case, equation (1.1) reduces to the familiar form

$$P[X_1 = x_1, \ldots, X_k = x_k] = \prod_{i=1}^{k} P_i(\theta)^{x_i}[1 - P_i(\theta)]^{1-x_i}. \qquad (1.2)$$

The data we observe when examinees or subjects respond to test or survey items can be thought of as i.i.d. samples from the marginal discrete multivariate distribution,

$$P[X_1 > c_1, \ldots, X_k > c_k] = \int P[X_1 > c_1, \ldots, X_k > c_k|\theta]\, dF(\theta), \qquad (1.3)$$

where the integrand comes from either equation (1.1) or equation (1.2), and $dF(\theta)$ represents the distribution of θ in the population of interest. The assumptions of unidimensionality, monotonicity and local independence can be relaxed in various ways (e.g., Sections 2.2, 3.2 and 4.1 below), but this basic model has been the foundation of much nonparametric scale construction.

Nonparametric scale construction

For dichotomous items, a nonparametric theory of scale construction has existed at least since Mokken (1971; 1997; see also Molenaar, 1991; 1997). The principal tools of that theory are adaptations of Loevinger's (1948) H coefficients, comparing the marginal covariance Cov (X_i, X_j) of each item pair with the maximum covariance Cov$_{max}(X_i, X_j)$ possible, preserving the margins of the observed $X_i \times X_j$ table. The bound Cov$_{max}(X_i, X_j)$ is obtained by adjusting the table to remove Guttman errors (e.g., Molenaar, 1991); and indeed the original formulas for the H coefficients were expressed as ratios of Guttman errors (Mokken, 1997).

A related modeling condition for dichotomous items is *invariant item ordering* which says that IRF's do not cross: for every pair i and j, either $P_i(\theta) \leq P_j(\theta)$ or $P_j(\theta) \leq P_i(\theta)$ holds uniformly for all θ. Rosenbaum (1987a,b) and Sijtsma and Junker (1996, 1997) explore and extend this idea and provide scale construction examples. Mokken's *double monotonicity* model incorporates both the monotone homogeneity and invariant item ordering assumptions.

The H coefficients are directly sensitive only to high or low correlations between items, rather than to local independence given θ as in equations (1.1) and (1.2). If the correlations are near zero, we may be unsatisfied to assume that such a θ exists (see for example the discussion of co-monotonicity in Junker & Ellis, 1997). While a perfect Guttman scale would produce H coefficients equal to one, large H coefficients provide only indirect evidence of a θ "explaining" covariation in the item responses in the sense of local independence.

More direct attacks on the problem of establishing such a θ from data analysis have been pursued by Stout, Ramsay and their students and colleagues. Stout (1990; and subsequent work, for example Stout et al., 1996) constructs a proxy for θ from the total score on a specially-selected subset of the items and uses it to test a weakened version of monotone homogeneity, Stout's essential unidimensionality model. Ramsay (1991) constructs nonparametric regression estimates of item (step) response functions using total score or rest score (see Section 2.1 below for definitions) as a proxy for θ, which allows one to explore non-monotonicity. Ramsay (1995) constructs nonparametric local smoothing estimates of the joint response surface of all items on the test. Stout's and Ramsay's methods are generally more computationally complex, and seem to require larger examinee and item sample sizes, than the methods initiated by Mokken and developed by Molenaar, Sijtsma, Rosenbaum and their colleagues and students. Thus the Mokken techniques have been more widely used in smaller social survey and experimental psychology settings.

Molenaar (1991) provided a direct generalization of the H coefficients to the case of polytomous responses, and developed an effective computational method for obtaining $\text{Cov}_{max}(X_i, X_j)$ in the polytomous case. Hemker, Sijtsma and Molenaar (1995) apply these ideas to scaling polytomous items. Generalizing invariant item ordering to the polytomous case turns out to be somewhat delicate. The most successful approach to date is based on preserving the order of expected item scores, e.g. $E[X_i|\theta] \leq E[X_j|\theta]$ uniformly in θ for fixed i and j; Sijtsma and Hemker (1998) compare this approach with other approaches (e.g., Molenaar, 1997; Scheiblechner, 1995) that impose order restrictions directly on the ISRF's $P[X_i > c|\theta]$, uniformly in θ.

Another approach to understanding scaling by the Mokken model—in dichotomous, polytomous, and more general settings—was initiated by Holland (1981), Rosenbaum (1984), and Holland and Rosenbaum (1986; see also Meredith, 1965). Junker (1993), Ellis and Junker (1997) and Junker and Ellis (1997) continued this approach; they combine Holland and Rosenbaum's (1996) *conditional association (CA)* condition

$$\begin{aligned}&\forall \text{ partitions } \boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{Z}),\ \forall\, f,\ g \text{ non-decreasing; } \forall h(\boldsymbol{Z}),\\ &\text{Cov}\,(f(\boldsymbol{Y}), g(\boldsymbol{Y})|h(\boldsymbol{Z})) \geq 0,\end{aligned} \tag{1.4}$$

with a *vanishing conditional dependence* condition

$$\begin{aligned}&\forall k,\ \text{as } m \to \infty,\ (X_1, \ldots, X_k) \text{ become independent,}\\ &\text{given } (X_{k+1}, \ldots, X_{k+m}),\end{aligned} \tag{1.5}$$

to obtain a complete characterization of an infinite-item-pool formulation of the basic monotone homogeneity model, in which θ is both genuinely latent and consistently estimable, in terms of the joint distribution of observable item responses.

Stochastic ordering

A side effect of the effort to understand how to characterize and test the monotone homogeneity model has been a selection of other model testing criteria, such as Junker's (1993) "manifest monotonicity" property for dichotomous items following the monotone homogeneity model,

$$P[X_i = 1|X_+^{(-i)} = s] \text{ is non-decreasing in } s. \tag{1.6}$$

This property, and examples showing that it does not hold when the "rest score" $X_+^{(-i)} = \Sigma_{j\neq i} X_j$ is replaced by the total score $X_+ = \Sigma_{i=1}^k X_i$, are included in unpublished work of Molenaar and Tom Snijders (Junker, 1993; Junker & Sijtsma, 2000).

Existing proofs of (1.6) hinge on establishing a "stochastic ordering" property for θ, given the total score X_+ [or equivalently the rest score $X_+^{(-i)}$]:

$$P[\theta > c|X_+ = s] \text{ is non-decreasing in } s,\ \forall\, c. \tag{1.7}$$

Hemker, Sijtsma, Molenaar and Junker (1996, 1997) call this property "SOL" (Stochastic Ordering of the Latent trait by the sum score), and show that, surprisingly, this property does not generalize to "most" nonparametric ordered-polytomous response IRT models. Thus for example, rules based on cutoffs for

X_+ need not be most powerful for "mastery decisions" in the sense of $\theta > c$; on the other hand, such cutoff rules for X_+ *are* most powerful for mastery decisions in the nonparametric dichotomous response case (Grayson, 1988; Huynh, 1994).

In the process of developing these stochastic ordering ideas, Hemker et al. (1997) and Hemker and Sijtsma (1999) have developed a taxonomy of nonparametric and parametric item response models, that usefully complements the taxonomy of Thissen and Steinberg (1986). The Hemker taxonomy is based on the cumulative, continuation-ratio, and adjacent-category logits that are commonly used to define parametric families of polytomous IRT models. Common forms of graded response models (GRM; Samejima, 1997 for example), sequential models (SM; Tutz, 1990; Mellenbergh, 1995; Samejima, 1969, 1995), and partial credit models (PCM; Masters 1982) assume, respectively, that the logit functions $\operatorname{logit} P[X_i > c|\theta]$, $\operatorname{logit} P[X_i > c|X > c-1, \theta]$, and $\operatorname{logit} P[X_i = c+1|X_i \in \{c, c+1\}, \theta]$ are *linear* in θ. Hemker's analogous nonparametric model classes, the np-GRM, np-SM and np-PCM, assume only that these logits are *non-decreasing* in θ.

This taxonomy is a powerful way to organize ideas about model definition and model development in applications of both parametric and nonparametric IRT. It follows that the np-PCM class is nested within the np-SM class, which is nested within the np-GRM class; moreover all three linear-logit families above (GRM, SM and PCM) are in fact subclasses of the np-PCM class. As Hemker and Sijtsma (1999) and Van der Ark (1999) show, this approach also highlights links between polytomous IRT and the machinery of generalized linear models (McCullagh & Nelder, 1989), just as it has long been realized that parametric dichotomous IRT is basically multivariate mixed effects logistic regression (e.g., Douglas & Qui, 1997; see also Lee & Nelder, 1996). It is important to realize however that of all of the models studied by Hemker and his colleagues, only the parametric PCM and its special cases, have the nice stochastic ordering property SOL (see 1.7).

2.2 Some Interesting Questions

An ongoing question in this area is developing adequate data analysis methodology. Most of what can now be done, in the dichotomous and polytomous cases, is encoded in the computer program MSP (Molenaar & Sijtsma, 1999). Sijtsma (1998) provides an excellent survey of nonparametric IRT approaches to the analysis of dichotomous item scores; Molenaar (1997) and Sijtsma and Van der Ark (this volume) survey extensions to the polytomous case. Snijders (this volume) introduces Mokken scaling tools for multilevel data as well. Ellis (1994) has re-examined Mokken's hypothesis testing framework for the H coefficients, and developed, in principle, new tests based on the theory of order-restricted inference of Robertson, Wright and Dykstra (1988). The same methods may be useful to develop tests of manifest monotonicity. In addition to Holland and Rosenbaum's (1986) applications of the Mantel-Haenzel test, Yuan and Clarke (1999) have developed asymptotic theory for testing the conditions of Junker (1993), that should be adaptable to the conditions of Junker and Ellis (1997). A more direct application of the theory of order-restricted inference to testing CA and related conditions is given by Bartolucci and Forcina (in press).

Sijtsma and Van der Ark (this volume) discuss progress on several problems related to the lack of SOL (1.7) in ordered polytomous IRT models and to the sensitivity and specificity of the manifest monotonicity condition (1.6) for detecting (violations of) the monotone homogeneity model. One of the strengths of the nonparametric approach to dichotomous IRT is that it usually assures us, under very general circumstances, that simple summaries of the data are informative about inferences we wish to make, yet the current evidence suggests that there are no such simple summaries for inferences about ordered polytomous data. Understanding the impact that this has on nonparametric polytomous IRT modeling will surely entail facing and solving the problems that Sijtsma and Van der Ark discuss.

Finally, some of the machinery developed to characterize monotone homogeneity models seems ready to apply to common modifications of this basic model. For example, conditional association (1.4) is basically an extreme sharpening of the well-known fact that inter-item correlations are nonnegative under monotone homogeneity. Post (1992; Post & Snijders, 1993) has established a similar fact about a class of nonparametric probabilistic unfolding models: the inter-item correlation matrix has a band of positive correlations near the main diagonal, bordered by bands of negative correlations. Is there a sharpening of the Post result analogous to conditional association? Could this be combined with the VCD condition of (1.5) to produce a characterization of Post's models? Could such a result follow from a "folding" of the monotone homogeneity model to produce nonparametric unfolding models, along the lines of Verhelst and Verstralen (1993) or Andrich (1996)? In another direction, much of the work in Ellis and Junker (1997) and Junker and Ellis (1997) does *not* depend on the latent variable θ being unidimensional. Junker and Ellis (1997) for example point out that their item-step "true scores" $P[X_i > c|\tau(X)]$ should form a manifold of the same dimension as the underlying latent variable θ. A characterization theorem may again result, if a weakening of conditional association to accommodate multidimensional θ could be developed. Such theorems help to distinguish among monotone homogeneity models, probabilistic unfolding models, and multidimensional nonparametric IRT models, on the basis of observable data.

3 Parametric IRT: Modeling Dependence

Parametric IRT, as surveyed for example in the edited volumes of Fischer and Molenaar (1995) and Van der Linden and Hambleton (1997), is a well-established, wildly successful statistical modeling enterprise. A basic and familiar model in this area is the "two-parameter logistic", or 2PL, model for dichotomous item response variables (e.g., Chapter 1 of Van der Linden & Hambleton, 1997), given by the monotone homogeneity assumptions in Section 2 and the assumption of a logistic form for the item response functions, $P[\, X_{vi} = 1 \mid \theta_v, \alpha_i, \beta_i] = 1/\{1 + \exp(-\alpha_i[\theta_v - \beta_i])\}$, describing the dichotomous response of examinee S_v to item I_i. The "discrimination" parameter α_i controls the rate of increase of this logistic curve, and is directly related to the Fisher information for estimating θ_v, and the "difficulty" parameter β_i is the location on the θ_v scale at which this information is maximal; note also that at $\theta_v = \beta_i$, $P[X_{vi} = 1] = 1/2$. The 3PL (three-parameter logistic) model extends the 2PL model by adding a non-zero lower asymptote to each item response function; on the other hand the Rasch or 1PL (one-parameter logistic) model is a restriction of the 2PL model obtained by setting α_i identically equal to some constant, usually 1.

Such parametric IRT models, extended by hierarchical mixture/Bayesian modeling and estimation strategies, make it possible in principle and in practice to incorporate covariates and other structure. Many violations of the basic local independence assumption of IRT models are in fact due to unmodeled heterogeneity of subjects and items, that can now be explicitly modeled using these methods. These models have greatly extended the data analytic reach of psychometricians, social scientists, and educational measurement specialists.

The main purpose of this section is to introduce a general modeling framework and highlight a few developments in parametric IRT, some old and some new, that will be relevant to my discussion of applying IRT and related models to cognitive assessment problems in Section 4 below. My summary of parametric IRT will be even less complete, relative to the vast parametric IRT literature, than my summary of nonparametric IRT.

3.1 Two-Way Hierarchical Structure

The estimation of group effects and the use of examinee and item covariates in estimating item parameters plays an important role in the analysis of large multi-site educational assessments such as the National Assessment of Educational Progress (NAEP; e.g., Algina, 1992; Johnson, Mislevy and Thomas, 1994; and Zwick 1992). These efforts, which go back at least to Mislevy (1985; see also Mislevy & Sheehan, 1989) can be recognized as the wedding of hierarchical linear or multi-level modeling methodology with standard dichotomous and polytomous IRT models. The general model is a two-way hierarchical structure for n individuals and k response variables, as in Table 1.1, where $P(\theta_v; \boldsymbol{\gamma}_i)$ is the IRF, depending on person parameters θ_v and item parameters $\boldsymbol{\gamma}_i$ (e.g., $\boldsymbol{\gamma}_i = (\alpha_i, \beta_i)$ in the 2PL model above), and where independence is assumed between i's conditional on θ_v at the first level and between v's at the second level. Terms in the first level, for example, are multiplied together to produce the usual joint likelihood for the $n \times k$ item response matrix $[X_{vi}]$; the second and third levels can be used to impose constraints on the first level parameters and latent variable, to deduce what integrations are needed for marginal likelihood approaches, etc. λ_f and λ_g represent sets of hyperparameters needed to specify these person distributions f_v and item distributions g_i, with hyperprior distributions ϕ_f and ϕ_g, respectively. The model in Table 1.1 is expressed for dichotomous items, for simplicity of ex-

First level:	X_{vi}	$\sim$	$P(\theta_v;\gamma_i)^{X_{vi}}[1-P(\theta_v;\gamma_i)]^{(1-X_{vi})}$, $v=1,\ldots,n;\ i=1,\ldots,k$
Second level:	θ_v	$\sim$	$f_v(\theta\|\lambda_f)$, each v
	γ_i	$\sim$	$g_i(\gamma\|\lambda_g)$, each i
Third level:	λ_f	$\sim$	$\phi_f(\lambda_f)$
	λ_g	$\sim$	$\phi_g(\lambda_g)$

TABLE 1.1. two-way hierarchical structure for n individuals and k dichotomous response variables.

position, but can easily be generalized to polytomous items, or combinations of item types (see for example Patz & Junker, 1999a; 1999b). It is also usual to assume for $f_v(\theta)$ a single latent trait distribution not depending on v, and similarly for g_i.

We may relax these assumptions by allowing the distributions $f_v(\cdot)$ of θ to depend hierarchically on examinee covariates, to model population heterogeneity, as in the multi-group IRT models of Mislevy (1985) and Bock and Zimowski (1997), or to reflect hierarchical linear structure as in Fox and Glas (1998). We may also elaborate $g_i(\cdot)$, for example by building linear structure into the item parameters. For example in the 2PL model, where $\gamma_i = (\alpha_i, \beta_i)$, we might take

$$(\beta_1 \quad \beta_2 \quad \cdots \quad \beta_{k-1} \quad \beta_k)' = Q (\psi_1 \quad \psi_2 \quad \cdots \quad \psi_m)', \tag{1.8}$$

where Q is an appropriate design matrix of full column rank, to reflect common sources (ψ_ℓ's) of item difficulty (β_i's) across items. In the case of Rasch (1PL) IRF's, this is the linear logistic test model (LLTM; Scheiblechner, 1972; Fischer, 1973). This model and its various generalizations (e.g., Glas & Verhelst, 1989; Patz & Junker, 1999b) continues to be used for psychological experiments with multiple outcomes per subject (e.g., Fischer and Molenaar, 1995 and the references therein) and for research in cognitively-motivated test design (Embretson, 1995; 1999). There is no reason to restrict attention to the β's, and for example Embretson (1999) has explored a similar decomposition of the α's in a 2PL model.

These generalizations of the basic IRT model both simplify and unify parametric approaches to many thorny test analysis questions, including differential item functioning and item parameter drift, nonequivalent groups and vertical equating, two-stage testing and matrix-sampled educational assessment survey work, etc. The computer programs ConQuest (Wu, Adams, & Wilson, 1997) and BILOG-MG (Zimowski, Muraki, Mislevy & Bock, 1997) provide fairly general E-M based solutions when the underlying IRT model is the 1PL (ConQuest), or 2PL or 3PL (BILOG-MG). Fox and Glas (1998) and Patz and Junker (1999a; 1999b) give two different Markov chain Monte Carlo (MCMC) approaches to the problem.

Generalization to polytomous items, facets-style rated response models, and mixtures of item types are conceptually, and often computationally, straightforward; see for example Glas and Verhelst (1989) and Patz and Junker (1999a; 1999b).

3.2 Some Multidimensional Models

Research in multidimensional IRT models has concentrated on additive and conjunctive combinations of multiple traits to produce probabilities of response. Additive models, known as *compensatory* models in much of the literature, replace the unidimensional latent trait θ with an item-specific, known (e.g., Stegelmann, 1983; Embretson, 1991; Kelderman and Rijkes, 1994; and Adams, Wilson, & Wang, 1997) or unknown (e.g., Reckase, 1985; Wilson, Wood, & Gibbons, 1983; Fraser & MacDonald, 1988; Muraki & Carlson, 1995) linear combination of components $a_{i1}\theta_1 + \cdots + a_{id}\theta_d$ of a d-dimensional latent trait vector, for example in the dichotomous response case $P[X_{vi} = 1|\theta_{v1}, \ldots, \theta_{vd}] = P(a_{i1}\theta_{v1} + \cdots + a_{id}\theta_{vd} - \beta_i)$, where $P(\cdot)$ might be the logistic or probit response function for example. Béguin and Glas (1998) survey the area well (see also several contributed chapters in Van der Linden & Hamilton, 1997) and give an MCMC algorithm for estimating these models; Gibbons and Hedeker (1997) pursue related developments in biostatistical and psychiatric applications.

Conjunctive models are often referred to as *noncompensatory* or *componential* models in the literature. These models (e.g., Embretson, 1997) combine unidimensional models for components of response multiplicatively, so that $P[X_{vi} = 1|\theta_{v1}, \ldots, \theta_{vd}] = \prod_{\ell=1}^{d} P_{i\ell}(\theta_{v\ell})$ where $P_{i\ell}(\theta_{v\ell})$ are parametric unidimensional dichotomous response functions. The usual interpretation is that the $P_{i\ell}(\theta_{v\ell})$ represent skills or subtasks all of which must be performed correctly in order to generate a correct response to the item itself. Janssen and De Boeck (1997) give a recent application.

Compensatory structures are attractive because of their conceptual similarity to factor analysis models. They have been very successful in aiding the understanding of how student responses can be sensitive to major content and skill components of items, and in aiding parallel test construction when the underlying response behavior is multidimensional (e.g., Ackerman, 1994). Noncompensatory models are largely motivated from a desire to model cognitive aspects of item response, a topic to which we will return in Section 4. Embretson (1997) also reviews blends of these two approaches (her general component latent trait models; GLTM).

3.3 Models That Accommodate Extra Behavioral Features of Assessment

In addition to providing a way to model dependence of item responses on specific examinee and item covariates, the hierarchical or multi-level approach to IRT also allows us to model extra behavioral features of assessment. This largely unexplored area is worth further study, since these features can affect both the certainty with which we make inferences from assessment data and the kinds of inferences we make; current models and methods largely relegate them to the "error distribution" of the model.

One example of this sort of work involves recent efforts to more elaborately model the behavior of raters in rated item response data. When only one rater rates each item, it may be sufficient to treat each rating as a different, locally independent pseudo-item—so that the first level in Table 1.1 contains one factor for each rater × item × examinee combination—and to model the rater effect as a linear influence on the item's difficulty parameter β_i. Mathematically this is equivalent to the LLTM model sketched above, but it has come to be known in this setting as the "Facets model" (e.g., Linacre, 1989; Engelhard, 1994).

For both formative and summative evaluation of raters, a number of multiple-read rating designs are now commonplace (Wilson & Hoskens, 1999), including designs with as many as six raters per item (e.g., Sykes & Heidorn, 1999). Thus each examinee performance is measured several correlated but fallible times. Junker and Patz (1998) showed that the usual Facets model formulation in which the likelihood is the product of LLTM-style factors for each rating of each item, accumulates information about θ too optimistically, so that even with only one item response, in the limit as the number of raters grows, the standard error for estimating or predicting θ apparently goes to zero, contradicting the notion (e.g., Junker, 1993) that the number of items should tend to infinity in order to make the error of estimation of θ vanishingly small. Instead, models are needed that appropriately accumulate information from multiple ratings to the single item response being rated, and then accumulate information across item responses to learn about θ itself.

Wilson and Hoskens' (1999) rater bundle model (RBM) attacks this problem by replacing the Facets product across raters for each item in level one of Table 1.1 with a log-linear model that models the dependence between ratings of the item, conditional on θ. Their approach seems very useful for, e.g., modeling "table effects" and other rater dependence phenomena that follow when raters are allowed or encouraged to discuss ratings amongst themselves to increase rating quality and inter-rater reliability.

Patz, Junker and Johnson's (1999) hierarchical rater model (HRM) provides an alternative approach that posits a "latent rating" ξ_{vi} (not unlike Maris', 1995, notion of latent responses; also note the connection with data augmentation methods in Bayesian computation, e.g., Tanner, 1996) that follows a conventional IRT model, such as $P[\xi_{vi} = 1|\theta_v, \alpha_i, \beta_i] = 1/\{1 + \exp(-\alpha_i[\theta_v - \beta_i])\}$ (or a polytomous variant), and a "signal detection model" for each rater r rating that item response such as $P[X_{vir} = 1|\xi_{vi}] = p_{11r}^{\xi_{vi}} p_{10r}^{1-\xi_{vi}}$, where p_{11r} is the probability that rater r rates a response with latent rating $\xi_{vi} = 1$ as a 0, and p_{10r} is the probability that rater r rates a response with latent rating $\xi_{vi} = 0$ as a 1. The factors $P[X_{vir} = 1|\xi_{vi}]$ now appear at level one of the hierarchy in Table 1.1, and the factors $P[\xi_{vi} = 1|\theta_v, \alpha_i, \beta_i]$ form a new level between levels one and two of Table 1.1. This produces a model that behaves, for multiple discrete ratings, much as a standard generalizability theory model would behave for multiple continuous ratings (see Verhelst & Verstralen, this volume, for a similar development). In addition to providing an appropriate way to combine information from multiple raters to learn about student performance, the HRM makes possible calibration and monitoring of individual rater behavioral effects such as, in the case of polytomous responses, rater severity and rater precision.

3.4 Some Current Questions

Almost any assessment phenomenon—from between-examinee dependence due to institutional or sociological factors, to behavioral aspects of raters, to the analysis of item responses into requisite examinee skills or item features—can be expressed in the hierarchical mixture/Bayes modeling framework, because of its conceptual simplicity. Recent advances in computation, and MCMC methods in particular, have made it possible to estimate a vastly wider variety of these models than would have been imaginable even ten years ago. Questions motivating this work inevitably involve identifying phenomena that are worth detailed parametric modeling, and seeing if the computational machinery can be pushed to estimate models of these phenomena. Recent examples include multidimensional (Gibbons & Hedeker, 1992) and hierarchical (Bradlow, Wainer, & Wang, 1999) modeling of testlets; blending IRT and hierarchical linear models, and behavioral models such as the rater models described above.

Speeding up the computations with approximations (including formal and informal applications of Laplace's method such as Rigdon & Tsutakawa, 1983 and Kass, Tierney, & Kadane, 1990; blends of Monte Carlo and E-M approaches as surveyed in Tanner, 1996; and variational methods, e.g., Jaakkola & Jordan, 1999) continues to be an essential and fruitful avenue of research. An avenue that has not been explored as much in IRT work is choosing models for which sufficient statistics are simple and interpretable. The most familiar "basic model" of this sort is of course the Rasch model, but as we shall see in the next section some current cognitively-motivated measurement problems may require a new "basic model".

4 Measurement Challenges Posed by Cognitive and Embedded Assessments

In recent years, as cognitive theories of learning and instruction have become richer, and computational methods and machinery to support assessment have become more powerful, there has been increasing pressure to make assessments truly criterion referenced, that is, to "report" on student achievement relative to theory-driven lists of examinee skills, beliefs and other cognitive features needed to perform tasks in a particular assessment domain. For example Baxter and Glaser (1998) and Nichols and Sugrue (1999) present compelling cases that assessing examinees' cognitive characteristics can and should be the focus of assessment design. In a similar vein, Resnick and Resnick (1992) advocate standards-referenced or criterion-referenced assessment closely tied to curriculum, as a way to inform instruction and enhance student learning.

Appropriate criterion-referenced testing can also be an effective teaching tool when embedded directly in teaching practice. Indeed there is substantial argument and evidence, as summarized for example by Bloom (1984), that part of what distinguishes higher student achievement in “mastery learning” and individualized tutoring settings as opposed to the conventional classroom, is the use of frequent and relatively unobtrusive formative tests coupled with feedback for the students and corrective interventions by the instructor, and follow-up tests to determine how much the interventions helped. This approach continues to be advocated as part of a natural and effective apprenticeship style of human instruction (e.g., Gardner, 1992), and it is the basis of many computer-based intelligent tutoring systems (ITS’s, e.g., Anderson, 1993; and more broadly Shute & Psotka, 1996). Here too, a decomposition of assessment items into appropriate cognitive attributes is important: feedback and/or corrective action in a mastery class or from an ITS depends on knowing which cognitive attributes the examinee has and has not mastered.

Cognitive assessment models must generally deal with a more complex goal than linearly ordering examinees, or partially ordering them in a low-dimensional Euclidean space, which is what IRT has been designed and optimized to do. The goal of cognitive assessment can be thought of producing, for each examinee, a checklist of skills or other cognitive attributes that the examinee may or may not possess, based on the evidence of tasks performed by the examinee. The checklist of cognitive features in a cognitive assessment generally comes from an analysis of the cognitive attributes needed to successfully perform each task in a domain of interest. For a particular set of tasks, this analysis can be encoded in an incidence matrix, the *Q-matrix* (e.g., Tatsuoka, 1990), which we will write as a $k \times d$ matrix $Q = [Q_{i\ell}]$ of 0’s and 1’s with entries

$$Q_{i\ell} = \begin{cases} 1, & \text{if attribute } \ell \text{ is required by task } i \\ 0, & \text{if not} \end{cases} \tag{1.9}$$

While the Q-matrix does not capture all of the structure we may be interested in (Q treats the skills in a flat, non-time-ordered manner, and there may be both hierarchical and time-order structure in the skills as they are applied to a task), it is a useful bookkeeping device.

Many attempts (e.g., as surveyed by Roussos, 1994) to blend IRT and cognitive measurement are based on a linear decomposition of item parameters, as in the LLTM, or on a linear decomposition of the latent trait θ, as in the multidimensional compensatory IRT models. Compensatory IRT models, like factor analysis models, can be sensitive to relatively large components of variation in examinee ability or propensity to answer items correctly. LLTM-style models can be sensitive to finer components of variation *among items* but are not at all sensitive to components of variation *among examinees*. Noncompensatory approaches (e.g., Embretson, 1997) are intended to be sensitive to finer variations among examinees, in situations in which several cognitive components are required simultaneously for successful task performance. Of course, whether cognitive assessment data actually supports models that track this finer level of variation is an empirical matter.

4.1 Three Approaches to Cognitive Assessment

To illustrate the differences between traditional IRT approaches and cognitively-motivated approaches to assessment, we consider two published models intended to deal with essentially the same data: task performance by students learning the LISP programming language using one of the computer based intelligent tutoring systems developed by John R. Anderson and his colleagues at Carnegie Mellon University (e.g., Anderson, Corbett, Koedinger, & Pelletier, 1995). The first model is the assessment model actually embedded in the tutoring software, as described by Corbett, Anderson and O'Brien (1995); the second is an IRT-based model for essentially the same data, as presented by Draney, Pirolli and Wilson (1995). Then we consider a third model to illustrate that, although the modeling traditions in IRT are not much like those in cognitive assessment modeling, the fundamental techniques and concepts from IRT modeling are quite useful in cognitive assessment.

The Corbett/Anderson/O'Brien model

The "knowledge tracing model" embedded in the LISP tutor software (Corbett, Anderson, & O'Brien, 1995) treats successful performance of a task at time t, $t = 1, 2, 3, \ldots$, as an uncertain indicator of possession of the necessary underlying skills at time t, according to the model

$$P[X_{vi}(t) = 1] = P[\xi_{vi}(t) = 1](1 - s) + (1 - P[\xi_{vi}(t) = 1])g, \tag{1.10}$$

where $X_{vi}(t) = 1$ or 0 indicates whether examinee S_v performed task I_i correctly at time t, $\xi_{vi}(t) = 1$ or 0 indicates whether examinee S_v possesses the requisite skills for task I_i at time t, and parameters s and g in (1.10) are universal "slipping" and "guessing" probabilities that accommodate uncertainty in predicting task performance $X_{vi}(t)$ from $\xi_{vi}(t)$. $P[\xi_{vi}(t) = 1]$ follows a simple conjunctive model (though more complex expressions along the lines of Mislevy, 1996, could be imagined):

$$P[\xi_{vi}(t) = 1] = \prod_{\ell=1}^{d} P[\alpha_{v\ell}(t) \geq Q_{i\ell}]. \tag{1.11}$$

with latent variables $\alpha_{v\ell}(t) = 1$ or 0, indicating whether or not student S_v possesses skill ℓ at time t, $t = 1, 2, 3, \ldots$ Entries of a Q-matrix, $Q_{i\ell}$, indicate whether skill ℓ is needed for task i, and skills combine conjunctively to predict task performance. It is worth noting that the $\xi_{vi}(t)$'s [or the $\alpha_{v\ell}(t)$'s] can be interpreted as playing the role of Maris's (1995) latent responses (they can also be interpreted in terms of the method of data augmentation in Bayesian computation, e.g., Tanner, 1996). Given current estimates of $P[\alpha_{v\ell}(t) \geq Q_{i\ell}]$, the tutor can both identify which skills need additional practice, and select items of suitable difficulty that exercise those skills, to assign to the student next.

Corbett et al. were particularly interested in how to gather evidence about $P[\alpha_{v\ell}(t)]$ as the number of opportunities t to apply rule ℓ increases—i.e. they are interested in modeling learning. To account for the order in which correct and incorrect actions are observed when skill ℓ is called for, they treat each $\alpha_{v\ell}(t)$ as a two-state hidden Markov chain in t, with an absorbing state at $\alpha_{v\ell}(t) = 1$, with corresponding observable evidence $a_{v\ell}(t) = 1$ or 0 that the correct action was performed when skill ℓ was called for. For brevity, denote $\omega = P[\alpha_{v\ell}(t) = 1 \mid \bar{\alpha}_{v\ell}(t-1) = 0]$, let "$\alpha_{v\ell}(t)$" stand for "$\alpha_{v\ell}(t) = 1$" and "$\bar{\alpha}_{v\ell}(t)$" stand for "$\alpha_{v\ell}(t) = 0$", and similarly for $a_{v\ell}(t)$ and $\bar{a}_{v\ell}(t)$, in what follows. They apply the Law of Total Probability and Bayes' Rule to relate the observable data to the hidden Markov states:

$$
\begin{aligned}
P[\alpha_{v\ell}(t) \mid a_{v\ell}(t)] &= P[\,\alpha_{v\ell}(t-1) \mid a_{v\ell}(t)\,] \\
&\quad + (1 - P[\,\alpha_{v\ell}(t-1) \mid a_{v\ell}(t)])\,\omega \\
P[\alpha_{v\ell}(t) \mid \bar{a}_{v\ell}(t)] &= P[\,\alpha_{v\ell}(t-1) \mid \bar{a}_{v\ell}(t)\,] \\
&\quad + (1 - P[\,\alpha_{v\ell}(t-1) \mid \bar{a}_{v\ell}(t)])\,\omega \\
P[\alpha_{v\ell}(t-1) \mid a_{v\ell}(t)] &= \{(1-s)P[\,\alpha_{v\ell}(t-1)\,]\} \\
&\quad / \{(1-s)P[\,\alpha_{v\ell}(t-1)\,] \\
&\quad + g\,P[\,\bar{\alpha}_{v\ell}(t-1)\,]\} \\
P[\alpha_{v\ell}(t-1) \mid \bar{a}_{v\ell}(t)] &= \{sP[\,\alpha_{v\ell}(t-1)\,]\} / \{sP[\,\alpha_{v\ell}(t-1)\,] \\
&\quad + (1-g)\,P[\,\bar{\alpha}_{v\ell}(t-1)\,]\}.
\end{aligned}
\tag{1.12}
$$

Given *a-priori* fixed values of ω, s, g, and a probability p_0 that each skill is in the learned state before the tutoring begins, we may substitute p_0 for $P[\alpha_{v\ell}(t-1)]$ when $t = 1$, and the above formulas give an algorithm for recursively updating $P[\alpha_{v\ell}(t)]$ on the basis of the observed sequence of correct and incorrect actions in the first t opportunities to apply rule ℓ. This is a particularly simple and fast computational method, capable of updating the tutor's model of the student's skills in real time as the student works with the tutor.

It is interesting to note that in order to produce a well-fitting model, Corbett et al. had to allow the probabilities ω, s, g, and the probability p_0 that each skill was already in the learned state before the tutoring began, to be perturbed differently from overall population values for each student S_v; effectively, they allowed these parameters to become random effects. Thus, in addition to the individual differences in skills acquisition that the model had been designed to detect, there were substantial individual differences in starting knowledge of the students, in tendency to slip or guess, and in the rate of learning, under this model.

The Draney/Pirolli/Wilson model

Draney, Pirolli and Wilson (1995) develop an LLTM-style model to analyze essentially the same data. The model they consider begins with an indicator $a_{vi\ell}(t) = 1$ if student S_v performs correctly when the tth opportunity to use skill ℓ occurs, under condition i; and $a_{vi\ell}(t) = 0$ otherwise. These $a_{vi\ell}(t)$ differ from Corbett et al.'s $a_{v\ell}(t)$ only in that the task that provides a context for performing the skill is allowed to affect the difficulty of correct skill performance. In the Draney et al. model, the "skill response functions" are given by

$$P[a_{vi\ell}(t) = 1 \mid \theta_v, \tau_i, \delta_\ell, \gamma] = \frac{1}{1 + \exp(-\theta_v + \tau_i + \delta_\ell - \gamma \log(t))}.$$

This model essentially decomposes the difficulty parameter β in the Rasch model according to a Q matrix, as in equations (1.8) and (1.9), in terms of parameters for task, τ_i, skill, δ_ℓ, and slope of the learning curve, γ. The logarithmic dependence on t is intended to follow the development of Anderson (1993, Appendix to Chapter 3). Also, if the decomposition of tasks into skills is complete, and the skills are of a suitable granularity, Anderson's ACT-R theory predicts that skill "performances" will be approximately independent of one another, given the relevant difficulty and student parameters. This is essentially a statement of local independence, so that the "skill response functions" above may be multiplied together in the usual way to form an IRT likelihood.

It is interesting to compare the two modeling approaches. For example, for a data set similar to that analyzed by Draney et al., the assessment model of Corbett et al. (p. 32; see also Draney et al., p. 115) would employ essentially four continuous latent variables, and 33 dichotomous latent attribute indicators, *per student tested*—in addition to 132 parameters to characterize features of the skills being assessed. Draney et al. provided equivalent or better fit to learning curves, employing *one* continuous latent variable per student tested (Draney et al., p. 109), and 36 parameters for the skills being tested (op. cit., p. 115). Thus, if the goal is to model learning curves, clearly the more complex Corbett et al. model is not needed.

However, the uses to which the two models can be put, and the substantive interpretations of estimated parameters in the two models, are very different. The Corbett et al. model is immediately useful for diagnosing individual differences in student task performance behavior by relating it directly to specific skills in the task decomposition of their task domain, but can only indirectly assess the validity and reliability of the tasks in the assessment, through fit to learning curves or other summaries of student task performance.

The Draney et al. model is not immediately useful for student diagnosis, since its student parameter is one-dimensional. After fitting the model, Draney et al. go back and rank students based on (empirical Bayes) estimated θ's (in the context of learning curves analysis, the θ's essentially code students' initial facility in the task domain prior to tutoring, much as the random effect version of p_0 does in the Corbett et al. model), and compare estimated skill performance difficulties to the students' aggregate θ distribution; see for example their Figure 5.1, p. 112. Such displays allow us to predict which skills a "typical" student with some fixed value of θ might be expected to perform correctly, and are very useful communication devices. However, detailed cognitive diagnosis of individual students on the basis of the θ's is not possible, without a post-hoc analysis of some sort. Indeed, for assessing whether individual students have learned particular skills, Draney et al. replace the IRT model with a Bayesian inference network that is focused on inferring the probability that an individual has learned each skill, using priors constructed from the fitted IRT model and new data from further attempts to perform the skill(s).

The utility of the Draney et al. IRT model for analyzing important task performance features, and aggregate examinee behaviors, should not be minimized however. For example, Junker, Koedinger and Trottini (2000) are using essentially the same modeling framework to develop a semi-automatic stepwise variable construction/model selection procedure, with the goal of identifying skills that were either too narrowly or too broadly defined in a cognitive tutor (these result in stylized deviations, or "blips" from the theoretically predicted learning curves for the skills). In a similar vein, Huguenard, et al. (1997; see also Patz et al., 1996) applied a polytomous version of the LLTM to study the relationship between task features and working memory load, using the IRT θ parameter to soak up residual between-subjects variation not modeled by experimental and working memory factors.

An IRT-like cognitive assessment model

As the preceding example makes clear, traditional parametric IRT approaches may not be well suited to individual assessment tasks in computer based intelligent tutoring systems. The same issues can also arise in standalone assessments that are designed to assess presence or absence of specific skills—rather than to sort examinees along a linear scale—based on performance on a fixed set of tasks given as a standalone test.

To illustrate this we consider a modified version of the assessment model of Corbett, Anderson and O'Brien (1995). We omit the hidden-Markov learning model and assume that examinee behavior is only observed at the task level, not the skill level. The new model, whose compoments are summarized in Table 1.2, can also be connected to multidimensional noncompensatory IRT models, since it is interpretable as a simplified version of Embretson's (1997) multicomponent latent trait (MLTM) model. A related, more general class of models is the class of probability matrix decomposition models (Maris, De Boeck, & Van Mechelen, 1996).

The goal is to try to estimate the $\alpha_{v\ell}$'s, or more precisely $P[\alpha_{v\ell}=1]$, from the task performance data. Our basic response model [level one in the hierarchy of Table 1.1] is

$$P[X_{vi}=1|\boldsymbol{\xi},\boldsymbol{s},\boldsymbol{g}] \quad = \quad \xi_{vi}(1-s_i)+(1-\xi_{vi})g_i \;=\; (1-s_i)^{\xi_{vi}}g_i^{1-\xi_{vi}},$$

and so for the entire examinees by tasks matrix $[X_{vi}]$ of task responses,

$$\begin{aligned} &P[X_{vi}=x_{vi},\ \forall\, v,i \mid \boldsymbol{\xi},\boldsymbol{s},\boldsymbol{g}] \\ &\quad= \prod_v \prod_i \left[(1-s_i)^{\xi_{vi}}g_i^{1-\xi_{vi}}\right]^{x_{vi}} \left[1-(1-s_i)^{\xi_{vi}}g_i^{1-\xi_{vi}}\right]^{1-x_{vi}} \\ &\quad= \prod_v \prod_i \left[(1-s_i)^{x_{vi}}s_i^{1-x_{vi}}\right]^{\xi_{vi}} \left[g_i^{x_{vi}}(1-g_i)^{1-x_{vi}}\right]^{1-\xi_{vi}} \end{aligned} \tag{1.13}$$

X_{vi}	=	1 or 0	indicating whether or not student S_v performed task I_i correctly
$Q_{i\ell}$	=	1 or 0	indicating whether or not task I_i requires skill ℓ
$\alpha_{v\ell}$	=	1 or 0	indicating whether or not student S_v possesses skill ℓ
ξ_{vi}	=	$\prod_{\ell:\, Q_{i\ell}=1} \alpha_{v\ell}$	indicating whether or not student S_v has the skills needed for task I_i
s_i	=	$P[X_{vi} = 0 \vert \xi_{vi} = 1]$	a per-problem slip parameter
g_i	=	$P[X_{vi} = 1 \vert \xi_{vi} = 0]$	a per-problem guessing parameter

TABLE 1.2. Components of a simple cognitive assessment model for a stand-along test, with data collected at the task level (X_{vi}) rather than at the skill level ($\alpha_{v\ell}$).

To see how estimation of the parameters depends on the data, it is instructive to set up the first stages of an estimation algorithm for the model. In particular, let's assume that prior distributions have been chosen [for levels two and three in the hierarchy in Table 1.1], so that $s_i \sim \pi_s(s_i)$, $g_i \sim \pi_g(g_i)$, $\alpha_{v\ell} \sim \pi_\ell^{\alpha_{v\ell}}(1-\pi_\ell)^{1-\alpha_{v\ell}}$, and perhaps $\pi_\ell \sim \pi(\pi_\ell)$. These prior distributions will be used as "placeholders" in the notation below; their particular form will not affect our conclusions in any substantial way. We will calculate the so-called "complete conditional" distributions (e.g., Gelman, Carlin, Stern, & Rubin, 1995) of each parameter, given the data and the rest of the parameters. Such a calculation is directly useful in setting up an MCMC algorithm for Bayesian estimation of the model parameters (e.g., Patz & Junker, 1999a; 1999b), and is also useful in some versions of the E-M algorithm, such as ECME (e.g., Liu & Rubin, 1998). However, even when E-M is "possible" for a model like (1.13), it need not be practical, due to the need to sum over all configurations of the latent $\boldsymbol{\alpha}$ vector (see equation 1.14 below). For cases in which relatively few $\boldsymbol{\alpha}$'s carry most of the latent trait distribution, MCMC can be a more efficient—albeit approximate—way to estimate the model. This phenomenon has been found in other models with complex discrete latent structure as well (e.g., Seltman, 1999; Snijders & Nowicki, 1997; and Ter Hofstede, Steenkamp, & Wedel, 1999).

As usual in setting up an MCMC calculation, we note that the complete conditional distribution for each parameter or latent variable is proportional to the product of only those likelihood and prior factors in Table 1.1 depending on that parameter. Thus we obtain for example

$$\begin{aligned} p(s_i|\text{rest}) &\propto (1-s_i)^{\sum_v x_{vi}\xi_{vi}} s_i^{\sum_v (1-x_{vi})\xi_{vi}} \pi_s(s_i), \\ p(g_i|\text{rest}) &\propto g_i^{\sum_v x_{vi}(1-\xi_{vi})} (1-g_i)^{\sum_v (1-x_{vi})(1-\xi_{vi})} \pi_g(g_i), \end{aligned}$$

$i = 1, \ldots, k$, where "rest" stands for the data and the rest of the parameters in the model. From these it is easy to see the intuitively plausible facts that that the slip parameter s_i is sensitive only to successes and failures of those examinees S_v who we hypothesize (through the values of ξ_{vi} upon which we have conditioned) do have the requisite skills to perform task I_i, and similarly the guessing parameter g_i is sensitive only to successes and failures of examinees who we hypothesize do not have the requisite skills.

More central to the goal of assessing examinees, we see that the complete conditional distributions for $\alpha_{v\ell}$ are of the form:

$$\begin{aligned} p(\alpha_{v\ell}|\text{rest}) \\ \propto \prod_{i:\ Q_{i\ell}=1} \left[(1-s_i)^{x_{vi}} s_i^{1-x_{vi}}\right]^{\alpha_{v\ell}\xi_{vi}^{(-\ell)}} \left[g_i^{x_{vi}}(1-g_i)^{1-x_{vi}}\right]^{1-\alpha_{v\ell}\xi_{vi}^{(-\ell)}} \\ \times\ \pi_\ell^{\alpha_{v\ell}}(1-\pi_\ell)^{1-\alpha_{v\ell}}, \end{aligned}$$

where $\xi_{vi}^{(-\ell)} = \prod_{m\neq\ell:Q_{im}=1} \alpha_{vm}$, which indicates presence of all skills needed for task i, *except* for skill ℓ. From these complete conditionals we can see that when $\xi_{vi}^{(-\ell)} = 1$, the suggested conditional model for α_{vi} is some sort of Bernoulli, which makes sense. Also, when there are no tasks such that both $Q_{j\ell} = 1$ and $\xi_{vi}^{(-\ell)} = 1$, then $\alpha_{v\ell}$ is sensitive only to its prior distribution $\pi_\ell^{\alpha_{v\ell}}(1-\pi_\ell)^{1-\alpha_{v\ell}}$: no learning from data occurs. This observation is really a version of the credit/blame assignment problem (e.g., VanLehn & Niu, in press): we cannot infer whether $\alpha_{v\ell}$ was learned, if we are hypothesizing that another needed skill is still unlearned. Roughly, there must be information in the task performance data to allow us to assign credit (when a task is performed correctly) or blame (when it is performed incorrectly) to every cognitive attribute related to the task.

There are essentially two ways to avoid the credit/blame problem. In some situations, skills can be scored directly; this is possible for example within Anderson's ITS's for LISP, algebra and geometry, because students are required to successfully perform each subgoal/skill, with hints and repeated attempts if necessary, before moving on the next subgoal in the task (e.g. Embretson, 1997, pp. 309ff.). If one cannot score the tasks subgoal by subgoal, one can try to design the assessment (by hand or using methods from traditional statistical experimental design) so that the tasks, taken together, the task performance data informs us about each skill. VanLehn, Niu, Siler, and Gertner (1998) illustrate the inferential difficulties that can result when item sets are *not* constructed with the goal of designing around the credit/blame problem.

Finally if we want to estimate the skill base rates π_ℓ in the population (a measure of skill difficulty) we may include a fourth set of complete conditionals $p(\pi_\ell|\text{rest}) \propto \pi_\ell^{c_\ell}(1-\pi_\ell)^{N-c_\ell}\pi(\pi_\ell)$, where $c_\ell = \sum_v \alpha_{v\ell}$ is the number of students who are presently estimated to have skill ℓ.

4.2 A Role for Nonparametric IRT Methods?

Such models as (1.13) may seem very far removed from the setting in which nonparametric IRT methods are familiar. I want to suggest several ways in which they are not so far removed.

First, suppose that the skill variables $\alpha_{v\ell}$ vary independently of one another in a population of students or examinees, as would be consistent with e.g., Anderson's (1993) ACT-R theory, and suppose that the slip and guessing parameters s_i and g_i are fixed and satisfy the plausible inequality $(1 - s_i) \geq g_i$. A model for the task performance of a randomly-sampled examinee from the population would then be

$$P[\boldsymbol{X} = \boldsymbol{x}] \;=\; \sum_{\boldsymbol{\alpha}} \prod_i P_i(\xi_i(\boldsymbol{\alpha}))^{x_i} \left[1 - P_i(\xi_i(\boldsymbol{\alpha}))\right]^{1-x_i} \; p(\boldsymbol{\alpha}) \qquad (1.14)$$

where $p(\boldsymbol{\alpha})$ is a product measure over the space of binary skills $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$, and $P_i(\xi_i(\boldsymbol{\alpha}))$ is the monotone function

$$P_i(\xi_i(\boldsymbol{\alpha})) = (1 - s_i)^{\xi_i(\boldsymbol{\alpha})} g_i^{1-\xi_i(\boldsymbol{\alpha})}$$

of $\xi_i(\boldsymbol{\alpha})$; hence $P_i(\xi_i(\boldsymbol{\alpha}))$ is also monotone in the coordinates α_ℓ of $\boldsymbol{\alpha}$ (the index v has been dropped from $\alpha_{v\ell}$ to indicate that we are only working with the marginal model in what follows). Note the similarity of equation (1.14) to (1.2) and (1.3).

It follows from Lemma 2 of Holland and Rosenbaum (1986) that for any non-decreasing summary $g(\boldsymbol{X})$ of $\boldsymbol{X} = (X_1, \ldots, X_k)$, $E[g(\boldsymbol{X}) \mid \boldsymbol{\alpha}]$ is non-decreasing in each coordinate α_ℓ of $\boldsymbol{\alpha}$; this implies the SOM (Stochastic Ordering of the Manifest score X_+ by the latent trait) property of Hemker et al. (1997), namely that $P[X_+ > c \mid \boldsymbol{\alpha}]$ is non-decreasing in each coordinate α_ℓ of $\boldsymbol{\alpha}$. Not much is known about the inverse and more useful property SOL (see 1.7) when the latent "trait" is multidimensional. We might conjecture for example that $P\left[\ \alpha_\ell = 1 \ \middle|\ \sum_{i:Q_{i\ell}=1} \ X_i = s\ \right]$ would be non-decreasing in s. As in conventional dichotomous IRT models we hope that such a property is true, because it means that a most powerful test of mastery of skill ℓ can be based on a simple total of correctly-performed tasks involving skill ℓ.

It also follows from Theorem 8 of Holland and Rosenbaum (1986) that since $\boldsymbol{\alpha}$ is a collection of independent, or more generally, associated, random variables, the collection of response variables $\boldsymbol{X}$ is also associated, that is, for any two non-decreasing summaries $f(\boldsymbol{X})$ and $g(\boldsymbol{X})$, that $\mathrm{Cov}\,(f(\boldsymbol{X}),\, g(\boldsymbol{X})) \geq 0$. Obtaining CA (or a suitable generalization in case of multidimensional $\boldsymbol{\alpha}$) for $\boldsymbol{X}$ is a more difficult challenge.

An invariant item ordering property, such as $P_i(\xi_i(\boldsymbol{\alpha})) \leq P_j(\xi_j(\boldsymbol{\alpha}))$ uniformly in $\boldsymbol{\alpha}$, will follow if we assume for example that $1 - s_i$, g_i, *and* $\xi_i(\boldsymbol{\alpha})$, are comonotone as i varies, *for all* $\boldsymbol{\alpha}$. This is a kind of Guttman scaling condition on the latent responses ξ_i, that says for example that easier guessing is associated with lower skill requirements, and vice-versa. Thus, our cognitive assessment model (1.13) provides fertile ground for thinking about invariant item ordering, without necessarily being tied to preconceptions about continuous unidimensional IRT models.

More broadly, Hoijtink and Molenaar (1997) show how nonparametric model features such as conditional association (1.4) and manifest monotonicity (1.6) can be directly relevant to assessing model fit in a parametric Bayesian setting. Given a complete and interesting set of nonparametric model features for models like (1.13), a similar approach to model fit may be taken here.

Finally, among many who work in the nonparametric IRT traditions of Mokken, Molenaar, Sijtsma, Holland, Rosenbaum and their colleagues, it is the source of some bemusement that we work so hard to establish that scores such as X_+, which are perfectly good in the Rasch model, the most stringent of logistic IRT models, also suffice to order examinees, assess monotonicity properties of the underlying model from observable data, make mastery decisions, etc., in general nonparametric settings. Certainly comparisons between Rasch and Mokken scaling are not new (e.g., Meijer, Sijtsma, & Smid, 1990), and aspects of both Holland's (1990) "Dutch identity" work and Scheiblechner's (1995; see also Junker, 1998) "ISOP-model" work can be seen partly as attempts to formalize the connection between Rasch and nonparametric IRT methods.

I believe that the connection is actually rather simple, and is nearly obvious from Holland's (1990) work: The Rasch model is a very well-behaved exponential family model with immediately understandable sufficient statistics for items, given person parameters, and immediately understandable sufficient statistics for persons, given item parameters. Much of the work on monotone homogeneity models and their cousins is directed at understanding just how generally these understandable, but no longer formally sufficient, statistics yield sensible inferences about examinees and items. The model (1.13) provides us with a different "basic" model for cognitive assessment, in which parameters depend on immediately understandable summaries of the data, as illustrated by the complete conditional calculations above. We may ask how complex the relationship between the examinee skill parameters $\boldsymbol{\alpha}$ and the task performance data $\boldsymbol{X}$ can get, and still have these summaries be informative about guessing, slips presence or absence of skills, etc. We may also ask whether this is the "right" parametric model on which to base a nonparametric theory of cognitive assessment. For example, other possible models we might consider instead of (1.13) as a starting point for such a nonparametric theory include the constrained latent class model of Haertel (1989) and the hybrid model of Yamamoto and Gitomer (1993).

In addition to the intrinsic interest of this enterprise, the resulting nonparametric theory of cognitive assessment may have some practical utility. The relative ease with which the original Corbett, Anderson and O'Brien (1995) model can be estimated is a consequence of *both* its careful tailoring to the psychological theory, *and* the fact that data could be collected at the skill level (same granularity as the psychological theory), rather than at the task level, by the LISP tutor. But other assessment settings may not permit such tight binding of data collection design and psychological theory; and our discussion of the model (1.13) shows that even relatively minor modifications along these lines can make the inferential task more difficult. Estimation probably requires E-M, MCMC, or some other computationally intensive method, and great care must be taken so that every parameter is identifiable from the data. Models being proposed in the psychometric literature today to help unify the traditional IRT and discrete-cognitive-attributes approaches, such as the Unified Model of DiBello, Jiang and Stout (in press), also appear to require such treatment. In many cases, the estimation method, while feasible, may well be too slow for use in real-time feedback, as with computer based intelligent tutoring systems, and too complicated for teacher scoring of assessments embedded in instruction. A clear theory of which faster data summaries are relevant to the cognitive inferences we wish to make, over a wide variety of cognitive assessment models, would be an important contribution from the interface between nonparametric and parametric "IRT" models.

Acknowledgments: The author is on leave at Learning Research and Development Center, 3939 O'Hara Street, University of Pittsburgh, Pittsburgh PA 15260. Some of the work reported here was performed in the course of preparing a commissioned paper for the National Research Council Committee on Educational and Psychological Foundations of Assessment, United States National Academy of Sciences. Preparation of this chapter was supported in part by National Science Foundation grants #SES-99.07447 and #DMS-97.05032, and the hospitality of the Learning Research and Development Center. Thanks to Klaas Sijtsma, Lou Mariano, Matt Johnson, L. Andries van der Ark, and the editors of this volume for their comments and suggestions.

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

Algina, J. (1992). Special issue: the National Assessment of Educational Progress (Editor's Note). *Journal of Educational Measurement, 29,* 93–94.

Anderson, J. R. (ed.) (1993). *Rules of the mind.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: lessons learned. *The Journal of the Learning Sciences, 4,* 167-207.

Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology, 49,* 347–365.

Bartolucci, F., & Forcina, A. (in press). A likelihood ratio test for MTP_2 within binary variables. In press, *Annals of Statistics*.

Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice, 17,* 37–45.

Béguin, A.A., & Glas, C.A.W. (1998). *MCMC estimation of multidimensional IRT models*. (Research Report 98-14). Department of Educational Measurement and Data Analysis, University of Twente, the Netherlands. [Online]. Available: `http://to-www.edte.utwente.nl/TO/omd/report98/rr9814-.htm`. Retrieved 28 April 2000.

Bloom, B. S. (1984). The 2-sigma problem: the search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13,* 4–16.

Bock, R. D., & Zimowski, M. F. (1997). Multi-group IRT. In W. J. Van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York: Springer Verlag.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153–168.

Cliff, N., & Donoghue, J. R. (1992). Ordinal test fidelity estimated by an item sampling model. *Psychometrika, 57,* 217–236.

Corbett, A. T., Anderson, J. R., & O'Brien, A. T. (1995). Student modeling in the ACT programming tutor. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 19–41). Hillsdale, NJ: Lawrence Erlbaum Associates.

DiBello, L., Jiang, H., & Stout, W. F. (in press). A multidimensional IRT model for practical cognitive diagnosis. In press, *Applied Psychological Measurement.*

Douglas, J., & Qui, P. (1997). *Generalized linear factor analysis with Markov chain Monte Carlo.* Unpublished manuscript.

Draney, K. L., Pirolli, P., & Wilson, M. (1995). A measurement model for a complex cognitive skill. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 103–125). Hillsdale, NJ: Lawrence Erlbaum Associates.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19,* 143–165.

Ellis, J. L. (1994). *Foundations of monotone latent variable models.* Nijmegen: Nijmegen Institute for Cognition and Information.

Ellis, J. L., & Junker, B. W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika, 62,* 495–523.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56,* 495–515.

Embretson, S. E. (1995). Developments toward a cognitive design system for psychological tests. In D. Lubinski, & R. V. Dawis (Eds.), *Assessing individual differences in human behavior: new concepts, methods and findings* (pp 17–48). Palo Alto CA: Davies-Black Publishing.

Embretson, S. E. (1997). Multicomponent response models. In W. J. Van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305–321). New York: Springer Verlag.

Embretson, S. E. (1999). Generating items during testing: psychometric issues and models. *Psychometrika, 64,* 407–433.

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with many-faceted Rasch models. *Journal of Educational Measurement, 31,* 93–112.

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

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

Fox, G. J. A., & Glas, C. A . W. (1998). *A multi-level IRT model with measurement error in the predictor variables.* (Research Report 98-16). Department of Educational Measurement and Data Analysis, University of Twente, the Netherlands. [Online]. Available: `http://to-www.edte.utwente.nl/TO-/omd/report98/report98.htm`. Retrieved 28 April 2000.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research 23,* 267–269.

Gardner, H. (1992). Assessment in context: the alternative to educational testing. In B. R. Gifford, & M. C. O'Connor (Eds.), *Changing assessments: alternative views of aptitude, achievement, and instruction* (pp. 77–119). Norwell, MA: Kluwer Academic Publishers.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* New York: Chapman and Hall.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57,* 423–436.

Gibbons, R. D., & Hedeker, D. R. (1997). Random effects probit and logistic regression models for three-level data. *Biometrics, 53,* 1527–1537.

Glas, C. A. W., & Verhelst, N. D. (1989). Extensions of the partial credit model. *Psychometrika, 54,* 635–659.

Grayson, D. A. (1988). Two-group classification in latent trait theory: scores with monotone likelihood ratio. *Psychometrika, 53,* 383–392.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26,* 301–321.

Hemker, B. T., & Sijtsma K. (1999, July). *A comparison of three general types of unidimensional IRT models for polytomous items.* Paper presented at a Symposium on Nonparametric Item Response Theory Models in Action, European Meeting of the Psychometric Society in Lüneburg, Germany.

Hemker, B. T., Sijtsma, K., Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement, 19,* 337–352.

Hemker, B. T., Sijtsma K., Molenaar, I. W., & Junker, B. W. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika, 61,* 679–693.

Hemker, B. T., Sijtsma K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika,* 62, 331–347.

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika, 62,* 171–189.

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika, 46,* 79–92.

Holland, P. W. (1990). The Dutch identity: a new tool for the study of item response models. *Psychometrika, 55,* 5–18.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent trait models. *Annals of Statistics, 14,* 1523–1543.

Huguenard, B. R., Lerch, F. J., Junker, B. W., Patz, R. J., & Kass, R. E. (1997). Working memory failure in phone-based interaction. *ACM Transactions on Computer-Human Interaction, 4,* 67–102.

Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. *Psychometrika, 59,* 77–79.

Jaakkola, T. S., & Jordan, M. I. (1999). Bayesian parameter estimation via variational methods. In press, *Statistics and Computing*. [Online]. Available: `http://www.cs.berkeley.edu/~jordan/publications.html`. Retrieved 28 April 2000.

Janssen, R., & De Boeck, P. (1997). Psychometric modeling of componentially designed synonym tasks. *Applied Psychological Measurement, 21,* 37–50.

Johnson, E. G., Mislevy, R. J., & Thomas, N. (1994). Theoretical background and philosophy of NAEP scaling procedures. In E. G. Johnson, J. Mazzeo, & D. L. Kline (Eds.), *Technical Report of the NAEP 1992 Trial State Assessment*

Program in Reading (pp. 133–146). Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *Annals of Statistics, 21,* 1359–1378.

Junker, B. W. (1998). Some remarks on Scheiblechner's treatment of ISOP models. *Psychometrika, 63,* 73–85.

Junker, B. W., & Ellis, J. L. (1997). A characterization of monotone unidimensional latent variable models. *Annals of Statistics*, *25,* 1327–1343.

Junker, B. W., Koedinger, K. R., & Trottini, M. (2000, July). *Finding improvements in student models for intelligent tutoring systems via variable selection for a linear logistic test model.* Paper presented at the Annual North American Meeting of the Psychometric Society, Vancouver, BC, Canada.

Junker, B. W., & Patz, R. J. (1998, June). *The hierarchical rater model for rated test items.* Paper presented at the Annual North American Meeting of the Psychometric Society, Champaign-Urbana IL, USA.

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24,* 65–81.

Kass, R. E., Tierney, L., & Kadane, J. B. (1990). The validity of posterior expansions based on Laplace's method. In S. Geisser, J. S. Hodges, S. J. Press & A. Zellner (Eds.), *Bayesian and likelihood methods in statistics and econometrics: Essays in honor of George A. Barnard* (pp. 473–488). New York: North-Holland.

Kelderman, H., & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59,* 149–176.

Lee, Y., & Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B, 58,* 619–678.

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

Liu, C., & Rubin, D. B. (1998). Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data. *Statistica Sinica, 8,* 729–747.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs, No. 7.*

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin, 45,* 507–530.

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60,* 523–547.

Maris, E., De Boeck, P., & Van Mechelen, I. (1996). Probability matrix decomposition models. *Psychometrika, 61*, 7–29.

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

Meijer, R. R. (1996). Person-fit research: An introduction. (Guest editor's introduction to the Special Issue: Person-fit research: Theory and applications.) *Applied Measurement in Education, 9,* 3–8.

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement, 14,* 283–298.

Meredith, W. (1965). Some results based on a general stochastic model for mental tests. *Psychometrika, 30,* 419–440.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80,* 993–997.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33,* 379–416.

Mislevy, R. J., & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation *Psychometrika, 54,* 661–679.

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* The Hague: Mouton.

Mokken, R. J. (1997). Nonparametric models for dichotomous items. In W. J. Van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351–368). New York: Springer Verlag.

Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden, 37,* 97–117.

Molenaar, I. W. (1997). Nonparametric methods for polytomous responses. In W. J. Van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern psychometrics* (pp. 369–380). New York: Springer Verlag.

Molenaar, I. W., & Sijtsma, K. (1999). *MSP for Windows.* Groningen: iecProGAMMA.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19,* 73–90.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2nd Edition). New York: Chapman and Hall.

Nichols, P., & Sugrue, B. (1999). The lack of fidelity between cognitively complex constructs and conventional test development practice. *Educational Measure-*

ment: Issues and Practice, 18, 18–29.

Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24,* 146–178.

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24,* xxx–xxx.

Patz, R. J., Junker, B. W., & Johnson, M. S. (1999, April). *The hierarchical rater model for rated test items and its application to large-scale educational assessment data.* Paper presented at a Symposium on Making the Most of Constructed Responses: The Information in Multiple Ratings, Annual Meeting of the American Educational Research Association, Montreal Canada.

Patz, R. J., Junker, B. W., Lerch, F. J., & Huguenard, B. R. (1996). *Analyzing small psychological experiments with item response models* (CMU Statistics Department technical report #644). [Online]. Available: `http://www.stat-.cmu.edu/cmu-stats/tr/tr644/tr644.html`. Retrieved 28 April 2000.

Post, W. J. (1992). *Nonparametric unfolding models. A latent structure approach.* Leiden: DSWO Press, Leiden University, The Netherlands.

Post, W. J., & Snijders, T. A. B. (1993). Nonparametric unfolding models for dichotomous data. *Methodika, 7,* 130–156.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56,* 611–630.

Ramsay, J. O. (1995). A similarity-based smoothing approach to nondimensional item analysis. *Psychometrika, 60,* 323–339.

Ramsay, J. O. (1996). A geometrical approach to item response theory. *Behaviormetrika, 23,* 3–17.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401–412.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: new tools for educational reform. In B. R. Gifford, & M. C. O'Connor (Eds.), *Changing assessments: alternative views of aptitude, achievement, and instruction* (pp 37–75). Norwell, MA: Kluwer Academic Publishers.

Rigdon, S. E., & Tsutakawa, R. K. (1983). Parameter estimation in latent trait models. *Psychometrika, 48,* 567–574.

Robertson, T., Wright, F. T., & Dykstra, R. L. (1988) *Order restricted statistical inference.* New York: Wiley.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*, 425–435.

Rosenbaum, P. R. (1987a). Probability inequalities for latent scales. *British Journal of Mathematical and Statistical Psychology, 40,* 157–168.

Rosenbaum, P. R. (1987b). Comparing item characteristic curves. *Psychometrika, 52,* 217–233.

Roussos, L. (1994). Summary and review of cognitive diagnosis models. Unpublished manuscript.

Samejima, F. (1997). Departure from normal assumptions: a promise for future psychometrics with substantive mathematical modeling. *Psychometrika, 62,* 471–493.

Scheiblechner, H. (1972). Das Lernen und Lösen komplexer Denkaufgaben. [The learning and solving of complex reasoning items.] *Zeitschrift für Experimentelle und Angewandte Psychologie, 3,* 456–506.

Scheiblechner, H. (1995). Isotonic ordinal probabilistic models (ISOP). *Psychometrika, 60,* 281–304.

Seltman, H. (1999). *Hidden Stochastic Models for Biological Rhythm Data.* Unpublished Ph.D. Dissertation, Department of Statistics, Carnegie Mellon University, Pittsburgh PA, USA.

Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement, 22,* 3–32.

Sijtsma, K., & Van der Ark, L. A. (this volume). Progress in IRT analysis of polytomous item scores: dilemmas and practical solutions. In A. Boomsma, T. Snijders, & M. Van Duijn (Eds.), *Essays in Item Response Modeling* (pp. xxx–xxx). New York: Springer-Verlag.

Sijtsma, K., & Hemker, B. T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika, 63,* 183–200.

Sijtsma, K., & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology, 49,* 79–105.

Sijtsma, K., & Junker, B. W. (1997). Invariant item ordering of transitive reasoning tasks. In J. Rost, & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 97–107). Münster: Waxmann Verlag.

Shute, V. J., & Psotka, J. (1996). Intelligent tutoring systems: Past, Present and Future. In D. Jonassen (Ed.), *Handbook of Research on Educational Commu-*

nications and Technology (pp. 570–600). New York: Macmillan Press.

Snijders, T.A.B. (this volume). Two-level non-parametric scaling for dichotomous data. In A. Boomsma, T. Snijders, & M. Van Duijn (Eds.), *Essays in Item Response Modeling* (pp. xxx–xxx). New York: Springer-Verlag.

Snijders, T.A.B., & Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification, 14* 75–100.

Stegelmann, W. (1983). Expanding the Rasch model to a general model having more than one dimension. *Psychometrika, 48,* 259–267.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589–617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55,* 293–325.

Stout, W. F., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20,* 331–354.

Sykes, R. C., & Heidorn, M. (1999, April). *The assignment of raters to items: controlling for rater bias.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Tanner, M. A. (1996). *Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions.* 3rd Edition. New York: Springer-Verlag.

Tatsuoka, K. K. (1990). Toward an integration of item response theory and cognitive error diagnosis. In N. Fredriksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ter Hofstede, F. Steenkamp, J.-B. E. M., & Wedel, M. (1999). Identifying spatially contiguous international target markets. Manuscript submitted for publication.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567-577.

Van der Ark, L. A. (1999, July). *A reference card for the relations between IRT models for polytomous items and some relevant properties.* Paper presented at a Symposium on Nonparametric Item Response Theory Models in Action, European Meeting of the Psychometric Society in Lüneburg, Germany.

Van der Linden, W. J., & Hambleton, R. K. (eds.) (1997). *Handbook of modern item response theory.* New York: Springer Verlag.

VanLehn, K., & Niu, Z. (in press). Bayesian student modeling, user interfaces and feedback: A sensitivity analysis. In press, *International Journal of Artificial Intelligence in Education.*

VanLehn, K., Niu, Z., Siler, S., & Gertner, A. (1998). Student modeling from conventional test data: a Bayesian approach without priors. In B. P. Goetl, H. M. Halff, C. L. Redfield, V. J. Shute (Eds.), *Proceedings of the Intelligent Tutoring Systems Fourth International Conference, ITS 98* (pp. 434–443). Berlin: Springer-Verlag.

Verhelst, N. D., & Verstralen, H. H. F. M. (1993). A stochastic unfolding model derived from the partial credit model. *Kwantitatieve Methoden, 42,* 73–92.

Verhelst, N.D., & Verstralen, H.H.F.M. (this volume). IRT models for multiple raters. In A. Boomsma, T. Snijders, and M. Van Duijn (Eds.), *Essays in Item Response Modeling* (pp. xxx–xxx). New York: Springer-Verlag.

Wilson, M. R., & Hoskens, M. (1999, April). *The rater bundle model.* Paper presented at a Symposium on Making the Most of Constructed Responses: The Information in Multiple Ratings, Annual Meeting of the American Educational Research Association, Montreal Canada.

Wilson, D., Wood, R. L., & Gibbons, R. (1983). TESTFACT: test scoring and item factor analysis. [Computer program]. Chicago: Scientific Software Inc. Online description available: `http://ssicentral.com/irt/testfact.htm`. Retrieved 28 April 2000.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). *ConQuest: Generalized item response modeling software.* ACER.

Yamamoto, K., & Gitomer, D. H. (1993). Application of a HYBRID model to a test of cognitive skill representation. In N. Fredriksen & R. J. Mislevy (Eds.), *Test theory for a new generation of tests* (pp. 275–295). Hillsdale, NJ: Lawrence Erlbaum Associates.

Yuan, A., & Clarke, B. (1999). Manifest characterization and testing for two latent traits. Manuscript submitted for publication.

Zhang, J., & Stout, W. F. (1996, April). A theoretical index of dimensionality and its estimation. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1997). BILOG-MG. [Computer program]. Chicago: Scientific Software Inc. Online description available: `http://ssicentral.com/irt/bilogmg.htm`. Retrieved 28 April 2000.

Zwick, R. (1992). Special issue on the National Assessment of Educational Progress. *Journal of Educational Measurement, 17,* 93–94.