From psychometrics through item response theory
to complex statistical analyses

Brian W. Junker
Department of Statistics
Carnegie Mellon University
Pittsburgh PA 15213, USA
(412) 268-8873
brian@stat.cmu.edu

March 20, 1997

Paper to be presented at Invited Paper Meeting #69: An Overview of Contemporary
Psychometrics, 51st Session of the International Statistical Institute, 18–26 August
1997, Istanbul, Turkey.

1. Overview. In this survey, we review some statistical models that link traditional psychometrics, which mostly revolves around issues of quality of measurement, to statistical analysis of complex repeated measures data—especially in large-scale educational research—which usually revolves around estimation of effects and prediction. We focus on item response theory models, which are related to generalized linear models for multivariate discrete responses with random person effects. This is an extended version of the paper submitted to the ISI proceedings. Both versions are available from the author or at http://www.stat.cmu.edu/~brian/isi.html.

2. The problem. Many models and data analysis problems in psychometrics have the structure

\[ P[Y_{ij} = y] = \text{link}(y; \text{subject}_i, \text{probe}_j), \]

representing the probability of response \( y \) by subject \( i \) to probe \( j \), where the form of the data is an \( I \times J \) array \( \{y_{ij}: i = 1 \ldots I; j = 1 \ldots J\} \) for \( I \) subjects and \( J \) probes or test items (the two terms will be used interchangeably throughout). This two-way structure, with no replication for any \( y_{ij} \), arises in large scale, standardized national exams, in sociological and survey research, in classroom exams and small scale psychology experiments, and in psychiatry. Complete crossing of subjects with probes usually fails in large scale educational
surveys, where not every student answers every test item, and in experimental settings where the “same” probe may be modelled differently depending on whether the subject is in a treatment group or a control group.

2.1. True score and factor models. If the $y_{ij}$ are continuous, a familiar model for $\{y_{ij}\}$ is

$$Y_{ij} = \alpha_i + \beta_j + \varepsilon_{ij},$$

where $\varepsilon_{ij}$ are typically iid normal errors. If the $\alpha, \beta$ are fixed effects, we have the standard additive ANOVA model. If $\alpha$ is a random effect—a latent variable—then we get classical true score theory or errors in variables models (e.g. Lord and Novick, 1968). A fancier model along the same lines is the linear factor analysis model (e.g. Bartholomew, 1987),

$$y_{ij} = \sum_{k=1}^{d} \alpha_{jk} \theta_{ik} + \varepsilon_{ij},$$

where the so-called factor loadings $\alpha_{jk}$ are regression coefficients to be estimated, and $\theta = (\theta_{i1}, \ldots, \theta_{id})$ are $d$-dimensional random effects (latent variables, factors). In both true score and factor models the emphasis has been on using the model to assess (or improve) the quality of the data, or data-gathering (i.e., measurement) method, rather than to make external predictions or extrapolations. This goal is usually couched in the language of reliability (briefly, error-free reproducability of the data), validity (relation of the data to constructs intended to be measured), scaling (identifying groups of items that fit a particular model) and estimating the dimensionality, $d$, of $\theta$.

Most of the ideas of latent variable modeling were first developed under the linear factor model (which is itself a quantification of early twentieth century theories of intelligence, from Spearman, 1904, to Thurstone, 1947). In this view, one first estimates the covariance matrix $\Sigma = \text{Cov}(Y)$ from the rows $y_{i1}, \ldots, y_{id}$ of $\{y_{ij}\}$, and then tries to identify the matrix $A = [\alpha_{ijk}]$ of factor loadings from an eigenvalue/eigenvector (principal components) decompositions of $\Sigma$ and related matrices. Many traditional psychometric methods, such as optimal scaling, Guttman scoring, etc., are variations on principal or singular value decompositions of some Gramian (symmetric, positive-definite) matrix of similarity coefficients or category weights, that aid understanding of the pairwise association properties of $\{y_{ij}\}$; e.g. Basilevsky (1994). A rich and general theory of data analysis and description can be built on such a linear algebra foundation (Gifi, 1990), but this approach avoids probability models which can be helpful in framing more complex estimation and prediction questions. Some
recent attempts to unify the Gifi approach with the probability modeling approach that follows include Junker (1996) and Anderson and Walker (1997).

2.2. Item response models. If $y_{ij} \in \{0, 1\}$, item response theory (IRT) models specify $P[Y_{ij} = 1|\theta_i; \beta_j] = P_j(\theta_i)$, where $\beta_j = (\alpha_j, \delta_j, \ldots)$ are parameters affecting the shape of the response function $P_j(\theta)$ for each item or probe. For example, the Rasch (1960) model is equivalent to $\text{logit} P_j(\theta) \equiv \log[P_j(\theta)/(1 - P_j(\theta))] = \theta - \delta_j$. More generally, the models $\text{logit} P_j(\theta) = \sum_{k=1}^d \alpha_{jk} \theta_k - \delta_j$, analogous to the factor analysis model above, are called unidimensional ($d = 1$) or multidimensional ($d > 1$) “two-parameter logistic” (2PL; Lord, 1980; Reckase and McKinley, 1991) models. Especially when $d = 1$, the slope parameters $\alpha_{jk}$ determine the discrimination among subjects at various $\theta$ levels by the $j^{th}$ item, and $\delta_j$ determines the item’s overall difficulty. The “three parameter logistic” (3PL; Lord, 1980) model adds a nonzero lower asymptote to $P_j(\theta)$ to model random response behavior by low-$\theta$ subjects. Modifications of these models to handle multicategory responses include Bock’s (1972) nominal model, Masters’ (1982) partial credit model, and Samejima’s (1969) graded response model (see also the recent surveys by Fischer and Molenaar, 1995; and van der Linden and Hambleton, 1997). Each of these has fixed-effects analogues that would be familiar to students of polytomous logistic regression.

3. Nonparametric IRT methods. Three properties underlying most psychometric models can be illustrated with the examples above: local independence (LI): $P(y_{i1}, \ldots, y_{iJ}|\theta_i; \beta_1, \ldots, \beta_J) \equiv \prod_j P(y_{ij}|\theta_i; \beta_j)$; monotonicity (M): usually, responses $y_{ij}$ are stochastically ordered by $\theta_i$; and unidimensionality (U): $d = 1$. Nonparametric IRT is the study of classes of IRT models in which $P_j(\theta)$ is left unspecified but LI, M and U (or at least $d$ very much smaller than $J$) are assumed to hold. Since data can only be observed from the marginal distribution $p(y_{i1}, \ldots, y_{iJ}) = \int P(y_{i1}, \ldots, y_{iJ}|\theta, \beta_1, \ldots, \beta_J) \pi(\theta|\lambda) d\theta$ [where $\pi(\theta|\lambda)$ is a population distribution for $\theta$ with parameters $\lambda$], there is an obvious indeterminacy in specifying $\theta$, and the nonparametric approach tries to sidestep this indeterminacy by focusing on observable consequences of LI, M and U. This approach provides a bridge between the classical, “measurement quality” goals of psychometrics, and the estimation and prediction goals of more complex analyses.
3.1. Nonparametric model analysis. At a very general level, one may ask what testable, data analytic features a particular class of models has. For example Junker and Ellis (1997) extend the work of Holland and Rosenbaum (1986) and others, to show that two simple-to-state covariance conditions essentially characterize distributions on \( \{y_{ij}\} \) that can be represented by LI, M, U models. Alternatively one can identify desirable model properties—e.g. for inferences on \( \theta \) one might wish that \( P[\theta_i > c | \sum_j y_{ij} = s] \) increases with \( s \)—and try to develop a large class of models that have these properties. Hemker et al. (1996, 1997) explore exactly this stochastic ordering property: it is well known to hold for any IRT model for dichotomous data satisfying LI, M and U (e.g. Grayson, 1988; Huynh, 1994), but it greatly restricts the variety of LI, M, U models available for ordered polytomous responses. In a similar vein, Bartholomew (1987) is lead to standard factor analysis models by a desire for simple sufficient statistics, and Goldstein (1980) is lead to a log-log response curve for binary data (instead of the conventional logistic response curve) by symmetry considerations. More recently, Love (1997) develops nonparametric criteria for constructing distractors for multiple-choice test items.

3.2. Nonparametric data analysis. There is a long history of parametric and nonparametric methods for scaling and dimension counting; only a few recent developments are mentioned here. Inspired by ideas of Humphreys (1982, 1985) and McDonald (1981), Stout and his students and colleagues (e.g. Nandakumar, 1994; Roussos, Stout, and Marden, 1994; Stout, 1987; 1990) have developed statistical theory and software for identifying approximately unidimensional groups of items. The main idea is to replace LI with an essential independence (EI) condition that requires the average covariance between items, given \( \theta \), to become vanishingly small as \( J \) grows, and directly test for EI. Mokken scaling (e.g., Mokken and Lewis, 1982; Molenaar et al., 1994) identifies unidimensional scales, by finding large marginal covariances between items. Both methods are related to Cronbach’s (1951) \( \alpha \) bound on reliability. Mokken scaling is very good at finding groups of items that are highly discriminating among subjects; Stout’s methods tend to better identify groups of items that narrowly satisfy LI/EI, M and U. Sijtsma and Junker (1996) and Sijtsma and Hemker (1997) survey methods for identifying groups of items that exhibit invariant item
ordering: \( P_{j_1}(\theta) \leq P_{j_2}(\theta) \) for all \( \theta \). Nonparametric, ordinal inference on \( \theta \) has been pursued by Cliff (1989; 1994; Cliff and Donoghue, 1992). Ramsay (e.g. Ramsay, 1991; 1995; 1996) has developed computational tools for estimating the \( P_j(\theta) \) nonparametrically, and for visualizing the dimensionality of \( \theta \) in terms of the surface in the \( J \)-dimensional unit cube generated by the likelihood \( P(y_{i1}, \ldots, y_{iJ} | \theta_i; \beta_1, \ldots, \beta_J) \) as \( \theta_i \) varies. This geometric approach makes possible a rigorous study of the change of dimensionality of a test over its total score range, as suggested by Reckase (1990).

4. Parametric IRT methods. Most modern parametric methods (cf. Holland, 1990b) impose the assumptions LI, M and U on the joint likelihood \( \prod_{i=1}^{I} P(y_{i1}, \ldots, y_{iJ} | \theta_i; \beta_1, \ldots, \beta_J) \). These models typically contain on the order of \( I + J \) parameters, with \( J \) considered fixed. Baker (1992), Bartholomew (1987), Fischer and Molenaar (1995), and van der Linden and Hambleton (1997) provide excellent reviews of the models and estimation techniques in use up to about 1995. Inconsistency problems as \( I \) grows (Neyman and Scott, 1948) are usually circumvented by conditioning or by averaging over the subject parameters.

For the Rasch model and its exponential family generalizations, conditional maximum likelihood (Andersen, 1972)—conditioning on row totals in \( \{y_{ij}\} \) to get rid of the \( \theta \)'s while making inferences on the \( \beta \)'s (or vice-versa)—has a long and successful tradition. For non-Rasch models, Bock and Aitkin (1981) developed an E-M (expectation-maximization; Dempster, Laird and Rubin, 1977) style algorithm they call marginal maximum likelihood for estimating the item parameters \( \beta \) and the population parameters \( \lambda \) by maximizing the marginal likelihood \( p(y_{i1}, \ldots, y_{iJ} | \beta_1, \ldots, \beta_J, \lambda) \) [with \( \theta \) integrated out as in Section 3]; inference on \( \theta \) then proceeds via Empirical Bayes. Research from Deleeuw and Verhelst (1986) through Lindsay, Clogg and Grego (1991) and Pfanzagl (1993) establishes the asymptotic equivalence of the conditional and marginal likelihood approaches, when both apply, and work from Tjur (1982) through Holland (1990a) shows the the equivalence of a semiparametric marginal likelihood for the Rasch model to a quasi-symmetric log-linear model. Joint asymptotics in \( I \) and \( J \) under which all parameters may be consistently estimated are delicate (Haberman, 1977), but recently Douglas (1997) has developed a general method justifying joint
maximum likelihood.

In the biostatistics literature Stiratelli, Laird and Ware (1984) adapt the method of Laird and Ware (1982) to develop E-M machinery similar to Bock and Aitkin’s for random effects models for longitudinal binary data. Agresti (1993) suggests application of the log-linear approach to the Rasch model in medical applications; Another method popular in biostatistics, generalized estimating equations (e.g. Liang and Zeger, 1986; Zeger and Liang, 1992), has not been as popular in psychometrics, since it doesn’t lead naturally to inferences on individuals’ $\theta_i$’s.

The recent development of Markov Chain Monte Carlo (MCMC) integration methods (Chib and Greenberg, 1995; Tanner, 1996; Tierney, 1994) provides an alternative to the E-M approach. Albert (1992) and Johnson 1996, 1997) show that data augmentation with Gibbs sampling is very effective for IRT style models; see also Bradlow and Zaslavsky (1997). Related methods are explored by Hoijtink and Molenaar (1997) for ordered latent class models, by Scheines, Hoijtink and Boomsma (1996) for structural equations models, and by Seltzer (1993) and Seltzer, Wong and Bryk (1996) for hierarchical linear models (HLM’s; Bryk and Roudenbush, 1992). Patz (1996) has developed a general methodology that replaces data augmentation with Metropolis-Hastings sampling; Douglas and Qui (1997) have developed a similar methodology for linear and nonlinear exploratory factor analysis models.

5. Complex applications of IRT. In recent years there has been an explosive growth in the application of IRT models, both within and outside of psychometrics. We indicate only a few of the more exciting recent developments.

In educational measurement, parametric and nonparametric IRT methods are routinely used to assess the quality of individual items and sets of items in educational measurement work (Linn, 1989), and to elucidate experimental design and statistical estimation issues when performances of examinees who sit for different versions of the same exam must be compared (Holland and Rubin, 1982; or more recently Little and Rubin, 1994); a similar problem arises in the scoring of adaptive, sequentially-designed examinations on the computer (e.g. Chang and Ying, 1996). IRT modeling has also elucidated research into the assessment of sociological bias on standardized test
items (Holland and Wainer, 1993; or more recently Bolt and Stout, 1996), and the detection and diagnosis of subject outliers (Levine and Drasgow, 1988; Molenaar and Hoijtink, 1990). Related work includes the study of rater effects (e.g. Engelhard, 1992; Johnson, 1996; Patz, 1996) and a modern approach to adjusting college GPA for the difficulty of courses taken (Johnson, 1997).

In the social sciences and related areas IRT has also played a substantial role. The Rasch model has long been applied in social survey work; three representative examples include Conaway (1992), Duncan (1985) and Thissen and Mooney (1989). Applications in multiple-recapture census work extend back at least to Sanathanan, 1973; recently Darroch et al. (1993) compare Rasch modeling of capture-recapture data with standard log-linear models. Analysis of the quality of items and scales with IRT has become more visible in psychiatry, from Gibbons et al. (1985) to Santor et al. (1995) and in epidemiology (e.g. Reiser, 1989; Darroch and McCloud, 1990). IRT methods are also applied in the analysis of item-level data in case-control experiments in medicine and psychology.


More complex applications have had to wait for the evolution of E-M and MCMC methods indicated in Section 4. These methods have made it feasible to incorporate covariates and other structure into IRT models, either through the item-specific parameters $\beta_j$, similarly to generalized linear models; or hierarchically through the subject-specific parameters $\theta$. The estimation of group effects and the use of examinee covariates in estimating item parameters (e.g. Mislevy, 1985; Mislevy and Sheehan, 1989) plays a large role in the analysis of large multisite educational assessments in the USA, such as the National Assessment of Educational Progress (NAEP; e.g., Algina, 1992; Johnson, Mislevy and Thomas, 1994; and Zwick 1992), which are much like multicenter clinical trials in biostatistics (e.g. Gray, 1994; see also Mislevy, 1991; and the references in Seltzer, Wong
and Bryk, 1996). The novelty of the NAEP analyses has been the departure from working with summary measures such as total scores on each individual (student or patient) studied—common in both educational and biostatistics examples—in favor of models that at least conceptually combine IRT and nonlinear factor analysis models at the “data level” with hierarchical structure on $\theta$ to reflect the multisite/multicenter nesting of observations, as in HLM’s (Patz, 1996). Indeed, Hemker et al. (1997) and Johnson (1997) have shown that total scores do not contain the same information, even ordinally, as underlying $\theta$ scores, when $J$ is small. In the coming years we can expect to see explosive growth in parametric models for multisite/multicenter studies that more fully marry the IRT and HLM approaches.
Bibliography.


Psychometrika, 57, 217–236.


Nandakumar, R. (1994). Assessing the dimensionality of a set of item responses—comparison of


