PSYCHOMETRICS: FROM PRACTICE TO THEORY AND BACK

15 Years of Nonparametric Multidimensional IRT, DIF/Test Equity, and Skills Diagnostic Assessment

William Stout

DEPARTMENT OF STATISTICS, UNIVERSITY OF ILLINOIS
AND
EDUCATIONAL TESTING SERVICE

This paper was originally published in the December 2002 issue of Psychometrika (Volume 67, Number 4). It is republished here to demonstrate the appearance of a paper that is prepared using the Pmetrika LaTeX Style File Package for Psychometrika Authors (Version 2b1) published by the Psychometric Society. A publication quality copy of William Stout's article can be obtained online at http://www.psychometricsociety.org /ARTICLEstout2002.pdf. The Pmetrika LaTeX Style File Package for Psychometrika Authors was designed and prepared by Tim Null, and it was based on original work done by Don Deland of Integre Technical Publishing. The package can be downloaded at http://www.psychometricsociety.org. If you have questions about the package you can contact Tim Null at tim@timnull.com.

This aricle is based on the Presidential Address William Stout gave on June 23, 2002 at the 67th Annual Meeting of the Psychometric Society held in Chapel Hill, North Carolina. —Editor

I wish to especially thank Sarah Hartz and Louis Roussos for their suggestions that helped shape this paper. I wish to thank all my former Ph.D. students: Without their contributions, the content of this paper would have been vastly different and much less interesting!

Requests for reprints should be sent to William Stout, Department of Statistics, University of Illinois, 725 S. Wright Street, Champaign IL 61820. E-Mail: stout@stat.uiuc.edu

Dedication: I want to dedicate this paper to my wife, Barbara Meihoefer, who was lost to illness in this year of my presidency. For, in addition to all the wonderful things she meant to me personally and the enormous support she gave concerning my career, she truly enjoyed and greatly appreciated my psychometric colleagues and indeed found psychometrics an important and fascinating intellectual endeavor, in particular finding the skills diagnosis area exciting and important: She often took time from her career as a business manager and entrepreneur to attend psychometric meetings with me and to discuss research projects with my colleagues and me. She would have enjoyed this paper.—William Stout

Abstract

The paper surveys 15 years of progress in three psychometric research areas: latent dimensionality structure, test fairness, and skills diagnosis of educational tests. It is proposed that one effective model for selecting and carrying out research is to chose one's research questions from practical challenges facing educational testing, then bring to bear sophisticated probability modeling and statistical analyses to solve these questions, and finally to make effectiveness of the research answers in meeting the educational testing challenges be the ultimate criterion for judging the value of the research. The problem-solving power and the joy of working with a dedicated, focused, and collegial group of colleagues is emphasized. Finally, it is suggested that the summative assessment testing paradigm that has driven test measurement research for over half a century is giving way to a new paradigm that in addition embraces skills level formative assessment, opening up a plethora of challenging, exciting, and societally important research problems for psychometricians.

Key words: nonparametric IRT, NIRT, latent unidimensionality, latent multidimensionality, essential unidimensionality, monotone locally independent unidimensional IRT model, MLI1, item pair conditional covariances, DIMTEST, HCA/CCPROX, DETECT, CONCOV, Mokken scaling, generalized compensatory model, approximate simple structure, DIF, differential item functioning, differential bundle functioning DBF, valid subtest, multidimensional model for DIF, MMD, SIBTEST, MultiSIB, Mantel-Haenszel, PolySIB, CrossingSIB, skills diagnosis, formative assessment, Unified Model, reparameterized Bayes Unified Model, MCMC, evidence centered design, ECD, PSAT Score Report Plus.

1. Introduction

As previous presidents of the Psychometric Society have emphasized, a presidential address is an occasion where one has a *duty* to present one's personal perspective about the past and future of psychometrics. In this tradition I participate enthusiastically. This paper is about my efforts and those of a large group of colleagues to individually and collectively address three core issues confronting educational and psychological testing. Most of these colleagues got their Ph.D.'s under my direction and, as such, have been members of the Statistical Laboratory for Educational and Psychological Measurement (the "Lab"), which I founded in the Department of Statistics at the University of Illinois in the late 1980s and which is currently co-directed by Jeff Douglas and Louis Roussos, former Ph.D. student Lab members and now faculty members at the University of Illinois.

During the first half of my research career, spent as a mathematics professor at the University of Illinois, I solved mathematics problems, motivated by their intrinsic intellectual appeal. In short, I was acting as artist rather than engineer. This pure-research-driven aesthetic motivation, which works so well for many, had in fact been increasingly sapping my research drive. My response was to switch fields from pure mathematics (probability theory, actually) to psychometrics. I was determined to do research that, in addition to being intrinsically intellectually interesting, would bear fruit that could be effectively *applied* to important societal problems.

Thus, my personal, belated, and enthusiastic entry into the field of psychometrics in the 1980s, after I had already become a full professor of mathematics, was my personal resolution of my professional "midlife crisis." In resolving this crisis, I discovered two important things about the process of doing psychometric research that are worth stressing. First, issues and problems arising out of the *actual practice* of educational testing provide fertile ground for the generation of excellent research problems, for which a deeply theoretical approach often produces effective and important solutions from the applications perspective. Second, enormous excitement and intellectual power can flow forth when a carefully selected group of researchers dedicates itself to collaboratively solving important and interesting psychometric research problems. In fact, after switching fields I was determined to intensify the collaborative research style I had learned while doing mathematical research, especially while working jointly and extensively with my mathematical colleague and friend Walter Philipp (see, e.g., Philipp & Stout, 1975).

Psychometrics at its best is a field that is both intellectually interesting and deeply relevant for a global society in which effective education and training are central prerequisites to progress. Indeed, psychometrics' intellectual appeal and great societal importance were what most motivated my choosing it over other applied fields like biostatistics. Interestingly, my aggressive shifting of research focus is strikingly congruent with Robert Sternberg's (1985; see especially pp. 50–51) discussion on the role of "selection" in his contextual theory of intelligence. Sternberg speaks of finding what one is "interested in," and pursuing it "relentlessly."

I decided the best way to maximize the likelihood that the Lab's psychometric research would be usefully applied to societally important "real-world" measurement problems (my personal goal) was to choose research problems wisely from the vast hurly-burly of practical problems and issues flowing out of actual measurement practice as conducted by testing

companies and other practitioners of the testing art. From this applications perspective, "choosing wisely" meant choosing research problems for which successful solutions would clearly have a widespread impact on bringing about improved educational testing and assessment.

In the case of the Lab, this principle of choosing wisely translated into the Lab focusing, over the decade and a half of its existence, on three core applied issues growing out of the practice of standardized testing: assessing the multidimensional structure of the latent ability space that stochastically drives test performance, assessing test fairness, and diagnosing examinee skills as a means of accomplishing formative assessment. By formative assessment, I mean the assessment of students while they are still learning, with the purpose of facilitating both teaching and learning.

In order to be effective at carrying out the research goals stated above and to provide a quality environment for Ph.D. study and research, from its inception the Lab has been managed according to certain principles: At any given time, it always had several Ph.D. students with research assistantship support. Funding from major testing companies was eagerly sought, as a source not merely of funding, but also of applied problems whose solution would be important. Students were encouraged to work cooperatively with each other and with outside researchers (including former Lab members). The Lab was, and is, viewed as part of the wider research community. It has always been organized around well-defined and evolving research goals.

Because my field of research had been probability theory, my tendency has always been to stress the importance of sophisticated probability modeling in addressing psychometric research. Thus, in each of the three identified research areas—latent structure, fairness, and skills diagnosis—the challenge of developing appropriate probability models, as well as theoretically deriving important implications of these developed models, became very important. Because I reside in a statistics department, another tendency has been to stress the importance of bringing modern statistical thought and methodologies to bear on psychometric research problems. This has also had a major influence on our progress in the three research areas and in addition has from time to time drawn research statisticians to the field of psychometrics. In this regard, I strongly believe, and have seen first hand, that psychometrics can powerfully appeal to research statisticians casting about for interesting areas of research and applications to ply their trade.

Subsequent to probability modeling and associated foundational analyses, the development of specialized and sometimes innovative statistical procedures, influenced by the statistics environment I reside in as just remarked, was required. Finally (the "back again" of the title), it was always judged vital to close the loop back to the applied setting that had precipitated the research problem by providing a practical and easy-to-implement solution to the practitioner's problem. That is, the goal has never been just the published papers we all cherish; rather, it has been to produce deeply successful applications played out in actual educational measurement practice. Moreover, if the research is to have an important impact on educational measurement, it must have easy and wide transferability to similar measurement settings. This transferability has been facilitated by making methodological software available to practitioners and researchers, as discussed below.

The remainder of this paper describes progress in the three applied research areas the Lab has focused on, with occasional suggested directions for future research given.

2. Nonparametric Latent Structure Assessment

We first consider the assessment of latent structure unidimensionality. Then more generally we consider the assessment of multidimensional latent structure.

2.1. Unidimensionality from the Weak LI Conditional Covariance Perspective

In his influential monograph *Applications of Item Response Theory to Practical Testing Problems* (Lord, 1980), Fred Lord stated, "There is a great need for a statistical significance test for the unidimensionality of a set of items." This strong statement, made when practical applications of unidimensional IRT modeling to testing was in its relative infancy, reminded the testing community of the great need to have a reliable statistical test of unidimensionality. Hypothesis test acceptance of unidimensionality would help to legitimize applying unidimensional logistic IRT-based calibration and prediction methodologies such as LOGIST and BILOG. Further, one important way to address the issue of whether it is appropriate to summarize examinee test performance with a single scale is by asking the psychometric question of whether the test data is unidimensional. Thus, the Lab's 15-year odyssey into nonparametric latent structure assessment began by addressing the question of how to statistically assess departures from latent unidimensionality.

The unidimensionality question is a prime example of sound theory being a prerequisite for good psychometric practice. My approach was both simple and nonparametric (Stout, 1987). By taking a nonparametric approach, one does not confound lack of model fit by a particular unidimensional parametric family of models with the data having been generated by an intrinsically multidimensional latent-trait model.

Conceptualizing latent unidimensionality, as must be done, requires one to step back and ask the foundational question of how best to define latent unidimensionality. The primary insight is that the issue of unidimensionality must be framed as whether the inferred manifest test distribution can be represented as a unidimensional, locally independent, monotone latent trait model. (In principle, by observing enough examinees, the manifest distribution can be statistically inferred from test data with any desired accuracy and hence becomes "manifest." In this sense, the unidimensionality problem is posed by presuming that a specific manifest test distribution has been specified.) It is interesting and important to note that being able to exhibit a two-dimensional locally independent, monotone latent-trait model that fits the given manifest distribution *does not* prove that unidimensionality fails for the given manifest distribution. By contrast, exhibiting a parametric unidimensional model fitting the manifest distribution does prove unidimensionality holds for the manifest distribution. In summary, the issue of unidimensionality is not addressed by showing the lack of fit of a particular unidimensional parametric family of IRT models to the given manifest distribution. Rather, the issue is whether any unidimensional IRT model *exists* that fits the manifest test distribution.

For simplicity, throughout the discussion of dimensionality we assume dichotomous item responses, although most of the procedures described and their associated theoretical underpinnings have polytomous versions too. Even though it is the test that is said to be unidimensional, as MIRT (M=multidimensional) developers Mark Reckase and Terry Ackerman

have stressed, the dimensionality lies in the interaction between the test structure (given by the item response functions; i.e., IRFs) and the latent ability structure (given by the latent ability examinee population distribution). In fact, the same test could be unidimensional for one examinee population and not for another.

We now define test unidimensionality. To avoid unneeded probabilistic and notational complexity, we make the easily removable assumption that all latent variables are random variables of continuous type.

Definition 1. A test $\mathbf{U}' = (U_1, U_2, \ldots, U_n)$ with specified manifest distribution $\mathrm{P}(\mathbf{U} = \mathbf{u})$ is said to be *unidimensional* if there exists a unidimensional random variable Θ with density denoted by $f(\theta)$ such that for all possible response patterns $\mathbf{u}$,

$$\mathrm{P}(\mathbf{U} = \mathbf{u}) = \int_{-\infty}^{\infty} \mathrm{P}(\mathbf{U} = \mathbf{u} | \Theta = \theta) f(\theta)\, d\theta \tag{1}$$

for which local independence (LI) and IRF monotonicity (M) relative to θ holds. Any IRT model satisfying LI and M for a unidimensional latent trait Θ is called a *monotone locally independent unidimensional model* and herein is denoted by MLI1.

The widely investigated MLI1 model goes back in IRT research at least to Mokken (1971). What is being denoted in this paper as a MLI1 model has many names in the literature including "monotone unidimensional latent trait model," "monotone homogeneity model," "monotone latent variable model," and "monotone IRT model." Thus, care is required in reading and interpreting the literature concerning the concept of a unidimensional latent model.

It is worth noting that if one drops M as a requirement, the conceptual idea behind the fact that one can always exhibit a unidimensional LI latent model for any given manifest test distribution can be easily explained, this result shown by Suppes and Zanotti (1981): Adopting Paul Holland's deterministically-responding but random-sampled examinee perspective (Holland, 1990a), the basic idea of the proof of unidimensionality is to assign to each examinee, as the value of his or her unidimensional latent variable, the binary expansion corresponding to his or her true deterministic knowledge state for all of the items (1 in i-th place of the binary expansion if item i answer "known" by the examinee, 0 if not known). The probability assigned to each such binary-expansion-represented latent state is the probability assigned by the manifest test distribution to the corresponding examinee item response pattern producing the "latent" binary expansion. It is then easy to see that LI holds and M fails for the resulting unidimensional model.

Generalizations of Definition 1 are possible by sensible weakening of either the concept of local independence or of monotonicity. One weakening, important from the practitioner perspective, is to replace the traditional definition of local independence with pairwise local independence or what is termed weak local independence (see McDonald, 1994).

Definition 2. A test $\mathbf{U}$ is said to be *(strongly) locally independent* (denoted LI, or SLI when it is necessary to contrast LI of this definition with weak LI of Definition 3 below) with respect to a

latent variableΘ if for all $\mathbf{u}$ and θ,

$$\mathrm{P}(\mathbf{U}=\mathbf{u}|\Theta=\theta) \;=\; \prod_{i=1}^{n}\mathrm{P}(U_i=u_i|\Theta=\theta). \tag{2}$$

Definition 3. A test is said to be *weakly locally independent* (WLI) with respect to a latent variable Θ if for all item pairs $i, i'$and all θ,

$$\mathrm{Cov}(U_i, U_i'|\Theta=\theta)=0. \tag{3}$$

In the dichotomous item-scoring case, it is trivial that WLI is equivalent to pairwise LI. That is, WLI holds if and only if pairwise LI holds for all item pairs i, i'. Namely, for all $\boldsymbol{\theta}$ WLI is equivalent to

$$\mathrm{P}(U_i=u_i, U_i'=u_i'|\Theta=\theta) \;=\; \mathrm{P}(U_i=u_i|\Theta=\theta)P(U_i'=u_i'|\Theta=\theta). \tag{4}$$

Clearly, from the practitioner's perspective a test could be viewed as unidimensional, or a latent model viewed as MLI1, if Definition 1 holds with local independence replaced by weak local independence. The careful reader will note that by making this replacement an empirical assumption has (sneakily) inserted itself: If a test is declared unidimensional according to Definition 1 with WLI (equivalently, pairwise LI) used as the definition of local independence, then such a declaration presumes it is also unidimensional (it supports a MLI1 IRT model) with SLI replacing pairwise LI. I agree with the assumption from the empirical viewpoint: From the measurement practitioner's perspective, WLI seems for practical purposes empirically equivalent (trivially, WLI and SLI are not mathematically equivalent) to the more complex (and much harder to verify statistically) reality of SLI.

In this regard it is useful to quote McDonald (1994), as he argues that this insertion of WLI in place of SLI amounts to "not changing our definition or our substantive conceptualization of latent traits" (p. 67). McDonald goes on to state that it is

> unlikely that investigators seriously imagine that the conditional covariances of the items vanish while they still possess higher order dependencies in probability. That is, we are unlikely to suppose that while every pair of items gives statistically independent responses [conditional on a latent variable], responses to some items [conditional on the latent variable] are dependent on responses to two or more other items. (p. 67)

Our acceptance of this practical equivalence of WLI and SLI undergirds our statistical approach to latent dimensionality assessment. That is, in order to study and statistically assess latent dimensional structure, we statistically investigate conditional covariances given an appropriately chosen latent trait θ, thereby investigating WLI instead of investigating the full SLI.

Thus, the practitioner-driven need for a useful statistical test of the unidimensionality of an educational test led me to formulate a theory of latent space dimensionality structure assessment based on conditional covariances. This, as will be discussed below, in turn led us to theoretically-defensible statistical procedures such as DIMTEST, DETECT, and HCA/CCPROX,

which are justified empirically via large-scale simulations and real-data-based standardized test applications.

In addition to a focus on conditional covariances as a basic dimensionality assessment tool, our efforts had refocused us on the importance of the fundamental MLI1 model that underlies the recent emphasis on foundational and applied nonparametric item response theory (NIRT) research on both sides of the Atlantic Ocean, with much cross-Atlantic cooperation (see especially Junker & Sijtsma, 2001). Indeed, our conditional covariance approach to latent dimensionality assessment has meshed nicely with this resurgence of interest in NIRT, propelled in part on the European side of the Atlantic by a long-standing focus on Mokken. The interested reader is in particular referred to the September 2001 *Applied Psychological Measurement* "Special Issue on NIRT"—edited by Brian Junker and Klaas Sijtsma—for a superb survey of modern NIRT research, and to a slightly earlier methodological NIRT survey paper by Sijtsma (1998).

2.2. Foundational Issues Facilitated by Infinite Test Length Unidimensional MLI1 Modeling

A flurry of foundational Lab-based work on NIRT emanated from the original DIMTEST paper (Stout, 1987), all of this work aimed at developing a clear foundational understanding of multidimensional latent trait modeling, especially as based on conditional covariances. Much of this work then formed the basis for further conditional-covariance-based statistical tools developed for practitioners in order to estimate important characteristics of a multidimensional latent space, especially HCA/CCPROX and DETECT.

Motivated by ideas of Lloyd Humphreys (expressed in, but transcending, the metaphor of classical factor analysis), early on I had been struck by the notion that unidimensionality, with either WLI or SLI as its underpinning, was too stringent a concept from the practical perspective of wanting to model and count only the important latent dimensions and in particular wanting to theoretically characterize a test consisting of only one important dimension. In particular, from the practitioner perspective, it is useful to dichotomize all the latent dimensions into the important (called "essential," "dominant," "major," etc.) dimensions and the unimportant (called "inessential," "weak," "nuisance," "minor," etc.) dimensions.

Thus, there was a need for a theoretical conception that would appropriately separate essential from inessential dimensions, counting only the number of essential dimensions and, in particular, defining in what sense a test can have just one "essential" dimension with possibly numerous inessential dimensions. This idea certainly has factor analytic roots: see, for example, Tucker, Koopman, and Linn (1969) for a factor analytic model distinguishing between minor (and hence inessential) factors and major factors. The resulting formal definition of essential unidimensionality (Stout, 1990) with respect to a latent variable θ uses conditional covariances and is based on the notion of modeling the dimensionality properties of a finite-length test $\mathbf{U}'_n = (U_1, U_2, \ldots, U_n)$ by representing $\mathbf{U}'_n$ as embedded in an infinite length test $\mathbf{U}'_\infty = (U_1, U_2, \ldots, U_n, \ldots) = (\mathbf{U}'_n, U_{n+1}, \ldots)$.

A philosophical remark is in order about this substitution of an infinite-length test. The notion of sampling examinees from a large finite population idealized as an infinite population of examinees is an almost universally accepted modeling device (for example, assuming that the latent distribution is standard normal and embracing the random sampling examinee perspective

presumes such). What is slightly less universally accepted is that there is another asymptotic aspect, that of a long test, or more abstractly and realistically, that of a long-test manufacturing process (as introduced in Stout, 1990; and nicely articulated in Douglas, 1997). Indeed, ability estimation asymptotics is impossible without it. In summary, one derives properties of a virtual test $\mathbf{U}_\infty$ of length ∞ in order to understand properties of a real length n test $\mathbf{U}_n$ (n large enough to be thought as a "long" test and hence well represented by $\mathbf{U}_\infty$).

This shift to studying $\mathbf{U}_\infty$ allows us to discuss the asymptotic consistency of ability estimation procedures. It also allows us to easily define the notion of a test being essentially unidimensional when the length n test is not strictly unidimensional. The point is that this infinite-length test abstraction, which in fact is just as valid as the infinite-population abstraction, permits a practically useful conception of a test having one dominant or essential dimension.

Of particular importance, consider the case of long test and large examinee sample *joint asymptotics*. A number of important parametric logistic IRF and nonparametric IRF results establishing joint consistency of item structure and examinee latent ability joint estimation require both examinee sample size and test length to cooperatively go to ∞ (see Douglas, 1997; Haberman, 1977; Trachtenberg & He, 2002). Indeed, joint examinee and item asymptotics is a powerful foundational modeling device, as discussed further below.

If one accepts the justification of an infinite length test formulation, a rigorous definition of when a test $\mathbf{U}_\infty$ is essentially unidimensional with respect to a latent variable θ can be given.

Definition 4. A test $\mathbf{U}_\infty$is *essentially unidimensional* with respect to the unidimensional latent random variable Θ if for all θ,

$$\frac{\Sigma_{1\leq i<i'\leq n}|\mathrm{Cov}(U_i, U_i'|\Theta=\theta)|}{\binom{n}{2}} \rightarrow 0 \tag{5}$$

as $n \rightarrow \infty$.

Here θ is conceptualized as the dominant dimension intended to be measured. In general, other existing (though asymptotically vanishing) dimensions force the conditional covariances to be nonzero with respect to θ. An excellent example is provided by the content dimensions of paragraphs upon which paragraph-based testlets are based in a reading comprehension test.

From Definition 4 flows the satisfying theoretical fact that total test score, appropriately rescaled, consistently estimates the *unique* essential latent dimension θ as test length $n \rightarrow \infty$, or, equivalently, that the number correct score consistently estimates ability on the latent true score scale (Stout, 1990). This mathematically proved uniqueness (modulo the equivalence class of monotone increasing transformations of θ of course) of the latent ability scale is philosophically and practically important because it justifies the notion that a unidimensional test measures a specific latent ability. This result, of course, does not imply any kind of asymptotic estimation efficiency of number correct (rescaled) for parametric models like 2PL and 3PL (such a claim being false). But, it is certainly important from the foundational perspective and useful from the NIRT modeling perspective, where the absence of a parametric model precludes using parametrically efficient estimators such as a maximum likelihood estimator (MLE).

A delicate foundational issue cannot be bypassed. It is the identifiability question for a unidimensional latent model, presuming that the latent distribution of Θ is specified in order to rule out trivial causes of nonidentifiablity. One realizes that some nonidentifiability for the IRFs of a MLI1 model of a finite-length test must exist when the family of permissible IRFs is allowed to be fully nonparametric. That is, different sets of IRFs for a MLI1 model with specified latent ability distribution can produce the given manifest distribution for $\mathbf{U}_n$. By contrast, for the infinite-test-length MLI1 $\mathbf{U}_\infty$ formulation, it is reassuring from the foundational NIRT perspective to learn that we have identifiability asymptotically for the infinite set of model-fitting IRFs. For a careful formulation and proof of this result, see Douglas (2001).

Jeff Douglas' result legitimizes, for a test of adequate length, the search for a statistical procedure to jointly estimate IRFs nonparametrically together with the estimation of examinee abilities (a sort of nonparametric LOGIST). In Douglas (1997), joint ability and IRF uniform asymptotic consistency is proved for a unidimensional class of kernel-smoothing-based NIRT IRF estimation procedures as test length and examinee sample size jointly and cooperatively approach infinity. Interestingly and predictably, this result holds in spite of the finite test length nonidentifiability of IRFs, a barrier removed by letting test length approach infinity.

The Douglas and Cohen (2001) paper presents a nonparametric IRF estimation procedure growing out of this joint estimation consistency result. This paper provides a kernel smoothing approach, inspired by Jim Ramsay's TESTGRAF (see Ramsay, 2000) is used together with a nonparametric hypothesis testing approach to assess lack of fit for parametric IRT models such as 1,2,3PL models. A NIRT model is fit using kernel smoothing, and its lack of fit to the closest-fitting logistic model is assessed. Implicit in this, I would propose, is a hybrid unidimensional statistical model-fitting approach: fit IRFs parametrically when the fit is sufficiently good and otherwise fit the IRFs nonparametrically, using the Douglas and Cohen approach or, using Jim Ramsay's functional data analysis approach (see Rossi, Wang, & Ramsay, 2002) to carry out the details.

Another foundational asymptotic result, an answer to a question posed by Paul Holland (1990b), is the establishment of the asymptotic posterior normality of Θ given examinee response pattern $\mathbf{U}_n = \mathbf{u}_n$ under any of a very large class of parameterized (that is, the IRFs are specified known parametric functions of θ) MLI1 IRT models, as defined by a broad and unrestrictive class of regularity assumptions (Chang & Stout, 1993). This result is both parametric and nonparametric. It is parametric because computation of the maximum likelihood estimator of θ, which requires knowledge of the parametric IRFs, is needed for practical applications of the result, and it is nonparametric in that its conclusion of asymptotic normally holds for a very broad class of MLI1 IRT models.

Brian Junker had been interested starting with his thesis in finding an empirical characterization (that is, determined by the "manifest" examinee response distribution as test length goes to infinity) of when a MLI1 IRT model is possible (see Junker, 1993, in particular). Brian Junker and Jules Ellis (Junker & Ellis, 1997; also Ellis & Junker, 1997) in one of several cross-Atlantic NIRT cooperations, used the concept of conditional association, due to Holland and Rosenbaum (1986), to produce an empirical characterization of when a MLI1 model is possible, namely that $\mathbf{U}_\infty$ must be conditionally associated and satisfy "vanishing conditional dependence" (see the Ellis & Junker papers for details).

Of particular modeling interest in the Junker and Ellis work is that their approach required use of theoretical constructs from advanced mathematical probability (in particular the concept of a tail σ-field) that in turn illuminate the foundational modeling issue of the stochastic-subject (within subject sampling) versus the random-sampling (of examinees) formulation for IRT modeling (see Holland, 1990a, for a thorough discussion of the sampling foundations of IRT).

More specifically, out of Ellis and Junker's sophisticated long test asymptotic approach comes a proposed new foundational paradigm uniting the stochastic subject versus randomly sampled subject latent trait modeling dichotomy posed by Holland. In particular, Ellis and Junker create the *stochastic meta-subject*. The intriguing idea is to create equivalence classes of subjects, where class membership is defined by members of the same class being indistinguishable with respect to their long-run test behavior. Each such equivalence class becomes by definition an indivisible (atomic) stochastic meta-subject. One is then free to interpret each such meta-subject conceptually either from the examinee random sampling perspective as a collection of many indistinguishable subjects to be sampled repeatedly or as a single stochastic subject to be sampled repeatedly (via the "washing the brain clean" virtual experiment imagined by the stochastic subject interpretation). Also foundationally intriguing is the briefly introduced construct of asymptotic specific objectivity.

2.3. Interpreting Conditional Covariances Geometrically to Assess Latent Multidimensional Structure

The weakening of strict unidimensionality (our version being the existence of a unidimensional *WLI* M IRT model, recall) to essential unidimensionality (the existence of a M IRT model with (5) holding) uses item pair conditional covariances in its formulation. In addition to helping answer issues of unidimensionality, one can legitimately ask how useful such conditional covariances can be for assessing the multidimensional structure when many dominant latent dimensions maybe present. That is, can we use these conditional covariances to infer important aspects about the multidimensional latent structure that is generating the data?

Zhang (see Zhang & Stout, 1999a) gives a strongly affirmative answer, powerfully showing that such conditional covariances do indeed provide important information about the latent multidimensional structure. Zhang adopts a semiparametric model formation that encompasses a broad class of parametric IRF models in forming the class of *generalized compensatory models*. The major assumptions that define this class of models are latent trait multinormality (one natural way to specify the multidimensional latent distributions), compensatory modeling in the sense that each item's monotone IRF is a linking function of a linear combination of the latent traits of the latent space (thus parameterizing the relative contribution of each latent dimension) to the probability of correct item responding, and monotone IRFs. Indeed, foundationally, the assessment of the multidimensional latent structure is a mathematically well-defined problem only when certain assumptions, such as requiring generalized compensatory IRFs, are made.

Definition 5. An MLI1 IRT model is *generalized compensatory* provided

$$P_i(\theta) = \mathsf{H}_i(\Sigma_{j=1}^{d} a_{ij}\theta_j - b_i), \;\; \mathsf{H}_i(-\infty) \geq 0, \; \mathsf{H}_i(\infty) = 1 \tag{6}$$

where each linking function $\mathsf{H}_i(\cdot)$ is required only to be monotone increasing, $\mathbf{a}_i' = (a_{i1}, a_{i2}, \ldots a_{id})$ is the item's discrimination vector with respect to the d-dimensional latent space indexed by $\boldsymbol{\theta}_d' = (\theta_1, \theta_2, \ldots, \theta_d)$, and b_i is a centering parameter associated with item difficulty.

Zhang views generalized compensatory models geometrically, each item being represented using its item discrimination vector $\mathbf{a}$ as having a direction in the latent space indexed by $\boldsymbol{\theta}$ as shown in Figure 1.

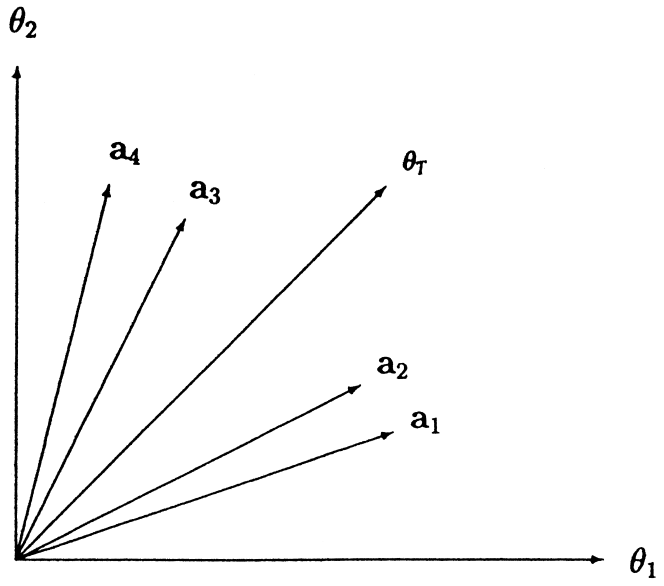

FIGURE 1.
Geometric representation of a four item two-dimensional test.

The notion of the *direction of best measurement* θ_T of the test score $X = \Sigma_{i=i}^{n} U_i$, or more generally of a subtest score Y, in the multidimensional latent space indexed by $\boldsymbol{\theta}$ is key to Zhang's development. In practice, a statistical procedure based on the conditional covariance perspective must estimate item pair conditional covariances given a direction of best measurement θ_Y for a specified subscore Y. The prototypical way to estimate such conditional covariances is to condition on the subtest score Y (possibly the test score) of a carefully selected subtest, and partition examinees based on the score. Then one can estimate the expected conditional covariance $E[\mathrm{Cov}(U_i, U_i'|\Theta_Y)]$ by first estimating the covariance for examinees within each partitioning interval (intervals determined by a lattice of y values of Y), namely $\mathrm{Cov}(U_i, U_i'|Y \approx y)$, in the ordinary way and then taking the weighted average of these estimated conditional covariances using the empirical distribution of the partitioning score Y over the partitioning intervals to determine the weights.

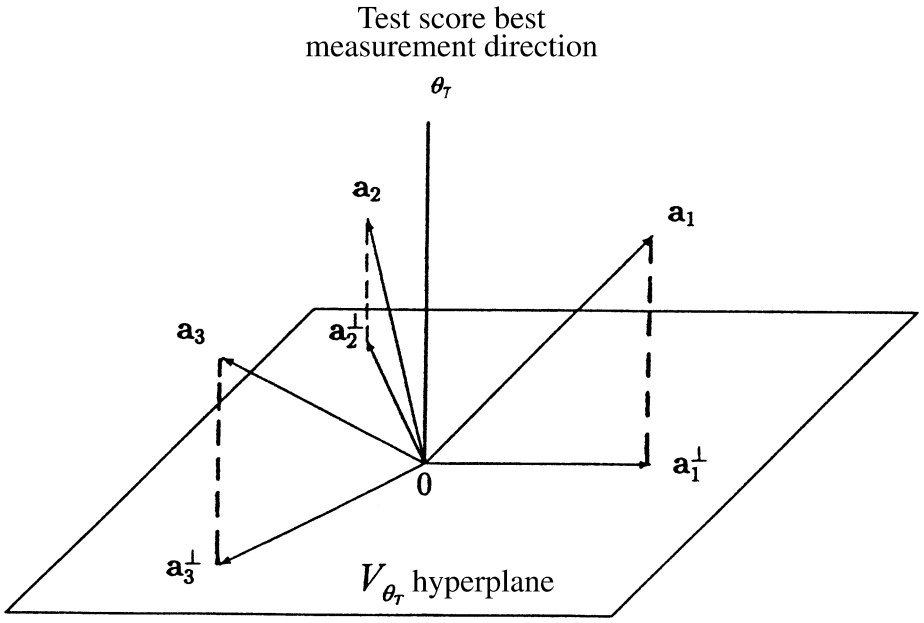

FIGURE 2.
A three dimensional test with projections of item discrimination vectors onto V_{θ_T} hyperplane.

The Zhang and Stout paper then studies how such expected conditional covariances given a direction of best measurement θ_T of the test score illuminate our understanding of the multidimensional latent structure of the test. In a set of theoretical results, striking both for their intrinsic simplicity and their practical usefulness, it is shown that the angles between the various item pair discrimination vectors for an item pair as projected on the hyperplane V_{θ_T}, defined as the hyperplane perpendicular to the direction of best measurement θ_T, actually reveal much information about the multidimensional latent structure.

That is, conditional covariances are useful in informing the practitioner about the latent multidimensional structure. In particular, when an item pair has a projected angle of less than 90 degrees, the items tend to combine to form the same dimension, and when the projected angle is greater than 90 degrees they tend to separate to form distinct dimensions. This flows nicely into a practical, geometrically-based definition of *approximate simple structure* (see Definition 2 of Zhang & Stout, 1999b) such that when the definition holds, items can be partitioned into dimensionally distinct but approximately unidimensional clusters. For example, the six items of Figure 3, as shown by their projection onto V_{θ_T}, form a three-dimensional approximate simple structure.

Interesting research questions remain to further develop this powerful geometric generalized compensatory approach of Jinming Zhang.

2.4. NIRT-Based Statistical Procedures, Emphasizing Conditional Covariances

Out of section 2.3's conditional-covariance-based semiparametric compensatory IRT modeling foundation, four useful statistical procedures have flowed. Together, they constitute a

coordinated statistical thrust to assess when unidimensionality holds and, when it is shown to fail, to assess important characteristics of the multidimensional latent trait structure.

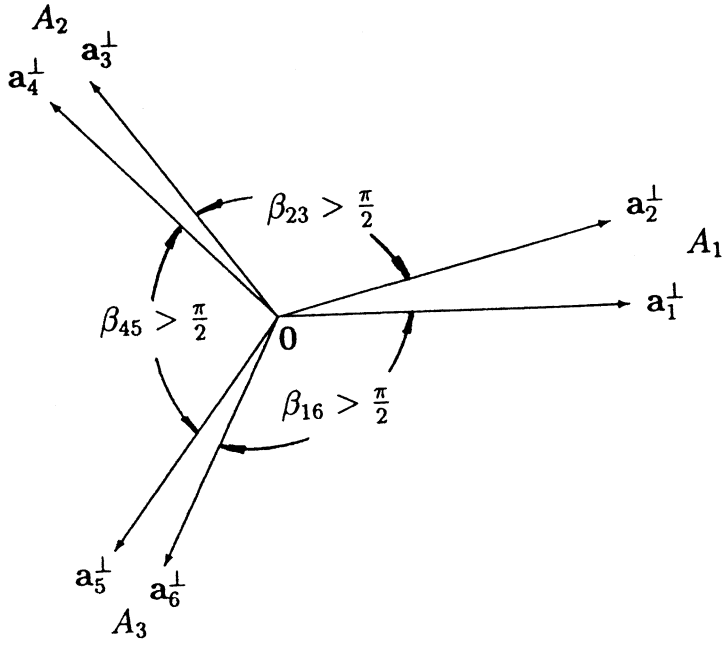

FIGURE 3.
Projection of item discrimination vectors onto V_{θ_T} hyperplance for a six item three-dimensional approximate sample structure.

DIMTEST, a conditional covariance based nonparametric statistical test of unidimensionality, was formulated by Stout (1987) and refined by Nandakumar and Stout (1993). Its statistical power depends on effective user selection (proceeding substantively or using exploratory statistical tools) of a set of items called the assessment subtest (AT1) that has been chosen to be dimensionally homogeneous and dimensionally distinct from the remaining items, referred to as the partitioning subtest (PT). This PT subtest's score is used as the conditioning subscore in the DIMTEST required conditional covariances. When unidimensionality holds, the approximate unbiasedness of the DIMTEST hypothesis testing statistic depends on the effective selection of a second bias correcting set of assessment subtest items, denoted AT2. The modern era of DIMTEST, fueled by work of Furong Gao and Amy Froelich (Stout, Froelich, & Gao,2001) replaces having to choose actual AT2 items, which sharply limits the range of applicability of DIMTEST, by estimating the needed AT2 based bias correction by means of a resampling scheme using nonparametric kernel smoothed estimation of AT1 IRFs to create a virtual AT2. The latest and by far most statistically effective methods for the choice of AT items (there now being only one assessment subtest) use variations of our various conditional covariance based procedures to

replace the linear factor analysis exploratory procedure for choosing AT1 originally used in DIMTEST (Froelich & Habing, 2002).

In Zhang and Stout (1999a), Zhang rigorously defends the capability of DIMTEST to have the statistical power to detect multidimensionality when there may be three or more dimensions (Stout's, 1987, theoretical defense of DIMTEST's statistical power was more informal and its consideration of statistical power was limited to there being two dimensions).

In his thesis, Louis Roussos developed a conditional-covariance-based hierarchical cluster analysis approach (its proximity measure uses conditional covariances), HCA/CCPROX, to sort items into relatively dimensionally homogeneous clusters (Roussos, Stout, & Marden, 1998). This approach can play a valuable exploratory role in discovering the tendency of items to form dimensionally distinct clusters that are internally dimensionally homogeneous, that is, approximate simple structure. Even when approximate simple structure fails, HCA/CCPROX is useful. In fact, the Zhang conditional covariance geometry provides a theoretical justification for HCA/CCPROX's capability to identify relatively dimensionally homogeneous item clusters, whether they are sufficiently dimensionally distinct to produce an approximate simple structure or not.

In her thesis, Hae Rim Kim (see Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996) developed the conditional-covariance-based DETECT procedure. In his thesis, Jimning Zhang refined DETECT and developed a theoretical defense of it (see Zhang & Stout, 1999b). DETECT counts the number of latent dimensions, provides an index that measures the strength of the departure from unidimensionality, and, when appropriate, partitions items into approximate simple structure item clusters. The concept of the strength of the departure from unidimensionally is distinct from the number of dimensions: For instance, a two-dimensional structure can strongly depart from unidimensionally, while a 10-dimensional structure can weakly depart from unidimensionality.

In his thesis, Brian Habing (Douglas, Kim, Habing, & Gao, 1998; and Habing, 2001) developed a local estimate of the conditional covariance function given θ. All the previously discussed conditional-covariance-based procedures are global in that they estimate the expected value of conditional covariances, or use other global averaging over θ. The resulting estimation procedure is called CONCOV. Some test phenomena, such as end-of-test speededness, can be investigated using Habing's estimated conditional covariances (see Douglas et al.).

The three global conditional covariance-based procedures are applied to real data in an integrated manner in Stout, Habing, Douglas, Kim, Roussos, and Zhang (1996), with a high point being the dimensional analyses of the analytical reasoning and reading comprehension sections of the LSAT. In these analyses, paragraphs strongly displayed themselves as contributing distinct testlet-based latent dimensions. In one of my favorite applied findings, DETECT combines two paragraphs as producing a single dimension, an apparent error until one discovers that both paragraphs are science-based (which may or may not be the true explanation, but, it is certainly suggestive). The Stout et al. paper is one of the best places to read about the three global conditional covariance procedures (DIMTEST, HCA/CCPROX, and DETECT).

Section 2 has summarized the 15-year research effort of the Lab to develop the IRT nonparametric assessment of the multidimensional latent structure that underlies any educational test. Philosophically, the substantive specifications of what a test measures and the statistical

description of the resulting latent IRT structure of the test should be consistent and should enhance and inform each other from the test construction and test scoring perspectives. It has been shown how foundational work, which draws heavily on the infinite-item test formulation, and a constellation of conditional-covariance-based nonparametric latent dimensionality assessment procedures, which are defended by theory—especially the geometrically based theory for conditional covariances, combine to provide a flexible, theoretically and empirically well supported, informative, and easy to apply set of nonparametric multidimensional latent structure assessment tools. As reported above, a mixture of large-scale simulations and real data analyses has been presented in the psychometric literature to demonstrate and justify the effectiveness of these conditional-covariance-based dimensionality procedures. In addition to the work reported upon above, much interesting, challenging, and, from the applied measurement perspective, important, work remains concerning conditional-covariance-based approaches to IRT multidimensional latent structure modeling and statistical analyses. One long-term research goal is to develop an effective and relatively complete NIRT item-level factor analysis methodology.

3. Test Fairness

High stakes testing is increasingly used worldwide to inform educational admissions, placement, and honors/awards decision-making processes. Moreover, the level of ethnic and cultural diversity in most countries that rely heavily on high stakes testing continues to increase. Thus, issues of test fairness are of vital and ever-increasing importance. The statistical analysis of item-level test fairness is universally called Differential Item Functioning (DIF). The role of psychometrics in informing our understanding of test fairness and in improving test fairness is often inappropriately compartmentalized, minimized, or even bypassed entirely. The test fairness challenge to psychometrics has been, and still is, to change this unfortunate state of affairs!

One of the subtle ways that DIF has been compartmentalized is the almost total disconnect that has evolved between substantive (content-based) and DIF (statistical) approaches to the understanding and practice of test fairness. For example, Paul Ramsey (1993), in a review of what was then the ETS sensitivity review process, states, "...there is no consistent effort to inform sensitivity reviewers what we are learning from DIF" (p. 385). Reacting to this reality, Robert Linn (1993) recommends "taking into account not only what has been learned from DIF analyses but [also] what has been learned from sensitivity reviews" when standardized tests are being designed (p. 364).

The applied DIF literature is strewn with examples of the failure to explain substantively why certain items display DIF and why certain items don't display DIF when substantive analyses suggest likely unfairness. Addressing this very point, William Angoff (1993), in his introductory article to the *Differential Item Functioning* monograph, states, "It has been reported by test developers that they are often confronted by DIF results that they cannot understand; and no amount of deliberation seems to help explain why some perfectly reasonable items [from the substantive perspective] have large DIF values" (p. 19).

In reacting to these enervating problems, the first of the Lab's central research goals was to facilitate the effective, and we believe necessary, integration of statistical and substantive approaches to test fairness. This was distilled into the psychometrically expressed goal of

developing a theoretical multidimensional IRT model that rigorously captures and explicates the intuitive notion that the cause of DIF in singly-scored tests is the presence of secondary dimensions, denoted by η's, other than the primary or essential dimension intended to be measured and denoted throughout by θ.

Thus, a second central DIF goal of the Lab was to use this model is to shift the psychometric DIF paradigm from a totally reactive (removing unfair items after they have been constructed and pretested) and single-item-based approach to a partially proactive (that is, also applied at the test design stage) and item-bundle-based approach to DIF. This approach stresses *substantively interpreted* latent-dimensionality-based explanations of causes of DIF that then can contribute to feedback loops for improving future test design specifications. In this manner, substantive and psychometric approaches to test fairness can be unified. For example, if reading-comprehension test items of paragraphs that discuss the physical sciences are discovered to display DIF against women, then the test specifications for future versions of such a reading comprehension test might exclude physical-science-based paragraphs.

Following in the tradition of Holland and Thayer's (1988) theoretically-defended Mantel-Haenszel IRT DIF approach (a foundational and practically important milestone, marking the beginning of the modern era of DIF from the psychometric perspective), a third DIF goal of the Lab was to embark on a lengthy effort to produce a constellation of nonparametric IRT DIF procedures to address various issues in conducting DIF analyses that were judged important and remained unsolved. These included crossing DIF, DIF for polytomously scored items, developing statistically robust DIF procedures, DIF for intentionally multidimensional tests, item-bundle-based DIF, DIF in CAT settings where examinees must be matched using IRT estimated θ, appropriately defining and estimating model-based theoretical DIF parameters for scaling various kinds of DIF in various settings, and local (in latent ability θ) DIF.

3.1. Multidimensional Model for DIF (MMD)

As is traditional, DIF is herein considered a phenomenon experienced by a targeted *focal group* (F), such as African Americans, when compared with a nontargeted *reference group* (R), such as Caucasians. Typically used focal groups include those defined by race, gender, ethnicity, disability, first language, and so forth. DIF for an item is defined to occur when the probability of a correct item response, for examinees with the same intended-to-be-measured ability θ, differs because of group membership. It is important to note that, *as is appropriate*, this definition has nothing to do with whether the focal and reference group distributions of Θ are identical or are stochastically ordered.

In two papers (Shealy & Stout, 1993a, 1993b; our MMD model was first discussed in Shealy, 1989), Robin Shealy and I laid out our multidimensional model for DIF, abbreviated MMD. The 1993b reference provides a detailed in-depth theoretical description of the model. It carefully derives various interesting consequences of a foundational nature that mathematically follow from the model. The 1993a reference provides a more brief and informal description of the aspects of the model needed to understand the SIBTEST DIF procedure. The most complete description of the MMD model from the applications perspective occurs in Roussos and Stout (1996a). It should be noted that Kok (1988) independently and concurrently to Robin Shealy and me

developed a similar multidimensional modeling approach to DIF. The Shealy/Stout MMD model calls for a new focus on DIF based on three related principles.

First, consistent with the test validity perspective of testing, and noting that educational and training decisions involving examinees are made on the basis of test scores rather than item scores, the assessment of test fairness should occur at the test score level rather than the item level. This insight led to the closely related modeling concepts of *differential bundle functioning* (DBF) and *differential test functioning* (DTF). DBF (analogously, DTF) is defined to occur for examinees of the same intended-to-be-measured ability θ when the expected item bundle subscore (analogously, test score) given θ for a carefully selected, and likely substantively and dimensionally homogeneous, bundle of items differs across group. DBF measures the combined amount of DIF at the item bundle score level experienced by examinees from different focal and reference groups. Of course, single-item DIF must always be a vital component of assessing test fairness; our approach has been to augment and embed such DIF considerations by including a needed additional focus on DBF and DTF.

Second, the explicitly multidimensional nature of the model allows, and exhorts, us to rigorously study and understand the necessary role of secondary dimensions in causing DIF, DBF, or DTF. In particular, when several items (perhaps forming a dimensionally homogeneous and substantively interpretable bundle, such as a set of items on a geography test that each require map-reading skills) each depend on the same secondary dimension, the possibility of a large amount of DBF experienced by the focal group at the bundle subtest score level caused by individual item *DIF amplification* (see Nandakumar, 1993) becomes an issue. Or, when the influences of multiple secondary dimensions interact, the possibility of *DIF cancellation* (see Nandakumar, 1993, again) of the influence of DBF-producing bundles at the test score level (such as a reading comprehension test where the paragraphs are carefully balanced by content, based on explicit consideration of gender) or cancellation of the influence of DIF-producing items at the bundle score level, becomes important. Since people are cognitively heterogeneous and since items cannot, and should not, be context-free, the notion of DIF cancellation and DBF cancellation is perhaps more important than casual thought first suggests.

Third, DIF is explicitly and correctly conceptualized by the MMD model to occur locally in θ. This incontrovertible and seemingly innocuous fact has important implications. It naturally leads us to the possibility of "crossing DIF" occurring, where, for example, an item could display DIF against low-scoring focal-group examinees while it also displays DIF against high-scoring reference group examinees. Further, local DIF at θ is caused by differing conditional distributions of a secondary dimension (for simplicity, we only consider the case of one secondary dimension) H given θ across the focal versus reference group (here H is the upper case version of η). This is counter to the widely and uncritically accepted informal model for what causes DIF, which presumes, usually implicitly, that differences in the marginal H distributions across the two groups are what causes DIF. Naive trust in this seductively intuitive, but sometimes inaccurate, informal model creates some striking paradoxes as discussed below.

In explaining the MMD of Shealy and Stout, for simplicity, we only consider the two-dimensional case, with θ denoting the dimension intended to be measured and the random variable H (taking on values η) denoting an additionally measured secondary dimension. H is often but not always thought of as a "nuisance dimension" being outside the

intended-to-be-measured construct. Indeed, a subtle, interesting, and important issue arises when H seems, or clearly is, part of the construct intended to be measured. IRF invariance is assumed to hold for all items with respect to the complete latent (θ, η) space. That is, for all items, the IRFs $P_i(\theta, \eta)$are the same for the focal and reference groups. Thus, when the latent space is completely specified, all group differences in item performances must, by definition, disappear.

In general, the focal and reference group Θ_F and Θ_R distributions will be different and often stochastically ordered. The possibly differing Θ_F and Θ_R distributions are *not* the source of DIF, DBF, or DTF, although such a difference in distributions clearly can contribute to group differences in test score distributions, along with the group differences in score distributions that are caused by the presence of DIF.

For the interested reader, I now briefly review the essence of the MMD model. The potential for DIF occurring at θ for an item with (group invariant) IRF $P(\theta, \eta)$ is caused by the conditional distribution of H_g *given* $\Theta_g = \theta$ differing for $g = R$ versus $g = F$. To see this, compute the group dependent marginal IRF with respect to θ;

$$P_g(\theta) = \int_{-\infty}^{\infty} P(\theta, \eta) f_g(\eta|\theta) d\eta, \tag{7}$$

where $g = R$ or F and $f_g(\eta|\theta)$ is the density of H_g given $\Theta_g = \theta$. Then the amount of DIF against F locally at θ is given by

$$B(\theta) = P_R(\theta) - P_F(\theta). \tag{8}$$

Further, the average amount of unidirectional DIF against F is given by

$$\beta_{\mathsf{UNI}} = \int_{-\infty}^{\infty} B(\theta) f_F(\theta) d\theta \tag{9}$$

where $f_F(\theta)$ denotes the density of Θ_F. β_{UNI} is the fundamental DIF index of MMD. It is what the SIBTEST DIF procedure estimates and about which it tests hypotheses. β_{UNI} for an item bundle is defined analogously with $B(\theta)$ denoting the difference of reference and focal expected item bundle scores at θ.

The fact that DIF is due to a difference in the conditional distributions of H_g given θ across group rather than to differences in the marginal H_g distributions across group can be the source of paradoxical situations where differing marginal H_g distributions across group fail to translate into DIF and where nondiffering marginal distributions of H_g nonetheless end up being accompanied by DIF (as discovered and explained by Louis Roussos; see Roussos & Stout, 1996a). This just-referenced paper gives a plausible example of this latter and most paradoxical case, using some seemingly contradictory findings of O'Neill and McPeek (1993) and Douglas, Roussos, and Stout (1996): O'Neill and McPeek found that some types of items, intended to measure verbal reasoning (θ), tended to exhibit DIF in favor of males (R) versus females (F) when the items concerned "practical affairs," (H), for which money is mentioned as a typical component. The finding of DIF in favor of males on these items thus indicates that $\mu_{\mathsf{H}_R} > \mu_{\mathsf{H}_F}$ must hold by the informal DIF model perspective that differing across-group marginal distributions for H_g cause DIF, where H denotes familiarity or knowledge of money and other practical affairs. By contrast, and paradoxically, Douglas et al. (1998) reported that, contrary to what the informal

model for DIF predicts, items intended to measure logical reasoning (θ) where the context is money or finances (H), showed little DIF against either males or females even though $\mu_{\mathsf{H}_R} > \mu_{\mathsf{H}_F}$ is believed to hold.

We resort to the MMD model to explain this apparent paradox. Now, if (Θ_g, H_g) are bivariate normal for each group g (a reasonably innocuous assumption), we have for both tests that

$$E(\mathsf{H}_R|\Theta_R = \theta) - E(\mathsf{H}_F|\Theta_F = \theta) = (\mu_{\mathsf{H}_R} - \mu_{\mathsf{H}_F}) - \rho(\mu_{\Theta_R} - \mu_{\Theta_F}) \tag{10}$$

for all θ. A close examination of (10) reveals a possible explanation for the apparent contradiction between observed DIF behavior for the verbal reasoning items and the logical reasoning items. As already stated, for both the logical reasoning items and the verbal reasoning items the first term on the right hand side of (10) resulting from the nuisance dimension H is positive. It is in fact the second term that explains the paradox. In the case of the verbal reasoning items, because females are as proficient or more proficient in verbal reasoning (θ) on average, the second term on the right hand side of (10) (including its minus sign) is positive as well, almost ensuring DIF in favor of males. However, in the case of the logical reasoning (θ) items, in which males seem to have a higher mean proficiency, the now *negative* second term on the right hand side of (10) tends to cancel the influence of the positive first term. Thus, the genuinely paradoxical, but logically correct, conclusion is that DIF against F can be reduced, at least to some extent, if the ability distribution on the primary (intended-to-be-measured) dimension θ also favors R. Note, however, that this is a delicate matter depending on the size of ρ.

3.2. Model-Based Parameterization of the amount of DIF in Various Settings

One important advantage of having a believable and realistic mathematical model for a complex real world phenomenon is that important and sometimes subtle aspects are quantified within the model and hence can be better understood and inferred from data. In the case of the MMD model, model-based scaling of various kinds of DIF, as the character of the DIF varies, the number of items involved varies, the number of dimensions intended to be measure by the test varies, and so forth, becomes possible.

The fundamental parameter β_{UNI} of the MMD model measures the amount of unidirectional DIF against F averaged over all levels of θ. Similarly, within the framework of MMD, as already remarked, an analogous parameter measuring unidirectional DBF is easily defined (see Shealy & Stout, 1993a). Parameters measuring the amount of crossing DIF for an item and for a bundle of items are defined by Li and Stout (1996). The polytomous case is handled by using the expected item score given θ as the parameter providing the DIF metric (Chang, Mazzeo, & Roussos, 1996). In Stout, Li, Nandakumar, and Bolt (1997), the parameter measuring the amount of DIF when the test is designed to measure two dimensions is defined carefully by modifying the basic formulation by replacing the unidimensional θ by a two-dimensional (θ_1, θ_2). For example, the test might be a mathematics test designed to measure geometry, θ_1, and algebra, θ_2.

Finally, in an important foundational paper with important implications for certain test fairness applications, Roussos, Schnipke, and Pashley (1999) develop (from first principles and a heuristic asymptotic argument) the DIF parameter measuring the amount of DIF that is in effect estimated by the widely used Mantel-Haenszel (MH) odds ratio DIF estimator. One practical

consequence based on the findings of the Roussos et al. paper is that significant caution should accompany the use of the MH odds ratio as an index of DIF when the 3PL model is used.

3.3. MMD- Inspired DIF Statistical Procedures

The initial methodological challenge for the Lab was to develop a nonparametric estimation and hypothesis-testing procedure to assess item DIF and bundle DBF. After having "rediscovered" the Dorans and Kulick (1986) standardization metric for DIF together with being impressed with the power of the theoretical rationale used to justify the MH procedure, as articulated by Holland and Thayer (1988), the Lab had the task of producing a robust and theoretically defensible standardization-based metric for DIF by modifying and augmenting the standardization statistic as needed. The resulting SIBTEST procedure is described in Shealy and Stout (l993a).

Recall that, by definition, DIF occurs when examinees, matched on the latent variable θ *that the test is intended to measure*, perform differentially depending on their group membership. Hence, the SIBTEST strategy was, as best as one could, to match examinees on a *valid subtest* whose score measures θ without serious contamination from secondary $\eta_1, \eta_2, \ldots$, dimensions. Here, the MMD model helps us see that the contamination issue is at the matching subtest score level rather than the individual item level. Hence, although the ideal is to have only items measuring θ alone in the valid subtest, having items present that are influenced by secondary dimensions is also an option as long as their influence approximately cancels out at the valid subtest score level.

We chose the terminology of "valid subtest" carefully, intending to stress that the matching criterion used by SIBTEST, MH, and other procedures should be selected with validity-based care, rather than automatically using total test score (perhaps with the studied item removed). Potential SIBTEST users sometimes misunderstand this terminology and think that SIBTEST, in contrast to other DIF procedures, cannot be used without its having a truly valid subtest. Rather, the point is that SIBTEST,*in common with all other DIF procedures as well*, is only as effective as the matching subtest is in matching examinees on the construct intended to be measured. In this regard, serious contamination of the valid subtest caused by secondary, nuisance dimensions must not occur. Indeed, our intent in introducing the "valid subtest" terminology was to stress that practitioners' efforts to achieve validity in matching examinees is a vital prerequisite to the success of any DIF/DBF/DTF analysis. That is, at its heart, a test fairness analysis requires a prerequisite validity analysis.

How practical it is to find valid subtests is a partly experiential question. First, in the prototypical large standardized test setting where one wants to assess pretest items for possible DIF, matching examinees on the score on operational items, which have previously undergone careful test design and psychometric scrutiny, seems reasonable (as Paul Holland suggested in a private communication). If reasonably done, attempts to purify an initially proposed matching subtest by removing DIF items seem appropriate too. Regarding the choosing of the valid subtest, one should note that the SIBTEST procedure invites the user to specify exactly which items constitute the valid subtest.

Using the standardization metric of Dorans and Kulick (1986), the fundamental idea of

SIBTEST is simple and intuitive: Let X_V be the score on the user-chosen valid subtest judged to be measuring θ without serious contamination. (Borrowing from our NIRT work, one would hope that θ lies close to Zhang's direction of best measurement of X_V.) Consider a preselected bundle (could be a single item) of possible DIF items with bundle score denoted by Y. In particular, let $\bar{Y}_{gk}$denote the average bundle score of all Group g (either R or F) examinees having valid subtest score $X_V = k$. Then the proposed SIBTEST estimator of β_{UNI} is given by

$$\hat{\beta}_{\mathsf{UNI}} = \Sigma_k(\bar{Y}_{Rk} - \bar{Y}_{Fk})\hat{p}_{Fk} \tag{11}$$

where $\hat{p}_{Fk}$ denotes the proportion of focal group examinees for which $X_V = k$. The most common application is where the bundle consists of a single targeted possibly DIF item and $Y_{gk} = 0$ *or* 1 for each examinee.

In spite of the obvious intuitive appeal of the estimator in (11), it turns out to be seriously statistically biased when the Θ_R and Θ_F distributions are stochastically ordered, as is often the case. This is caused by the regression to the mean phenomenon: The conditional distribution of Θ_g given test score X_V depends on the differing means of the Θ_R and Θ_F distributions, as Figure 4 shows. Hence, because the distribution of $\bar{Y}_{gk}$ depends on the conditional distribution of Θ_g given test score X_V, statistical bias ($\bar{Y}_{Rk} > \bar{Y}_{Fk}$) is expected to occur when there is no DIF at θ if the marginal distribution of Θ_R is stochastically larger than that of Θ_F.

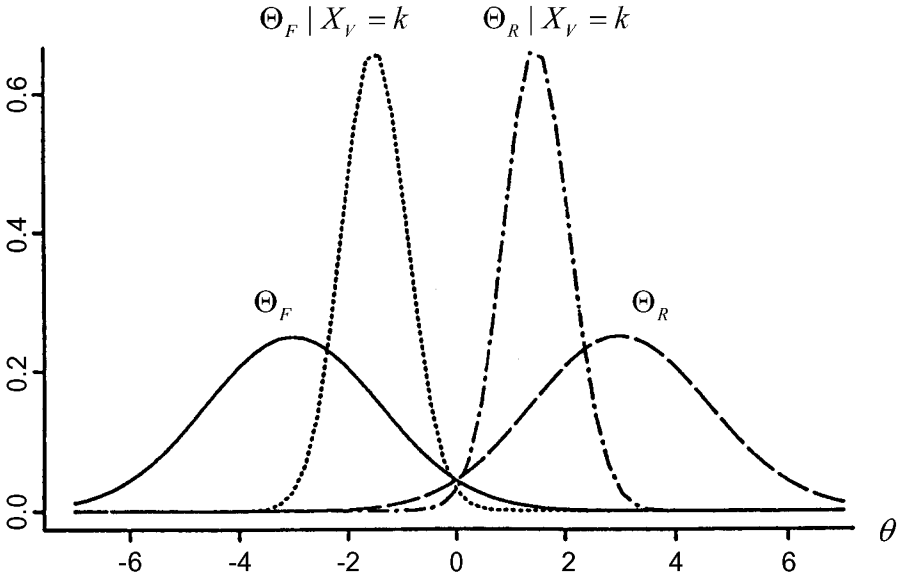

FIGURE 4.
Comparison of Θ_F and Θ_R distribution with $\Theta_F|X_V = k$ and $\Theta_R|X_V = k$ distributions.

Clearly the proposed SIBTEST statistic $\hat{\beta}_{\mathsf{UNI}}$ is biased unless test length is very large, when in fact the matching (valid) subtest score becomes completely reliable and the regression to the mean influence on $\bar{Y}_{Rk}$ and $\bar{Y}_{Fk}$ becomes negligible. The solution, as detailed in Shealy and Stout

(1993a), is to shift the $\bar{Y}_{gk}$'s for $g = R$ and F according to a heuristically-justified *regression correction*(to undo the regression to the mean influence). Dividing the regression-corrected $\hat{\beta}_{\mathsf{UNI}}$ by an appropriate estimated standard error produces a hypothesis-testing statistic. When there is no DIF, a heuristic argument shows this statistic to be standard normal asymptotically as reference and focal group sample sizes go to infinity. That the SIBTEST estimator $\hat{\beta}_{\mathsf{UNI}}$ and its associated hypothesis testing statistic perform well, as predicted by the heuristic asymptotics, is confirmed in Shealy and Stout (1993a) by a large-scale simulation study showing that the SIBTEST estimator is relatively unbiased, and that the SIBTEST hypothesis testing procedure is powerful and adheres well to the nominal level of significance in its simulated Type I error behavior over a wide range of realistic DIF models, including models allowing up to one standard deviation of difference ($d = (\mu_{\Theta_R} - \mu_{\Theta_F})/\sigma_{\Theta_g}$) in the Θ_R *and* Θ_F means, where it is assumed $\sigma_{\Theta_R} = \sigma_{\Theta_F} \equiv \sigma_{\Theta_g}$.

The large-scale SIBTEST simulation study included a study of bundle DBF as well as individual item DIF. This study showed, as expected, that the hypothesis-testing power is much greater for a bundle of DIF items contrasted with DIF items analyzed singly. The SIBTEST DIF/DBF simulation study also demonstrated valid-subtest robustness in the sense that minor contamination of the valid subtest by DIF items measuring η in addition to as θ did not produce serious deterioration of the SIBTEST performance (see Shealy & Stout, 1993a).

In Roussos and Stout (1996b), a large scale Type I error study was done using the linear regression correction version of SIBTEST, demonstrating that SIBTEST with the linear regression correction is more robust than the MH procedure in the presence of sizeable group differences in the $\Theta_g$distribution. However, even in the case of SIBTEST, some of the observed Type I error rates were unacceptably high (when group ability differences are large and an item is highly discriminating, for example), thus motivating the Jiang and Stout nonlinear regression correction version of SIBTEST: In Jiang and Stout (1999), the linear regression correction of Shealy and Stout is replaced by a more effective nonlinear correction, as shown in the paper's simulation study.

SIBTEST is based on a heuristically defended normal distribution large-sample theory. Unfortunately, in order to develop a variation of SIBTEST to detect crossing DIF, asymptotic normality fails and a randomization-test-based hypothesis testing approach was thus required. In Li and Stout (1996), through a series of simulation studies, the "CrossingSIB" procedure is shown to have good Type I error behavior and good power for detecting crossing DIF or DBF. Of some methodological interest is the fact that the Type I error portion of the simulation study included simulated examinee responses generated via nonlogistic IRFs obtained using nonparametric IRFs estimated by TESTGRAF (Ramsay, 2000).

In Chang, Mazzeo, and Roussos (1996), a polytomous version of SIBTEST called "polySIB" is developed and its performance studied. In Douglas, Stout, and DiBello(1996), the nonparametric estimation of the amount of DIF locally in θ is considered. Being able to assess DIF/DBF locally is important in situations where the test is targeted to measure ability accurately over a limited ability range such as the PSAT when used in America to award National Merit Scholarships to high ability examinees. Using nonparametric kernel smoothing estimation, the amount of DIF/DBF, as given by the function $B(\theta)$ defined above, is estimated in Douglas et

al. (1998).

In Nandakumar and Roussos (2002), a version of SIBTEST is developed where matching is on $\hat{\theta}$, thus allowing SIBTEST to be applied in CAT settings. Finally, in Stout, Li, Nandakumar, and Bolt (1997), a version of SIBTEST called "MultiSIB" is developed where the ability intended to be measured is two-dimensional and hence two-dimensional matching is required so that matching validity is maintained. Again, a thorough simulation study is done, which showed that DIF/DBF estimator bias is low and that reasonable Type I error behavior and hypothesis-testing power are observed in realistically simulated settings. This procedure is of special foundational interest because it is a particularly compelling instance of the fundamental validity mandate that the matching criterion has to be appropriately developed if the real purpose (achieving test fairness) of the DIF/DBF analysis is to be achieved.

For example, if a test is designed to measure both algebra and geometry, then clearly large DIF could occur if we match on total test score, even though the test is in fact measuring exactly what it is supposed to—namely algebra and geometry. Of course, in the case of a test designed to measure both algebra and geometry, a study where one matches on total score can be interesting from the cognitive perspective; we are, however, not assessing test fairness in the process! Indeed, in the simulation-based Type I error studies of DIF hypothesis testing with nominal level of significance set at $\alpha = 0.05$, the observed Type I error of 0.059 for two-dimensional matching was increased to 0.39 when incorrect total score matching was substituted for the valid two-dimensional matching on two subscores.

3.4. Implementation of DIF/DBF Procedures

The development of a DIF model, DIF parameters, and DIF statistics are all important DIF analysis components, but it is still the full integration of these components that is required to produce a fully developed implementation procedure for DIF analyses. Perhaps the most important in this regard are the MMD paradigmatic implications and imperatives for conducting test fairness research and practice. In summary, these include (i) integrating psychometric DIF/DBF and substantive analyses of test bias—in particular addressing underlying substantive causes of DIF/DBF; (ii) using feedback from DIF/DBF analyses to influence future test design and assembly; (iii) matching examinees in DIF/DBF studies in ways that make the matching procedure consistent with modern test validity considerations as espoused by Lee Crombach, William Angoff, Sam Messick (see chapters 1–3 of the Wainer and Braun edited *Test Validity*, 1988), and others; (iv) carrying out bundle DBF analyses where the bundles are selected to be homogeneous based on statistical, substantive, or blended grounds; (v) recognizing that the practical test validity implications of DIF/DBF are expressed at the test score level even though the "atoms" of DBF are of course the items causing DIF; (vi) addressing the potential for amplification and cancellation of item DIF at the bundle score level and of bundle DBF at the test score level; and (vii) taking advantage of the potential for increased statistical DIF/DBF detection power when one works at the item bundle level versus the individual item level.

From this test equity paradigm-changing perspective, five papers stand out. The first paper, by Terry Ackerman (1992), expresses the Shealy and Stout MMD model for DIF geometrically using the logistic multidimensional IRT model (MIRT) and studies various cases of DIF from the

perspective of studying the behavior of

$$E(\mathsf{H}_R|\Theta_R = \theta) - E(\mathsf{H}_F|\Theta_F = \theta) \tag{12}$$

as a function of θ under the assumption of bivariate normality of (Θ_g, H_g) for both groups.

Ackerman's introduced concept of the *validity sector* is potentially useful. It is consistent with the notion of essential unidimensionality presented in section 2 and can be thought of as a hypercone in the (θ, η) space having a reasonably small vertex angle. It assumes that the valid subset consists of items that are reasonably dimensionally homogeneous and reasonably close to the θ axis. Figure 5 shows a validity sector for a two dimensional test.

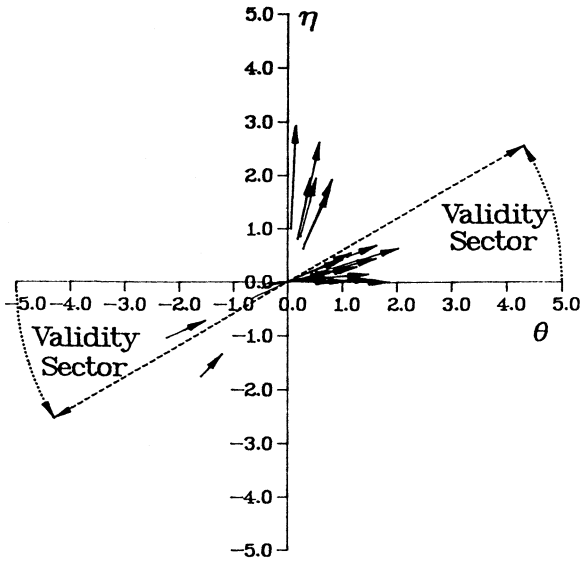

FIGURE 5.
Item discrimination vectors of a 22 item validity sector.

Three other three papers are the Douglas, Roussos, and Stout (1996) methodological paper on doing bundle DBF analyses, the Nandakumar (1993) paper on amplification and cancellation, and the Bolt, Froelich, Habing, Hartz, Roussos, and Stout (2002) paper presenting an in-depth application and further development of the bundle DBF SIBTEST methodology, applied to the GRE-Q (quantitative) exam.

The fifth paper, Roussos and Stout (1996), is the clearest and most complete statement of the MMD paradigm. It stresses that a standardized test given periodically can see its level of fairness improved over time through a principled application of substantively and statistically suggested DBF hypotheses that when affirmed (e.g., through a SIBTEST DBF analysis) can then be incorporated into the test specifications of future versions of the test. There are at least two distinct approaches to forming item bundles hypothesized to display DBF. First, as laid out in Douglas, Roussos, and Stout (1996), one can take a confirmatory approach based on the opinion

of experts. In the paper, this approach is applied to male/female DBF for eight (expert chosen) bundle-defining categories (items involving social issues, the military, technological sciences, health sciences, and so forth). Figure 6 shows the high correlation between the SIBTEST DBF $\hat{\beta}$ index and the panel of experts' own combined substantive index of the bundle's DBF.

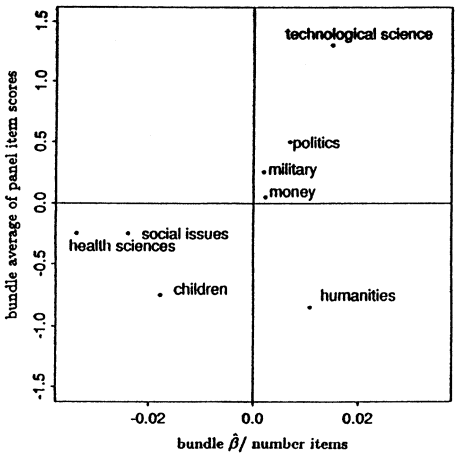

FIGURE 6.
Panel index versus bundle DBF $\hat{\beta}$/item.

The second approach uses DIMTEST and HCA/CCPROX to identify "suspect" bundles together with a substantive refinement of these statistically identified bundles suspected of DBF. In an example from the Douglas et al. paper, a bundle of six items on an NAEP history exam, which all pertained to early American documents, is identified by this blended statistical/substantive approach. Then, a SIBTEST bundle analysis of this six-item bundle produced an observed level of significance of 0.002, strongly indicating DBF against women.

Ratna Nandakumar (1993) presents real-data examples of both cancellation and strong amplification. Amplification can be of real importance in that many items having slight DIF against the same group can amplify to have a highly deleterious effect at the test score level: One question on baseball may not matter much on a sixth grade math test taken by boys and girls, but ten questions likely will.

The SIBTEST-based GRE bundle DBF analyses (Bolt, Froelich, Habing, Hartz, Roussos, & Stout, 2002) are of particular interest because of the very careful and detailed use of the bundle DBF approach as a possible aid to GRE test design from the test fairness perspective. The GRE study is the most complete example of applying the MMD paradigm in real data settings. For further interesting analyses of SIBTEST-based bundle DBF studies, the reader is referred to Gierl, Bisanz, Bisanz, Boughton, and Khatiq (2001); Gierl and Khaliq (2001); and Gierl, Bisanz, Bisanz, and Boughton (2002).

In my opinion, much remains to be done before proactive psychometric bundle-based approaches to DIF have been fully utilized in achieving test fairness. In fact, most of the

components of what we suggest above is a needed psychometrically driven paradigm shift in how to view and practice test equity has not been implemented.

4. Formative Assessment Skills Diagnosis: A New Test Paradigm

Out of the usually staid psychometric world we inhabit, perhaps the most intellectually exciting and practically important psychometric challenge since that of factor analyzing the intellect (which many characterize as the defining problem of twentieth-century psychometrics), has arisen largely unnoticed and certainly unheralded. The long dominant summative-assessment-focused and single-score-based testing paradigm that unidimensional IRT modeling so effectively addresses has begun to be challenged. Summative assessments focus on broad constructs such as mathematics, reading, writing, or physics and typically assess individual educational achievement as examinees exit one system and make the transition to another, such as from high school to college. Currently, the educational standards and instructional accountability driven drumbeat outside the psychometric community intensifies for including formative assessment testing, whose primary goal is assisting the teaching and learning process *while it is occurring.*This creates the psychometric challenge of converting each examinee's test response pattern into a multidimensional student profile score report detailing the examinee's skills learned and skills needing study. The new testing paradigm calls for *a blending* of summative and formative assessments, with sometimes the same test being used for both purposes and sometimes new tests being developed for formative assessment purposes alone.

Examinee skills-based formative assessments, often aggregated over classroom, school, school district, state, or other collective, as appropriate, can be used by students, teachers, curriculum planners, government departments of education, and others involved in the educational process, to facilitate teaching and learning. In particular, using feedback from a test-based formative assessment, each student learns what skills he or she has mastered and what skills need his or her further attention. For example, a student's algebra test score is supplemented or replaced by a skills profile indicating mastery of linear equations, nonmastery of factoring, and so forth, such information hopefully leading to informed remediation. At the classroom level, the summative-assessment-based classroom test score average and standard deviation (or other measure of score variation) are supplemented or replaced by formative-assessment-based classroom proportions of masters and nonmasters for each targeted skill. Used effectively, this can lead to immediate test-driven changes in instruction for the just-tested classroom and future changes in instruction for future classrooms in the same subject. Skills-level formative assessment can also help students meet the state educational standards that are receiving so much current emphasis. The notion of a "skill" as used here is generic from the cognitive psychology perspective; it refers to any postulated mental quality whose possession improves cognitive functioning, especially in a test setting.

There is currently an enormous push in American education to periodically conduct standards-based skills diagnostic assessments of students from kindergarten to the end of high school (K–12). The biggest push comes from the U.S. government's recently passed "Leave No Child Behind" legislation, mandating periodic standards-based testing for all American K-12 students. The emphasis on formative assessment skills diagnosis is emphasized in the U.S.

Department of Education's widely referred-to regulations for this legislation: "A State's academic assessment system must produce individual student interpretive, descriptive, and diagnostic reports that... allow parents, teachers, and principals to understand and address the specific academic needs of students."

It is interesting that many of the Web sites of large city and state education departments and of test providers claim to be doing skills-level assessment. The statistical sophistication of existing psychometric approaches such as the Robert Mislevy, Russell Almond, and Linda Steinberg Evidence Centered Design/Bayes Nets approach; the Lou DiBello, Sarah Hartz, Louis Roussos, and William Stout Bayes Unified Model Markov Chain Monte Carlo (MCMC) approach; or the Mark Wilson multidimensional graded-response Rasch Berkeley Evaluation and Assessment Research System (BEAR) approach raises the question of how so many organizations responsible for formative assessments can make the striking claim of their doing such assessments, given that they have not employed such sophisticated techniques as mentioned above. The answer is simple: Most are merely using skills based *subscoring.*

For example, let's examine North Carolina's end-of-grade test reporting for eighth grade mathematics, shown in Figure 7. Each of the 80 math questions is categorized as measuring one particular skill out of eight specified math skills, such as "word problems." Thus, a subtest of items is associated with each skill. A student's reported skills-mastery profile consists of his or her skill subscore on each of the eight targeted skills. Others (e.g., Missouri) avoid reporting subscores altogether by use of proficiency scaling, where one's overall test score is translated into a skills-mastery profile, based on the (often controversial) notion that one's unidimensional score location can be used to effectively create a profile of skills mastered and skills nonmastered. From the psychometric perspective, it is clear that in most skills diagnostic settings, more sophisticated approaches than skills-based subscoring or proficiency scaling are needed to avoid serious sub-optimality in skills diagnostic accuracy.

4.1. A Brief Survey of Psychometric Skills Diagnostic Models

This new and evolving field had its early and important psychometric modeling pioneers, including Gerhardt Fischer, Edward Haertel, Robert Mislevy, Susan Embretson, and Kikumi Tatsuoka. I'd like to indicate a few milestones in the psychometric history of cognitive modeling. The first skills-level (or "cognitive") psychometric model seems to be Gerhardt Fischer's (1973) linear logistic trait model (LLTM). Although LLTM is a unidimensional Rasch IRT model and as such is not designed to model, and hence support measurement of, multiple examinee skills, it does factor item difficulties into skills-based components, thereby providing a skills-based structural IRT model.

In an effort to develop a skills diagnosis modeling approach, which *a fortiori* postulates that examinees possess a multidimensional latent structure, Susan Embretson developed a series of multidimensional continuous trait (each such trait viewed as a skill) noncompensatory (in fact, conjunctive) logistic IRT models (Whitely, 1980; Embretson, 1984, 1985). Susan Embretson's diagnostic models are probably the first models capable of undergirding multidimensionality-based skills profiling using examinee test data.

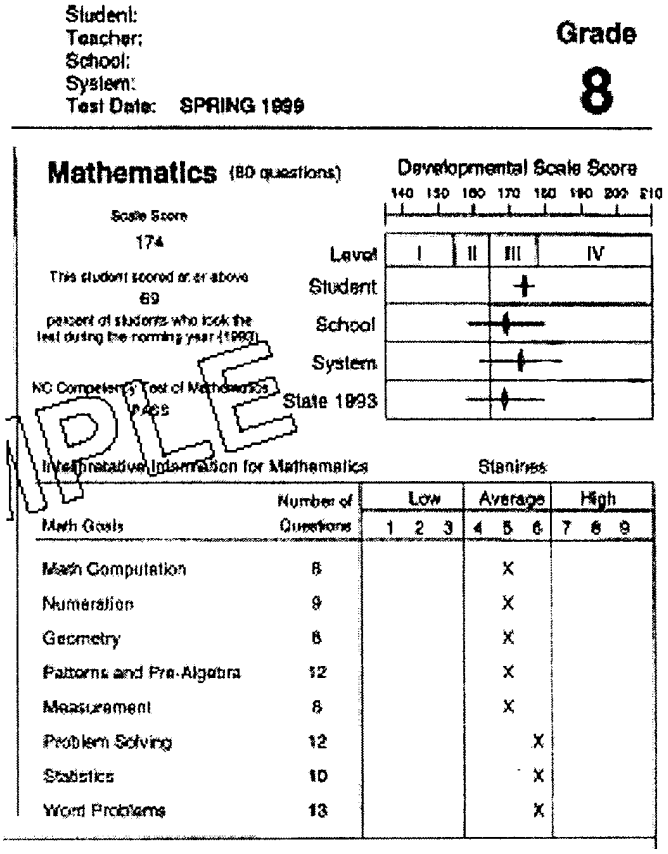

FIGURE 7.
North Carolina End-of-Grade Math Skills Test Subscores.

The important realization that the item/skill test structure (i.e., which items measure which skills) is important and inferentially useful information that should be built into one's skills diagnostic model was fully exploited by Kikumi Tatsuoka (1990, 1995). Her Rule Space pattern recognition approach dichotomizes skills as mastered (assigned a latent skill level of $\alpha = 1$) versus nonmastered (assigned a latent skill level $\alpha = 0$), thus forming a latent skill vector $\boldsymbol{\alpha}$ with dichotomous components. The method requires a user-supplied list of skills judged to be needed for performing well on the test being analyzed. This item/skill structure (usefully thought of as the skills-level test design) is represented by a user-supplied item by skill $\mathbf{Q}$ *matrix* of 0's and 1's, where the 1's in the i-th row identify which of the user-provided list of skills are needed for solving the i-th item. The underlying cognitive processing assumption of the Rule Space and certain other diagnostic models using the $\mathbf{Q}$ matrix approach (such as the Unified Model discussed below) is that correct examinee responding is *conjunctive* but random: The examinee noisily tends to get an item right if he or she has mastered *all* the skills identified by the $\mathbf{Q}$ matrix as required for the item and otherwise noisily tends to get the item wrong. Although it is a point of research contention whether conjunctive modeling is universally appropriate for

cognitive processing, it seems appropriate for many, perhaps most, skills-level test settings.

The most basic conjunctive modeling approach using such item/skill structural information is Edward Haertel's (1989) restricted latent class model, which assumes local independence, conditional upon an examinee's discrete latent class, and dichotomizes each examinee relative to each item by assigning two probabilities of correct item response, namely $g_i < 1 - s_i$, where g_i is the probability for an examinee who is a nonmaster of at least one of the required skills for item i and $1 - s_i$ is the probability for an examinee who has mastered all the required skills for the item. Here g is often called a guessing probability and s a slip probability.

The Rule Space model uses a continuous two-dimensional (θ, ζ) representation to facilitate inference about the examinee's latent skill vector $\boldsymbol{\alpha}$. Rule Space data fitting is accomplished by augmenting a standard unidimensional IRT statistical approach. In fact, inference about examinee skill mastery/nonmastery profiles are reduced to inferences based on an examinee's standardly estimated latent logistic model-based $\hat{\theta}$ and a closely related, continuous "caution index" $\hat{\zeta}$, which measures how Guttman-response atypical (overly inconsistent or overly consistent) the examinee response pattern is for an examinee of estimated ability $\hat{\theta}$. The basic assumption is that these two continuous indices can be used to estimate an examinee's latent skill vector $\boldsymbol{\alpha}$ effectively.

A version of the Rule Space approach developed by ETS scientists, especially Lou DiBello and Kikumi Tatsuoka, is in current operational use on the PSAT and as such seems to be the first psychometrically sophisticated large-scale standardized-test-based skills diagnostic application, a major pioneering milestone, both from the psychometric and the formative assessment perspectives.

For example (see Figure 8), an examinee taking the Math PSAT might be given the following description of skills to improve on: "organizing and managing information to solve multistep problems" and "applying rules in algebra and geometry." In fact, every PSAT test-taker is given advice on improving up to three skills identified by ETS's Rule Space algorithm. This major advance by ETS scientists, undergirded by the Tatsuoka Rule Space research, is an example of practically important and intellectually challenging psychometric research generated by an important test applications problem. In the context of this article, it is a particularly nice example of a "back again" success!

The Bayes net approach to skills diagnosis is the product of an extensive team-oriented research effort that has been vigorously put forward by Robert Mislevy and colleagues (especially Russell Almond and Linda Steinberg). An excellent source is Bob Mislevy's (1994) Psychometrika Society presidential address paper. This Bayes net approach to cognitive diagnosis has been applied in a variety of applied diagnostic settings, including the assessment of problem-solving skills of dental hygienists (Mislevy, Almond, Yan, & Steinberg, 1999) and airplane repair training.

A very recent exposition, which includes a description of Bayes nets, appears in Mislevy, Steinberg, and Almond (2002). The paper places a heavy emphasis on evidence-centered design (ECD). ECD is a principled approach to test design (and to assessment in general) that combines student modeling, evidence modeling, and task modeling, thus creating a model-based approach to complex assessments, as all skills diagnosis tests are. It is worth noting that ECD is a general approach, which does not suppose a Bayes net statistical analysis.

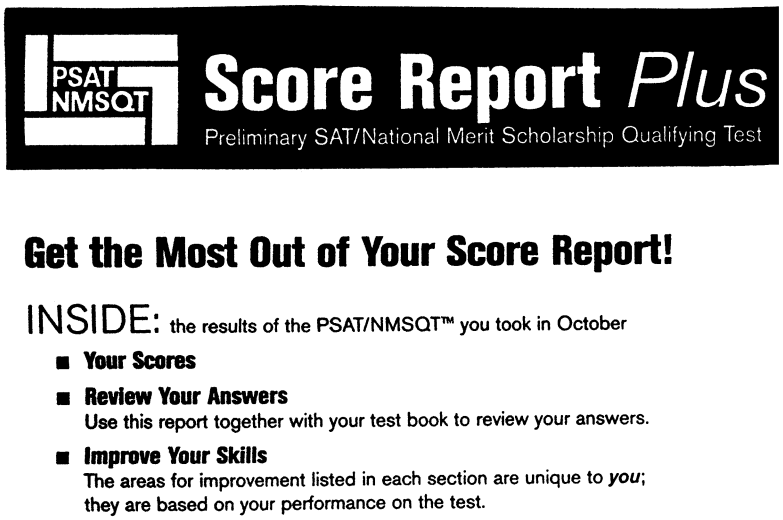

FIGURE 8.
PSAT Score Report *Plus* Skills Mastery Reporting.

It is useful and instructive to translate the three ECD modeling components into an IRT latent modeling setting. The *student model* becomes the possibly complex and hierarchical stochastic model for the latent ability space $p(\boldsymbol{\theta})$. The task model reduces and quantifies possibly rich and complex examinee responding to items (a task becoming an item) to a useful scoring metric. The last and most important component is the evidence model, which becomes the likelihood-based stochastic modeling link between the student model variable θ and the task modeled response variable $\mathbf{x}$, namely $p(\mathbf{x}|\boldsymbol{\theta})$. The overall goal of ECD from a psychometric modeling perspective is to link $\mathbf{x}$ to $\boldsymbol{\theta}$ as informatively as possible, namely, to provide a highly informative $p(\boldsymbol{\theta}|\mathbf{x})$ distribution. The ECD paradigm is expected to help guide future skills level formative assessment research and practice.

The Bayes net approach represents a joint probability latent cognitive model as a carefully sequenced recursive product of conditional probabilities, where the order of multiplied factors is carefully chosen to simplify the model representation as much as possible through local independence assumptions:

$$\begin{aligned} p(X_1, X_2, \ldots, X_n) = {} & p(X_n|X_{n-1}, X_{n-2}, \ldots, X_1) \\ & \times p(X_{n-1}|X_{n-2}, X_{n-3}, \ldots, X_1) \cdots p(X_2|X_1)p(X_1) \end{aligned} \tag{13}$$

Graph theory (in particular directed acyclic graphical networks) is used as a tool for discovering and representing the conditional dependencies and the simplifying conditional independences required to simplify (13).

Another important psychometric approach worth recognizing, especially because of the "back again" applications orientation of this paper, is the Berkeley Evaluation and Assessment Research (BEAR) embedded-assessment system using a graded-response multidimensional Rasch approach and under the leadership of Mark Wilson. This system has been successfully applied in precollege

science learning settings and draws upon the Rasch modeling tradition (see http://bear.berkeley.edu/pub.html for a long list of relevant papers).

It is interesting to note that there are approaches to skills diagnosis that are more expert systems and algebraic in their flavor than psychometric probability modeling. As a nice example, Jean-Caude Falmagne and colleagues have developed the ALEKS (Assessment and Learning in Knowledge Spaces) system, which provides course-based diagnostic assessments in a variety of subject areas, including algebra, trigonometry, and basic statistics (available from McGraw Hill and the ALEKS Corporation). For an exhaustive treatment of the complex combinatorial algebraic approach behind the ALEKS course assessment system, see Doignon and Falmagne (1999).

Two excellent surveys on the contributions of statistics and psychometric modeling to skills-level assessment, both surveys underscoring the importance of psychometric modeling of skills diagnosis, are (the first survey) Chapter 4 of the National Research Council's *Knowing What Students Know*, which in fact is based in part on a (the second survey) more technical survey by Brian Junker (1999). A nice foundational paper about skills-level latent modeling by Eric Maris (1995) is also highly recommended.

4.2. The Unified Model and Generalizations Making it Useful

Lou DiBello, Louis Roussos, and I set out to develop a **Q**-matrix-based parametric IRT skills diagnosis model with easily interpretable parameters for both users and psychometricians. Further, its parameters were expected to describe the major sources of stochastic departures from the deterministic pattern-recognition-responding predicted by the **Q** matrix. Now, I present a simplified version of the Unified Model (UM) we developed (see DiBello, Stout, & Roussos, 1995).

The UM characterizes each examinee by his or her multidimensional latent ability ($\boldsymbol{\alpha}$, θ) in the traditional sense that LI given latent ability ($\boldsymbol{\alpha}$, θ) is assumed. $\boldsymbol{\alpha}$ is a dichotomous latent vector characterizing an examinee's mastery profile on the skills specified as important by the user. As such, estimating $\boldsymbol{\alpha}$ is the goal of any skills diagnostic assessment.

From the statistical perspective, the amount of statistically recoverable skills-level information about examinees provided by a test of dichotomously scored items is by its intrinsic nature limited. Hence, if a skills-based psychometric model, such as the UM is to be effectively used to undergird a statistically effective skills diagnosis, the model must obey the principle of statistical parsimony and thus introduce only a statistically reasonable number of skills to be assessed. Hence, it is obvious that, in contrast to traditional cognitive psychology modeling of human intellectual performance (where there is no penalty for model complexity: see Koedinger and Maclaren, 2002, for an example of a complex cognitive model developed in pursuit of a deeper understanding of human problem-solving behavior in the "early algebra" stage), many of the skills influencing item performance must be intentionally left out of the psychometric model.

It is this profoundly simple insight, viewed from our perspective as psychometricians, that mandates a certain parametric simplicity of our skills diagnostic models. Indeed, the skills diagnosis modeling challenge is to achieve as much parametric model complexity as possible using interpretable, informative, and useful (for the practitioner) parameters, while still retaining statistical and computational tractability.

To incorporate explicitly the influence of the intentionally omitted (and usually numerous) skills upon item responding, the UM introduces a residual continuous ability θ that unidimensionally summarizes the examinee ability level on the large collection of skills left out of the UM. That is, θ compensates for the cognitive-processing *incompleteness* of the specified skills vector $\boldsymbol{\alpha}$.

Although in the UM we dichotomize a skill as being mastered or nonmastered by an examinee, the UM was developed from the philosophical position that an examinee can be a nonmaster of a skill but still (noisily) apply it correctly in attempting to solve an item. Similarly a master of a skill can (noisily) fail to apply it correctly. This phenomenon of deviating from the mastery/nonmastery patterns as predicted by $\mathbf{Q}$ for an examinee with skill vector $\boldsymbol{\alpha}$ is called *positivity*.

Although some might view positivity as merely amounting to guesses and slips, actually something cognitively more subtle and important is being modeled. That is, the presumed dichotomization of examinees into masters and nonmasters of a skill is an idealization that partially breaks down in reality. To illustrate, some algebra items may require such daunting uses of the "rules of exponentials" skill that many "masters" of the skill will fail to apply it correctly and thus answer many of these items incorrectly. Likewise, for some items requiring use of exponents, the correct application of the skill will be so routine that many "nonmasters" of the skill will tend to apply the skill correctly for these items. It is this heterogeneity of the role of examinee mastery and nonmastery required across items that leads to the positivity parameters of the UM.

Modeling positivity and incompleteness led to the UM's evidence model for an item IRF:

$$P(U_i = 1|\alpha, \theta) = \Pi_k \, \pi_{ik}^{\alpha_k} r_{ik}^{(1-\alpha_k)} P(\theta + c_i) \tag{14}$$

where the product is over the attributes k required for item i as specified by Q, $0 \leq c_i \leq 3$ is the completeness parameter, $\alpha_k = 0$ or 1 denotes skill nonmastery or mastery respectively, and, by definition,

$$\pi_{ik} = P(\text{Attribute } k \text{ applied correctly to Item } i|\alpha_k = 1), \tag{15}$$

$$r_{ik} = P(\text{Attribute } k \text{ applied correctly to Item } i|\alpha_k = 0), \tag{16}$$

and

$$P(x) = \frac{\exp(1.7x)}{1 + \exp(1.7x)}. \tag{17}$$

Here Θ is assumed to have a standard normal distribution. Note that the correct applications of different skills are assumed independent given latent ability ($\boldsymbol{\alpha}$, θ) in (14).

For an item in which the required α_k's are relatively complete in the sense that the role of other attributes as captured via θ is minor, c_i will tend to be large (close to 3). Clearly an item with small r's, large π's, and a large c will be a highly discriminating item in its capacity to separate masters from nonmasters of the skills required by the item, and hence the item will be highly desirable for skills diagnostic purposes.

The positivity and completeness item parameters of the UM promise to be quite useful in providing a model to be used for skills diagnosis test performance evaluation and just as

importantly in providing a model to be used for skills-level test design purposes, just as the difficulty and discrimination parameters of logistic IRT models are useful for conventional unidimensionally scaled tests. This capability is very important because skills-based testing is a "whole new ball game" for which principles of item construction, test design, and test performance evaluation are totally undeveloped. In particular, the impressive body of knowledge about item construction, test design, and test performance evaluation for unidimensionally scaled summative tests may not be very transferable to the multidimensional discrete-skills formative assessment test setting. In summary, having estimable parameters that measure diagnostic effectiveness at the fine-grained item/skill level lays a sound foundation for evaluating skills-level test performance and for designing skills level tests.

Unfortunately, the 1995 version of the UM founders on the shoals of the credit/blame problem in that examinee item level correct/incorrect response data is intrinsically not rich enough in information to render all the π's and r's of (14) identifiable. In her thesis, Sarah Hartz (2002) reparameterizes the UM in a manner that produces identifiable and still nicely interpretable and highly useful parameters, these parameters continuing to quantify the key concepts of incompleteness and positivity.

Hartz also recast the UM as a hierarchical Bayes model and wrote an MCMC algorithm to successfully calibrate the resulting reparameterized UM. This was a major required step forward, making the model statistically tractable with highly interpretable parameters, useful to the practitioner designing or psychometrically evaluating the effectiveness of skill diagnostic tests. Henceforth, the reparameterized Bayes UM of Hartz will be referred to as the "Bayes UM."

Simulation studies by Hartz (2002) showed effective estimation of the model parameters by her MCMC procedure. In a 1500 examinee, 7 skills, 40 item, 2 skills/item on average setting with highly skills-discriminating items assumed ("*strong cognitive structure*"), the correct classification rate achieved by the MCMC algorithm averaged 95% across the seven skills with only about 4% of the examinee/skill pairs left unclassified on average across the seven skills because of lack of statistical information. Further, the evidence was strong that MCMC convergence to the stationary distribution for the non burn-in portion of the generated Markov chain had occurred. Interestingly and appropriately, when a weak cognitive structure setting (low item discrimination of skills) was simulated, then the classification accuracy dropped to 84% and the proportion of examinees not classified rose to 14%. Item parameters still tended to be well estimated.

One of the biggest concerns was that the prerequisite specification of the $\mathbf{Q}$ matrix might produce a high degree of nonrobustness. That is, if the $\mathbf{Q}$ matrix was inaccurate by a minor amount (as it surely must be!), then the diagnostic accuracy could be seriously compromised. Indeed, we tend to view a user-supplied $\mathbf{Q}$ matrix as a hypothesis, setting up a confirmatory situation where good statistical practice would allow data-driven minor modifications of the $\mathbf{Q}$ matrix.

With this in mind, Hartz carried out a number of simulation studies where the $\mathbf{Q}$ matrix was incorrectly specified. In all cases, the deterioration in skills classification accuracy was appropriately minor. Moreover, in an interesting finding, the presence of θ in the UM compensated nicely for 0s erroneously placed in the MCMC's presumed $\mathbf{Q}$ matrix—see Hartz (2002) for details. Further, the data-driven parameter-reduction statistical approaches built into the MCMC Bayes UM analysis, designed to remove unneeded parameters, nicely removed the

parameters corresponding to 1s erroneously placed in the MCMC's presumed **Q** matrix. The terms "placed erroneously" and "incorrectly specified" mean that the (correct) **Q** matrix entry of the simulation model generating the data differed from the corresponding entry in the **Q** matrix (incorrectly) presumed by the MCMC statistical analysis of the simulated data.

4.3. Application of the Unified Model to PSAT Data

Of course, it is a sort of psychometricians' joke that one can prove almost anything using simulation studies. Hence, we were very pleased when ETS presented us with the opportunity to try out the Bayes UM on experimentally obtained PSAT test/retest data. For example, the PSAT math test had 40 items and a **Q** matrix based on 16 skills (these skills carefully developed for ETS by teachers, educational psychologists, psychometricians, and cognitive scientists), having approximately three skills per item.

Of particular interest is the Bayes UM methodology's capacity to assess skill, and item, level performance aspects of a test being used for skills diagnostic purposes. Indeed, the Bayes UM parameters assess the discrimination of every item/skill pair. Thus, an item's effectiveness across the set of skills it is purported to measure, and a skill's effectiveness across the set of items purported to require the skill, can both be assessed for all the items and user-specified skills. To illustrate, according to the Bayes UM analysis of the PSAT Math test conducted using the Hartz MCMC approach, three items on one math form failed to display a useful amount of skill discrimination on any of the skills the **Q** matrix claimed they were measuring.

This result is not surprising in that the PSAT is designed to scale examinees on mathematics achievement unidimensionally (it is a summative assessment), rather than being designed to diagnose the 16 post-test created mathematics skills. Thus, these three items were contributing to the desired unidimensional scaling by measuring mathematics skills other than those specified by the **Q** matrix. From the skills test-design perspective, where one might want to modify a test to improve its diagnostic power on a set of specified skills, such skills-level information about particular items would be highly useful. The capability to assess test performance for each skill and for each item will be very useful for future skills-level test design, where one wants assurances that each targeted skill is being effectively measured.

The ETS PSAT setting was an experimental setting that went beyond operational data by providing retest data. This allowed the use for research purposes of the performance criterion of skill assessment agreement across the two tests each student took. That is, skill assessment agreement occurs when examinees are independently assessed by both tests to have mastered a skill or to have not mastered that skill. Although it is not the impossible gold standard of observing how reliably examinee/skill pairs are *correctly diagnosed*, it is an excellent stand-in. Naïvely, one might decide that anything above 50% agreement (the naively presumed chance agreement rate) across tests would provide evidence that the procedure is effectively classifying examinee skill masteries based on PSAT examinee performances. But, this is wrong of course: the correct baseline rate depends on the baseline rates of assigning skill mastery on each of the two tests (for example, 100% agreement across tests can be obtained by artificially labeling every examinee a master of all skills on both tests): Let p_i denote the assigned proportion of masters of

a skill on Test i. Then the baseline chance agreement rate (assigning (0,0) or (1,1)) is

$$b \stackrel{\text{def}}{=} p_1 p_2 + (1 - p_1)(1 - p_2). \tag{18}$$

Let a be the observed across-test agreement rate proportion for an skill. Then the percentile rank of a relative to the chance rate should be a good adjusted agreement rate index:

$$100\frac{a - b}{1 - b} \tag{19}$$

Using this index, one version of the Bayes UM analyses achieved a percentile index of 70% almost uniformly across the 12 skills (out of a possible 16) judged to be effectively measured-very respectable for a test not designed for skills diagnosis.

One could wonder whether labeling an examinee as a master versus a nonmaster of the entire set of skills required on an item is consistent with examinee performance on the item, in the sense that masters should perform well and nonmasters poorly. For, if either the Bayes UM fails to fit the data well or, even when it does, the ensuing statistical MCMC analysis fails to calibrate the model well or to use the well-calibrated model effectively to predict examinee skills well, then one would expect the statistically inferred masteries and nonmasteries of examinee skills to fail to be strongly consistent with examinee correct/incorrect performance on the item.

In this regard, for each item/examinee combination, we classified the combination as an *item mastery* or an *item nonmastery combination* depending on whether or not the MCMC analysis assessed the examinee as having mastered all the skills the **Q** matrix shows as required for the item. Further, the nonmastery category is split into two subcategories depending on whether an examinee nonmaster has mastered less than half or at least half of the required skills for the item, producing *high and low nonmastery*combinations. In this regard, analysis of both forms of the PSAT math and both forms of the PSAT writing test produced excellent results. For example, on the V2 math form, item mastery resulted in an 85% item/examinee combination correct rate, and item nonmastery in a 27% item correct rate, which split into 42% and 18% for high nonmaster and low nonmaster/examinee combinations respectively.

Clearly, examinees estimated to be item masters are performing very well on items and those estimated to be item nonmasters are doing relatively poorly, with low nonmasters doing extremely poorly. This result is very encouraging. It is relevant to ask whether such behavior holds up item by item. In this regard it is interesting to contrast two simulation studies performed by Hartz (2002), one assuming strong cognitive and one with weak cognitive structure. Even the weak cognitive structure case effectively produced high-performing masters and low-performing nonmasters for all the items in the simulated data, with reasonable variation from item to item. The results for the strong cognitive structure were even better.

Of particular interest (the "back again" perspective of this paper's title) was the real-data-based reenactment of the actual College Board PSAT student reporting process described above, where each student taking the PSAT has up to three nonmastered skills reported. This was carried out on the ETS PSAT data set using the Bayes UM model as calibrated from PSAT data via the MCMC statistical inference approach. The results were especially satisfying. About 90% of the test takers had at least one nonmastered skill "reported".

Among these examinees identified as lacking at least one needed skill, correct performance proportions (using the actual test data) on items involving skills examinees were judged to have not mastered was around 0.21 on average over the four tests (two math forms and two writing forms), a pleasingly low number. The most authentic aspect was that cross-validation was carried out, in the sense that we recorded the proportion of times a skill that was "reported" as nonmastered by the examinee on one test was inconsistently "reported"as mastered on its paired test. These two error-rate proportions (viewing each test of a test/retest pair as the report-generating test produces two numbers) for the PSAT Writing test were both around 0.06. On the PSAT Math test, for whatever reason, the resulting pair of test/retest error rates were still very good but less so, averaging 0.13.

The research on skills diagnosis seems the most compelling example of our research paradigm of going from practice (in this case the enormous need in the testing arena for sophisticated skills diagnostic IRT modeling and consequent statistical analyses of standardized test data to produce accurate skills diagnoses and effectively designed skills-level tests) to theory (the Unified Model of DiBello, Stout, and Roussos and its Bayes reparameterized form and MCMC procedure for data analysis using the Bayes UM of Hartz) and back again (the very satisfying and potentially useful analysis of the PSAT data).

4.4. Skills Diagnosis: The New Paradigm?

Is formative assessment skills diagnostics the new test paradigm, as I suggested in the opening paragraph of Section 4? I think so, and I hope the brief and somewhat informal description in this section convinces my psychometric colleagues of this claim. Further, I am optimistic that the Bayes UM MCMC approach will play a significant role, according to the stated criterion at the beginning of this paper that psychometric research is valuable when it is effectively used in actual educational practice. Of course, many vital research challenges remain in the skills-level formative assessment arena.

5. Dimensionality, Equity, and Diagnostic Software

The purpose of this section is to give a partial listing of who to contact concerning methodology software for procedures mentioned in the article. The conditional covariance based dimensionality assessment software, namely DIMTEST, HCA/CCPROX, DETECT, and CONCOV, is available from Assessment Systems Corporation (http://www.assess.com). The Mokken multidimensional scaling software is available from iecProGAMMA (http://www.gamma.rug.nl. The TESTGRAF nonparametric IRF and examinee ability estimation software is available through (http://www.psych.mcgill.ca /faculty/ramsay/TestGraf.html). The multiple variations of SIBTEST are available from Assessment Systems Corporation (http://www.assess.com). Mantel-Haenszel DIF software is available from many sources and in fact the MH DIF procedure is easy to program if needed. The MCMC Bayes UM skills diagnostic software is the property of the Educational Testing Service.

6. Concluding Remarks

I wrote the presidential address with several themes in mind. First, I wanted somewhat informally to survey progress made in three major research areas that I and former Statistical Laboratory for Educational and Psychological Measurement members worked on over many years. These areas have not only been intellectually interesting areas to work on from the psychometric perspective, but I see them as very important from the applied measurement perspective. That is, achieving the "back again" of the title is seen as truly important for each area. In summary, the three areas are (i) nonparametric latent structure modeling and dimensionality assessment from the nonparametric perspective, with the emphasis being on the use of item-pair conditional covariances, the conditioning being on an appropriately chosen subtest score, (ii) the nonparametric modeling and assessment of test fairness with the emphasis being on uniting substantive and statistical approaches to test equity via the MMD multidimensional model for DIF/DBF/DTF and on the SIBTEST family of DIF/DBF/DTF procedures growing out of the MMD model, and (iii) the parametric skills diagnostic IRT modeling and associated Bayes Unified-Model-based skills diagnostic methodology created by Sarah Hartz for carrying out skills-level formative assessment and skills-level test design.

Second, I wanted to suggest one appealing, and I think, effective, approach for selecting and carrying out psychometric research problems. This consists of selecting one's research problems as motivated by important applied educational testing and assessment research challenges, then bringing sophisticated probabilistic modeling and modern statistical thought to bear upon the research problems selected, and last, making the effectiveness of the research in improving educational measurement practice the ultimate criterion for judging its worth.

Third, I wanted to stress the great power and enjoyment that can result when a team of dedicated and talented researchers, as I have had the great privilege to be associated with regarding the work described in this presidential paper, cooperatively and collegially attacks carefully selected research problems in a determined manner. Of course, other examples of the team-oriented approach to psychometric research exist: The L.L. Thurstone Laboratory under the direction of David Thissen provides an excellent contemporary example of a group environment producing important psychometric research (see *Test Scoring*, 2001, for a body of psychometric research motivated by applied measurement problems and produced largely by members of the Thurstone Lab).

Fourth, I wanted to stress that there is a major paradigmatic shift broadening the nature of large scale educational testing and assessment that I personally believe has major implications for future psychometric IRT research. That is, the summative assessment paradigm for testing is being supplanted by a new blended summative assessment and formative assessment paradigm. For those interested in catching this research train, it promises to be an exciting and eventful trip, both because it is intellectually fascinating and challenging to psychometricians and statisticians and because it is of great importance to the future of education and training. In this regard, ETS's pioneering Score Report Plus Report developed for the College Board's PSAT Exam is truly the tip-of-the-iceberg.

References

Ackerman, T.A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*, 67–91.

Angoff, W.H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning*(pp. 3–24). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bolt, D., Froelich, A.G., Habing, B., Hartz, S., Roussos, L., & Stout, W. (in press). *An applied and foundational research project addressing DIF, impact, and equity: With applications to ETS test development* (ETS Technical Report). Princeton, NJ: ETS.

Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: an adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, *33*, 333–353

Chang, H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, *58*, 37–52.

DiBello, L., Stout, W., & Roussos, L. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. Nichols, S. Chipman, & R. Brennen (Eds.), *Cognitively diagnostic assessment* (pp. 361–389). Hillsdale, NJ: Earlbaum.

Doignon, J.-P., & Falmagne, J.-C. (in press). *Knowledge spaces*. Berlin Springer-Verlag.

Dorans, N.J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355–368.

Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika*, *62*, 7–28.

Douglas, J.A. (2001). Asymptotic identifiability of nonparametric item response models. *Psychometrika*, *66*, 531–540.

Douglas J.A., & Cohen A. (2001). Nonparametric ICC estimation to assess fit of parametric models. *Applied Psychological Measurement*, *25*, 234–243.

Douglas, J., Kim, H.R., Habing, B., & Gao, F. (1998) Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics*, *23*, 129–151.

Douglas, J., Roussos, L., & Stout, W., (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF. *Journal of Educational Measurement*, *33*, 465–484.

Douglas, J., Stout, W., & DiBello, L. (1996). A kernel smoothed version of SIBTEST with applications to local DIF inference and unction estimation. *Journal of Educational and Behavioral Statistics*, *21*, 333–363.

Ellis, J.L., & Junker, B.W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika*, *62*, 495–524.

Embretson (Whitely), S.E. (1980). Multicomponent latent trait models for ability tests *Psychometrika*, *45*, 479–494.

Embretson, S.E. (1984). A general latent trait model for response processes. *Psychometrika*, *49*, 175–186.

Embretson, S. E. (Ed.). (1985), *Test design: Developments in psychology and psychometrics* (pp. 195–218, chap. 7). Orlando, FL: Academic Press.

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

Froelich, A.G., & Habing, B. (2002, July). A study of methods for selecting the AT subtest in the DIMTEST procedure. Paper presented at the 2002 Annual Meeting of the Psychometrika Society, University of North Carolina at Chapel Hill.

Gierl, M.J., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice*, *20*, 26–36.

Gierl, M.J., & Khaliq, S.N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests. *Journal of Educational Measurement*, *38*, 164–187.

Gierl, M.J., Bisanz, J., Bisanz, G.L., & Boughton, K.A. (2002, April). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the DIF analysis framework. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Haberman, S.J (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, *5*, 815–841.

Habing, B. (2001). Nonparametric regression and the parametric bootstrap for local dependence assessment. *Applied Psychological Measurement*, *25*, 221–233.

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

Hartz, S.M. (2002). *A Bayesian framework for the Unified Model for assessing cognitive abilities: blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign, Department of Statistics.

Holland, P.W. (1990a). On the sampling theory foundations of item response theory models. *Psychometrika*, *55*, 577–601.

Holland, P.W. (1990b). 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 variable models. *The Annals of Statistics*, *14*, 1523–1543.

Holland, W.P., & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Earlbaum Associates.

Jiang, H., & Stout, W. (1998). Improved Type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioral Statistics*, *23*, 291–322.

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

Junker, B.W. (1999). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment*. Prepared for the National Research Council Committee on the Foundations of Assessment. Retrieved April 2, 2001, from http://www.stat.cmu.edu/∼brian/nrc/cfa/

Junker, B.W., & Ellis, J.L. (1998). A characterization of monotone unidimensional latent variable models. *Annals of Statistics*, *25*(3), 1327–1343.

Junker, B. W. & Sijtsma, K. (2001). Nonparametric item response theory in action: an overview of the special issue. *Applied Psychological Measurement*, *25*, 211–220.

Koedinger, K.R., & MacLaren, B.A. (2002). Developing a pedagogical domain theory of early algebra problem solving (CMU-HCII Tech. Rep. 02–100). Pittsburgh, PA: Carnegie Mellon University, School of Computer Science.

Li, H. & Stout, W. (1996). A new procedure for detecting crossing DIF. *Psychometrika*, *61*, 647–677.

Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent models* (pp. 263–275). New York, NY: Plenum Press.

Linn, R.L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 349–364). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F.M. (1980) *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates, Hinsdale, NJ.

McDonald, R.P. (1994). Testing for approximate dimensionality. In D. Laveault, B.D. Zumbo, M.E. Gessaroli, & M.W. Boss (Eds.), *Modern theories of measurement: Problems and issues* (pp. 63–86). Ottawa, Canada: University of Ottawa.

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

Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika*, *59*, 439–483.

Mislevy, R.J. Almond, R.G., Yan, D., & Steinberg, L.S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K.B. Laskey & H. Prade (Eds.), *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (pp. 437–446). San Francisco, CA: Morgan Kaufmann.

Mislevy, R., Steinberg, L. & Almond, R. (in press). On the structure of educational assessments. *Measurement: Interdisciplinary research and perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Molenaar, I.W., & Sijtsma, K. (2000). *User's manual MSP5 for Windows: A program for Mokken Scale Analysis for Polytomous Items. Version 5.0* [Software manual]. Groningen: ProGAMMA.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement*, *30*, 293–311.

Nandakumar, R., & Roussos, L. (in press). Evaluation of CATSIB procedure in pretest setting. *Journal of Educational and Behavioral Statistics*.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, *18*, 41–68.

O'Neill, K.A., & McPeek, W.M. (1993). Item and test characteristics that are associated with differential item functioning. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255–276). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pellegrino, J.W., Chudowski, N., & Glaser, R (Eds.). (2001). *Knowing what students know: The science and design of educational assessment* (chap. 4, pp. 111–172) Washington, DC: National Academy Press.

Philipp, W. & Stout, W. (1975). *Almost sure convergence principles for sums of dependent random variables* (American Mathematical Society Memoir No. 161). Providence, RI: American Mathematical Society.

Ramsay, J.O. (2000). TESTGRAF: *A program for the graphical analysis of multiple choice test and questionnaire data* (TESTGRAF user's guide for TESTGRAF98 software). Montreal, Quebec: Author. Versions available for Windows®, DOS, and Unix. The Windows® version was retreived November 11, 2002 from ftp://ego.psych.mcgill.ca/pub/ramsay/testgraf/TestGraf98.wpd

Ramsey, P.A. (1993). Sensitivity review: the ETS experience as a case study. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 367–388). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rossi, N., Wang, W. & Ramsay, J.O. (in press). Nonparametric item response function estimates with the EM algorithm. *Journal of Educational and Behavioral Statistics*.

Roussos, L., & Stout, W. (1996a). DIF from the multidimensional perspective. *Applied Psychological Measurement*, *20*, 335–371.

Roussos, L., & Stout, W. (1996b). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type 1 error performance. *Journal of Education Measurement*, *33*, 215–230.

Roussos, L.A., Stout, W.F., & Marden, J. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement*, *35*, 1–30.

Roussos, L.A., Schnipke, D.A., & Pashley, P.J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics*, *24*, 293–322.

Shealy, R.T. (1989). *An item response theory-based statistical procedure for detecting concurrent internal bias in ability tests*. Unpublished doctoral dissertation, Department of Statistics, University of Illinois, Urbana-Champaign.

Shealy, R.,& Stout, W. (1993a). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*, 159–194.

Shealy, R., & Stout, W. (1993b). An item response theory model for test bias and differential test functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning*(pp. 197–240). Hillsdale, NJ: Lawrence Erlbaum.

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

Sternberg, R.J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York, NY: Cambridge University Press.

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

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

Stout, W., Froelich, A.G., & Gao, F. (2001). Using resampling to produce an improved DIMTEST procedure. In A. Boomsma, M.A.J. van Duijn, T.A.B. Snijders (Eds.), *Essays on item response theory* (pp. 357–376). New York, NY: Springer-Verlag.

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

Stout, W., Li, H., Nandakumar, R., & Bolt, D. (1997). MULTISIB—A procedure to investigate DIF when a test is intentionally multidimensional. *Applied Psychological Measurement*, *21*, 195–213.

Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese*, *48*, 191–199.

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

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. Nichols, S. Chipman, & R. Brennen (Eds.), *Cognitively diagnostic assessment*. Hillsdale, NJ: Earlbaum. 327–359.

Thissen, D., & Wainer, H. (2001). *Test scoring*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Trachtenberg, F., & He, X. (2002). One-step joint maximum likelihood estimation for item response theory models. Submitted for publication.

Tucker, L.R., Koopman, R.F., & Linn, R.L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, *34*, 421–459.

Wainer, H., & Braun, H.I. (1988). *Test validity.* Hillsdale, NJ: Lawrence Erlbaum Associates.
Zhang, J., & Stout, W. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika*, 64, 129–152.

Whitely, S.E. (1980). (See Embretson, 1980)

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213–249.