Multidimensional Computerized Adaptive Testing in a Certification or Licensure Context

Richard M. Luecht

National Board of Medical Examiners

Multidimensional item response theory (MIRT) computerized adaptive testing, building on recent work by Segall (1996), is applied in a licensing/certification context. An example of a medical licensure test is used to demonstrate situations in which complex, integrated content must be balanced at the total test level for validity reasons, but items assigned to reportable subscore categories may be used under a MIRT adaptive paradigm to improve the reliability of the subscores. A heuristic optimization framework is outlined that generalizes to both univariate and multivariate statistical objective functions, with additional systems of constraints included to manage the content balancing or other test specifications on adaptively constructed test forms. Simulation results suggested that a multivariate treatment of the problem, although complicating somewhat the objective function used and the estimation of traits, nonetheless produces advantages from a psychometric perspective. *Index terms: adaptive testing, computerized adaptive testing, information functions, licensure testing, multidimensional item response theory, sequential testing.*

The merging of multidimensional item response theory (MIRT; e.g., Reckase, 1985) and computerized adaptive testing (CAT; e.g., Kingsbury & Weiss, 1983; Wainer, Dorans, Flaugher, Green, Mislevy, Steinberg, & Thissen, 1990; Weiss & Kingsbury, 1984) is an intriguing direction to explore (Bloxom & Vale, 1987; Miller, Reckase, Spray, Luecht, & Davey, in press; Segall, 1996). However, the problems of CAT item selection and parameter estimation become more complex in a multidimensional context. For example, unlike a unidimensional CAT, which merely administers items targeted to an examinee's location along a score scale, a MIRT CAT must locate an examinee's trait estimates on a plane or hyperplane and administer items that ideally minimize the joint estimation errors for those estimates.

Faced with these complexities, it is appropriate to ask whether MIRT CAT is worth the added complications, and Segall (1996) demonstrated that it may be. Segall compared a unidimensional CAT for nine power achievement subtests in an Armed Services Vocational Aptitude test battery to a multivariate CAT, fixing the covariance structure in the latter case so that the items in each subtest loaded on individual trait composites (i.e., oblique simple structure). He also implemented a Bayes modal estimation procedure that allowed the population covariances among the nine traits to enter into the solutions. By maximizing the determinant of the posterior variance-covariance matrix as the statistical objective function for the MIRT adaptive item selections, Segall demonstrated some detectable gains in the reliabilities of the outcome subscores when compared to simulated unidimensional CATs. The present study both cross-validated and extended Segall's work by introducing a somewhat different application, one that contends with mandatory, complex content constraints in a context related to certification or licensure testing—specifically, medical licensure.

Licensure/Certification Tests

Many professional certification or licensure tests consist of multifarious content structures covering a large domain of integrated knowledge and require applications of that knowledge to realistic problems in the profession. The multidisciplinary concepts and applications tend to be of a broad scope, in-depth, and

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 4, December 1996, pp. 389–404

0146-6216/96/040389-16$2.05

highly interrelated, comprised of items specifically written to an integrated content outline covering numerous hierarchical levels, with various combinations of crossed and nested specifications (e.g., Federation of State Medical Boards & National Board of Medical Examiners, 1996a, 1996b). However, the primary purpose of the assessment may be to produce a single total test score or decision (e.g., competency to be certified or licensed as a professional); this purpose implies that there is a single underlying statistical dimension of interest.

If the reported outcome measure is strictly univariate (e.g., a total test score or decision), there may be little apparent practical advantage in considering a complex multivariate framework that attempts to capture the salient multidimensionality of the test, subject, or course, to empirically evaluating for serious violations of the requisite assumptions underlying the use of a particular unidimensional item response theory (IRT) calibration or scaling model. For example, a test comprised of several hundred "cells" in a fully- or semi-crossed content specifications matrix would not be feasible to model by assigning latent trait dimensions for each content cell. At the same time, content experts might strongly question the validity of building test forms strictly to optimize the reliability of trait or factor scores for statistically determined factors (e.g., factors determined using exploratory factor analysis and therefore subject to arbitrary interpretation and capricious decisions about the number of factors to use, the type of rotation to use, and so forth).

That professional certification or licensure tests comprised of complex, integrated content are perhaps multidimensional is not the relevant issue. Rather, the question is whether there is any advantage to attempting to decompose the total test into arbitrary—and perhaps substantively meaningless—statistical multivariate latent structures when the most that could be accomplished would be to estimate a set of (probably unstable) coefficients or loadings for recombining the multivariate scores in some fashion to generate a total test composite score.

What is needed, therefore, is a reason to use a multidimensional statistical framework—a reason that arguably relates to reported scores. The relevant issue may be whether there exists an estimable and meaningful multivariate structure that can be used in a MIRT CAT to serve some auxiliary score reporting purpose, such as providing subscore profiles to help examinees diagnose their strengths or weaknesses or providing schools with aggregate performance feedback in what some might term *core areas.* Of course, a total test score applicable for making accurate decisions needs to be produced and the requisite content validity at the total test level needs to be maintained.

This type of dual purpose (i.e., making total test pass/fail decisions and also reporting subscores based on items that are logically assigned to various categories) is not uncommon in professional certification or licensure testing (e.g., Swanson, Case, Kelley, Lawley, Nungester, Powell, & Volle, 1991). For example, the integrated, applied content for a medical licensure test covering clinical science concepts and patient management could be inappropriately represented, from a content perspective, if the test were built strictly along traditional discipline dimensions (e.g., 50 items each covering medicine, pediatrics, surgery, obstetrics and gynecology, and so forth; see, e.g., Greenburg, Case, Golden, & Melnick, 1996). There are numerous other dimensions to cover on such a test (e.g., physician tasks, various organ systems, disease categories, sites of care).

The construction of test forms instead requires a more comprehensive system of content taxonomies and constraints to capture the breadth and depth of the clinical sciences in the medical profession (Federation of State Medical Boards & National Board of Medical Examiners, 1996b). However, examinees might still benefit from knowing how they performed on items logically related to those somewhat traditional disciplines. Medical schools might also request to use the aggregate information of their students' performance in such core areas for program evaluation or to inform aspects of the curriculum. It is suggested here that the secondary purpose of the test (i.e., generating subscores) could be used to establish a multivariate framework amenable to MIRT CAT.

It is important to make this distinction that relates the purpose(s) of the test to the scores that are reported. CAT can improve the reliability of scores using targeted item selections and heuristics for locally optimizing a particular statistical objective function (e.g., minimizing error variances); it does not necessarily improve validity in any substantive way. This perspective serves as an introduction to what will be a confirmatory view of multidimensionality and test construction following from the intended purpose of the test, the content specifications and, ultimately, the scores that are reported.

A Multidimensional Framework for Estimating Multiple Traits

A general form of the two-parameter logistic MIRT model, M2PLM, (e.g., Doody-Bogan & Yen, 1983; Reckase, 1985; Reckase & McKinley, 1983) can be expressed as

$$P\left(u_i = 1 \middle| \theta_1, \ldots, \theta_K,\, a_{i1}, \ldots, a_{iK},\, d_i\right) \equiv P_i = \frac{\exp\left(\mathbf{a}_i^{\mathrm{T}} \boldsymbol{\theta} + d_i\right)}{1 + \exp\left(\mathbf{a}_i^{\mathrm{T}} \boldsymbol{\theta} + d_i\right)}, \tag{1}$$

where

u_i is a dichotomously scored response to item i, $u_i \in \{0,1\}$, $i = 1, \ldots, n$;

$\boldsymbol{\theta}$ is a $K \times 1$ vector of latent traits (i.e., the traits in K dimensions);

$\mathbf{a}_i$ is a $K \times 1$ vector of item discriminations or coefficient loadings of item i on the latent traits, $\boldsymbol{\theta}$; and

d_i is a scalar analogous to the unidimensional location of item i.

The lower asymptote parameter, c_i, $i = 1, \ldots, n$, or c, a constant for item i, can be added to the model as a correction for assumed guessing. The simulations used here assumed no guessing and used the M2PLM; thus, no c parameter was used. Segall (1996) used the three-parameter model and the associated likelihood equations. Finally, for convenience and under the implicit assumption that items will be logically assigned to latent dimensions based on their intended inclusion in particular subscores (or other considerations), a vector of indexes, $\mathbf{v}_k$, denotes the items intended to load on each of the traits, θ_k, $k = 1, \ldots, m$.

Item Parameter Estimation

In practice, the item parameters are only estimates. Assuming that the parameters are well estimated, they can be treated as knowns in MIRT CAT. Bock (1985), Bock, Gibbons, & Muraki, (1988), and McDonald (1982) demonstrated that it is feasible to obtain reasonable estimates for the multivariate item discrimination and threshold parameters. Bock et al. presented a full information factor analytic solution that is used in the computer program TESTFACT (Wilson, Wood, & Gibbons, 1984). McDonald provided a convenient polynomial solution to approximate a cumulative normal density for a linear combination of latent multivariate traits. McDonald's solution was implemented by Fraser (1986) in the computer program NOHARM.

There is another option. As Segall (1996) demonstrated, if an oblique simple structure is hypothesized for the multidimensional space (i.e., correlated traits with items each loading on only one trait), unidimensional item parameters estimated uniquely for each trait may also be of practical use. A similar suggestion to consider using unidimensional parameter estimates to form oblique simple structure factors for items that cluster together in the multidimensional space was provided by Luecht & Miller (1992).

θ Estimation

The maximum likelihood estimators (MLEs) of the multidimensional traits, $\boldsymbol{\theta}$, closely resemble their unidimensional counterparts (e.g., Lord, 1980). Under the usual assumption of local independence, the multidimensional likelihood function for the M2PLM probability function is

$$L\left(\mathbf{U} \middle| \boldsymbol{\theta}, \mathbf{a}, d\right) \equiv L = \prod_{i=1}^{n} P_i^{u_i} Q_i^{1-u_i}, \tag{2}$$

where $Q_i \equiv 1 - P_i$. Recalling that $\mathbf{v}_k$ denotes the vector of indexes for items assigned to the K latent dimensions, the MLEs can be computed by taking the natural logarithm of Equation 2, differentiating with respect to each element in $\boldsymbol{\theta}$, and solving the equations when the partial first derivatives are set equal to 0:

$$\frac{\partial \ln(L)}{\partial \boldsymbol{\theta}} = G(\boldsymbol{\theta}) = \begin{bmatrix} \sum_{i \in v_1}^{n} \frac{u_i - P_i}{P_i Q_i} \left(\frac{\partial P_i}{\partial \theta_1} \right) \\ \vdots \\ \sum_{i \in v_K}^{n} \frac{u_i - P_i}{P_i Q_i} \left(\frac{\partial P_i}{\partial \theta_K} \right) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{3}$$

where $\partial P_i / \partial \theta_k = a_{ik} P_i Q_i$. The symmetric $K \times K$ matrix of second derivatives, $\mathbf{H}(\theta)$, has diagonal elements $\partial^2 \ln(L)/\partial \theta_K^2$, and off-diagonal elements $\partial^2 \ln(L)/\partial \theta_k \theta_l$ ($k = 1, ..., K$; $l = 1, ..., K$; $k \neq l$). By evaluating the second derivatives at the same point as the first derivatives, the expectation of $\mathbf{H}(\theta)$ can be found. $\mathbf{H}(\theta)$ can be well approximated by the negative information matrix under Fisher's concept of information; that is,

$$-\mathrm{E}\left[\mathbf{H}(\theta)\right] = \mathbf{I}\left(\theta, \hat{\theta}\right) = \begin{bmatrix} \sum_{i \in v_1} \mathbf{a}_{i1}^2 P_i Q_i & \sum_{i \in v_1, v_2} \mathbf{a}_{i1} \mathbf{a}_{i2} P_i Q_i & \cdots & \sum_{i \in v_1, v_K} \mathbf{a}_{i1} \mathbf{a}_{iK} P_i Q_i \\ \sum_{i \in v_2, v_1} \mathbf{a}_{i2} \mathbf{a}_{i1} P_i Q_i & \sum_{i \in v_2} \mathbf{a}_{i2}^2 P_i Q_i & \cdots & \sum_{i \in v_2, v_K} \mathbf{a}_{i2} \mathbf{a}_{iK} P_i Q_i \\ \vdots & \vdots & & \vdots \\ \sum_{i \in v_K, v_1} \mathbf{a}_{iK} \mathbf{a}_{i1} P_i Q_i & \sum_{i \in v_K, v_2} \mathbf{a}_{iK} \mathbf{a}_{i2} P_i Q_i & \cdots & \sum_{i \in v_K} \mathbf{a}_{iK}^2 P_i Q_i \end{bmatrix}. \tag{4}$$

The information matrix is also a $K \times K$ symmetric matrix. Segall (1996) provided more complete details on the derivations. Using the information matrix, Fisher's method of scoring can be implemented instead of the more common Newton-Raphson method as

$$\hat{\boldsymbol{\theta}}^{[s+1]} = \hat{\boldsymbol{\theta}}^{[s]} + \hat{\boldsymbol{\delta}}^{[s]}, \tag{5}$$

where the superscript $[s + 1]$ represents the updated estimates of $\boldsymbol{\theta}$ and

$$\hat{\boldsymbol{\delta}} = \mathbf{I}\left(\theta, \hat{\theta}\right)^{-1} G(\boldsymbol{\theta}). \tag{6}$$

Bayes mean estimators (e.g., Bock, 1985; Bock & Aitkin, 1981; Mislevy, 1986) can also be derived by direct extension of the unidimensional estimators, but for a multidimensional trait space. The posterior distribution,

$$p\left(\boldsymbol{\theta} \middle| \mathbf{a}, d, \mathbf{U}\right) \propto L\left(\mathbf{U} \middle| \boldsymbol{\theta}, \mathbf{a}, d\right) p\left(\boldsymbol{\theta} \middle| \mathbf{u}, \boldsymbol{\Sigma}\right) p(\mathbf{u}, \boldsymbol{\Sigma}), \tag{7}$$

is merely a joint posterior for the vector-valued $\boldsymbol{\theta}$, with a multivariate probability density, $p(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$, and with population parameters, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. It is necessary to integrate over the joint posterior to obtain the Bayes mean estimators [i.e., expected a posteriori estimates (EAPs)] and the associated estimation variance-covariance errors (Bock & Aitkin, 1981).

Bloxom & Vale (1987) provided closed-form approximations to the Bayes centroids and variances and covariances by extending Owen's (1975) formulas to multivariate, orthonormal traits. Although Bloxom and Vale's approximations to the centroids were not EAPs, strictly speaking, they serve as reasonable surrogates for obtaining provisional θ estimates and can be used to speed up the estimation process if the traits are orthogonal. In general, EAP estimates are often the preferred estimators over MLEs because: (1) they are available for all response patterns, including null or perfect response; and (2) it is impossible to improve on the minimum variance properties of EAPs, on average, over a population of traits (e.g., Bock & Mislevy, 1982). However, EAPs, when used as point estimates, also tend to overestimate the true correlations between the multivariate latent traits (e.g., Luecht & Miller, 1992; Segall, 1996). In those cases, a method suggested by Mislevy (1984), using pseudocounts computed over the joint posterior for all examinees, can be used to directly obtain better estimates of the population correlation or variance-covariance matrices.

Segall (1996) also presented the solutions for the multidimensional Bayes modal estimators, which generalize to orthogonal or oblique traits for a multivariate normal prior probability density. As Segall noted, estimating the modes of the posterior distribution may be easier to do in practice than estimating EAPs when the number of dimensions is more than two or three, because they avoid the need for integration.

Item Selection Heuristics in a Multidimensional Context

As presented above, it is relatively straightforward to derive multidimensional estimators for $\boldsymbol{\theta}$; however, obtaining stable estimates in a practical CAT environment (e.g., using less-than-infinite item banks) is principally an empirical issue that depends largely on how items are selected from a given bank. CAT is fundamentally a heuristic process by which items are selected sequentially to maximize or minimize a particular objective function, subject to various content or other constraints (e.g., word counts, the test length, or any of the usual adaptive stopping rules). Heuristics are used quite often in other types of automated test assembly (e.g., Luecht & Hirsch, 1992; Stocking & Swanson, 1993) and using them for CAT is a natural extension of the technology (Luecht, in press; Stocking & Swanson, 1993). Heuristics allow each item selection to be modeled as a local optimization problem having a statistical criterion—the objective function. Equally important in selecting items are content or other constraints that follow from test specifications.

An important challenge of using heuristics to locally optimize the item selections in MIRT CAT concerns the choice of the objective function. In unidimensional IRT, a MLE is asymptotically consistent and normally distributed around the true but unknown trait, θ. Where certain regularity conditions hold (e.g., Lehmann, 1983), the variance of the maximum likelihood estimate, $\hat{\theta}$, about θ is

$$\mathrm{Var}\left(\hat{\theta}\middle|\theta\right)=\frac{1}{\mathrm{E}\left\{\left[\frac{\partial\ln(L)}{\partial\theta}\right]^2_{\theta}\right\}}. \tag{8}$$

In Equation 8, ln(L) denotes the log-likelihood function; for example, the natural logarithm of Equation 2, but for a unidimensional likelihood function. As the number of items becomes very large, the mean of the sampling distribution of estimates approaches the true parameter, θ (i.e., $\mu\hat{\theta}\,|\,\theta \rightarrow \theta$) and the information that the MLE provides about θ reduces to

$$I\left(\theta,\hat{\theta}\right)=\sum_{i=1}^{n}\frac{\left(\frac{\partial P_i}{\partial\theta}\right)^2}{P_i\left(1-P_i\right)} \tag{9}$$

for a unidimensional item response function, P_i, with known item parameters (e.g., a_i and b_i). Equation 9 is the unidimensional IRT item information function (Birnbaum, 1968). Therefore, if the item parameters are assumed to be known, maximizing Equation 9 by selecting the most informative items in CAT will minimize the asymptotic error variance in Equation 9.

The unidimensional CAT objective function is therefore defined as maximizing the IRT test information function in Equation 9 (a single-valued function computed at the current MLE), selecting the item that contributes the most item information to the sum of the item information functions. This item selection strategy assumes minimum errors of estimation as a global objective function following the traditional least-squares principle for the deviations of the estimators about the true value of the parameter, even though the item selection mechanism only yields a locally optimal decision at the current, provisional estimate of θ. Because the numerator of Equation 9 is the squared regression slope (i.e., squared partial first derivative of the response function), by maximizing the squared slope of the likelihood function the error variance of the estimates about the true parameter is minimized (Birnbaum, 1968).

Generalizing to the multidimensional case, for some vector-valued trait, **θ**, the MLEs tend toward asymptotic multivariate normality under regularity conditions (e.g., Kendall & Stuart, 1967; Lehmann, 1983). The approximate asymptotic variance-covariance matrix of the sampling distribution of estimators about the true parameters (i.e., the dispersion matrix) is the inverse of the information matrix given in Equation 4. Because the objective function in CAT usually must be reduced to a single value, a composite function for locally optimizing the item selections must be found. The information criterion is represented by a matrix.

Miller, Reckase, Spray, Luecht, & Davey (in press) reported a variety of composite functions that operate on the information matrix, each intended to maximize some aspect of the information (e.g., maximizing the trace of the information matrix). Using simulation results, they concluded that maximizing the determinant of the inverse information matrix was the most effective approach to use, although computationally less intensive procedures also produced adequate results. Bloxom & Vale (1987) and Segall (1996) similarly recommended using the determinant of the inverse information matrix as an objective function for MIRT CAT; Segall further provided a complete statistical justification for this strategy.

Assuming a sequence of i = 1, ..., n item selections (for a test length of n items), a reasonable locally optimal strategy for selecting the ith item is, therefore, to maximize the determinant of the provisional information matrix, $\mathbf{w}(\theta, \hat{\theta}_i)$; that is,

$$\mathbf{W}(\theta,\hat{\theta}_i) = \left|\mathbf{I}(\theta,\hat{\theta}_{i-1}) + \mathbf{J}(\theta,\hat{\theta}_i)\right|, \quad \mathrm{i} = 1,\ldots,n, \tag{10}$$

where $\mathbf{J}(\theta, \hat{\theta}_i)$ denotes the information matrix for the unselected items evaluated at the current values of the provisional MLEs (Segall, 1996).

However, this is only part of the heuristic solution needed in the present context. Maximizing the determinant (i.e., Equation 10) satisfies the need in MIRT CAT for a single-valued objective function; it does not incorporate the necessary content constraints that ensure that the total test will fulfill the requisite content dimensionality (i.e., meeting the content requirements in terms of the primary total test specifications).

Swanson & Stocking (1993), Stocking & Swanson (1993), Luecht & Hirsch (1992), and Luecht (in press) provide more comprehensive heuristics that can be modified for this type of sequential test construction problem. Although a complete presentation of these heuristics is beyond the scope of this paper, the total test content constraints can be built into the objective function. Luecht presents a complete optimization framework and formally demonstrates how the entire process simplifies to generating a composite objective function using coefficient terms and user-defined weights. This type of composite objective function can incorporate the statistical criteria, all the current content needs, and the availabilities in the item bank. Ancillary specifications such as word counts can also be directly integrated into the composite objective function. A less satisfying approach (although easier to implement) would be to merely set upper bounds on the item frequencies within each of the "cells" of the content outline.

A final note on item selection mechanisms concerns the use of exposure controls. Methods suggested by Sympson & Hetter (1985), with extensions offered by Stocking (1993), provide a convenient way to introduce randomization into the adaptive item selection process; this randomization precludes the same items from being selected a majority of the time for most examinees and indirectly helps to make better use of the entire item bank. In general, exposure controls use a predetermined item statistic (usually an empirically derived probability value) and a uniform random generating function to randomly determine whether the next optimally selected item should be administered to the current examinee. This procedure controls the expected frequency of use for the item within an assumed population of examinees. (See Stocking & Lewis, 1995, for a somewhat different approach to exposure controls.)

Obviously, without some type of item exposure controls in place, the security of any type of high-stakes examination would quickly be breached due to examinees memorizing and sharing the items (especially if the statistically best items were always selected). However, exposure controls can counteract the statistical optimization in a CAT. The result of implementing exposure controls, in practice, may be to produce less reliable tests than might be achieved in ideal settings in which the item banks are infinite and no examinees ever cheat. The latter point is made to emphasize that for simulation results (such as those presented below), caution should be used so as not to overinterpret statistical efficiency gains that might not be achievable to the same degree in practice.

Method

A simulation study was conducted to compare the use of MIRT CAT to a traditional unidimensional CAT for a medical licensure exam. The outcome measures from the CAT were to be used for two purposes: (1) generation of total test scores with full content balancing; and (2) production of discipline-related subscores for items assigned to eight categories.

The Examination Context, Content, and Item Bank

2,458 previously administered items from the United States Medical Licensing Examination (USMLE) Step 1 were used as the item bank for the simulation study. USMLE is a joint testing program of the Federation of State Medical Boards (FSMB) and the National Board of Medical Examiners (NBME). USMLE Step 1 is the first of three examinations required to obtain a medical license in the U.S. The purpose of Step 1 is to measure examinees' understanding of important concepts of the basic biomedical sciences, emphasizing principles and mechanisms underlying health, disease, and modes of therapy. The examination is usually taken by U.S. medical students following their second year of medical school. Foreign medical graduates must also pass Step 1 as part of the requirements for a medical license in all 50 states. A typical Step 1 is administered in paper-and-pencil format over a two-day period and has over 600 scored items.

The paper-and-pencil tests are constructed according to detailed content specifications involving several hundred constraints and are designed to statistically optimize score precision in the region of the Step 1 pass score. Step 1 total test score reliability coefficients for the U.S. first-taker group are usually in the range of .96 to .97.

Table 1 summarizes the primary categories in the Step 1 content outline (Federation of State Medical Boards & National Board of Medical Examiners, 1996a) that were used for test construction in this simulation study. For the actual full-length Step 1 examinations, the content dimensions in Table 1 would be covered to an additional depth of four or five outline levels and additional content criteria would be used. The Secondary Processes Dimension was fully crossed with Categories 2 to 11 of the primary General Principles and Organ Systems dimension. Thus, a system of constraints covering 60 independent cells had to be developed for the simulated CATs. Every constructed CAT form was forced to meet all the constraints at the total test level. The General Principles section comprised approximately 45% of each test form. The Organ Systems (Categories 2 to 11) comprised the remaining 55% of the test length. The five Processes (A to E) were constrained to be marginally represented at approximately 15%, 25%, 35%, 15%, and 15%, respectively, across the Organ Systems categories and correspond to the approximate percentages used to construct the full-length USMLE Step 1 test forms (Federation of State Medical Boards & National Board of Medical Examiners, 1996a).

Table 1
USMLE Step 1 Primary Content Dimensions for the Total Test

Dimensions and Content Descriptions
Primary Dimensions: General Principles and Organ Systems
1. General Principles
1.1 Biochemistry and Molecular Biology
1.2 Biology of Cells
1.3 Human Development and Genetics
1.4 Biology of Tissues and Their Responses to Disease
1.5 Psychosocial, Cultural, and Environmental Influences on Behavior, Health, and Disease Processes
1.6 Multisystem Processes
1.7 Pharmacodynamic and Pharmacokinetic Processes
1.8 Microbial Biology and Infection
1.9 Immune Responses
1.10 Quantitative Methods
2. Hematopoietic and Lymphoreticular Systems
3. Central and Peripheral Nervous System
4. Skin and Related Connective Tissue
5. Musculoskeletal Systems
6. Respiratory System
7. Cardiovascular System
8. Gastrointestinal System
9. Renal/Urinary System
10. Reproductive System
11. Endocrine System
Secondary Processes Dimensions
A. Normal Development and Structure, and Age-Related Changes
B. Normal Processes
C. Abnormal Processes
D. Principles of Therapeutics
E. Psychosocial, Cultural, and Environmental Considerations

Note. The USMLE Step 1 Primary Content Dimensions are reprinted with the permission of The Federation of State Medical Boards of the United States and the National Board of Medical Examiners®.

In addition to reporting a total Step 1 test score based on all scored items and making a corresponding pass/fail decision, discipline-related subscores are produced to provide examinees with feedback in a graphical profile of their strengths and weakness. Within-school means, standard deviations (SDs), and graphical summaries are also supplied to medical schools for the discipline subscores. There are currently eight biomedical science "core" areas reported as subscores: physiology, biochemistry, pathology, microbiology, pharmacology, behavioral sciences, gross anatomy, and histology. Other subscores in various organ system areas are also produced; however, these were ignored for the present study.

There is a long-standing tradition behind reporting these "core" discipline subscores, even though it is well-recognized that a Step 1 examination is far more than the simple sum of eight discipline components. In fact, items are assigned by physicians and biomedical content experts to each of the eight categories specifically for the purpose of subscore reporting, based solely on the relevance of each item for that core area and without regard to overlap of items among the disciplines (i.e., even the item counts across disciplines do not sum to the total test length).

662 items in the item bank of 2,458 items had overlapping discipline category assignments. Each item had been classified previously by content experts as having substantive relevance for the discipline. Most items were assigned to a single category; no single item was assigned to more than three categories. Upper-bound constraints were established for each of the eight discipline categories, factoring in availabilities of items in

the bank, overlapping content, and the total test length for each of the simulated CATs. These constraints were relaxed enough to allow the heuristics some freedom in selecting items, but prevented any test form from containing a disproportionately large number of items in any particular discipline area, regardless of the statistical optimization. The same constraints were used throughout the simulation.

The 2,458 items in the bank had been calibrated previously using the Rasch model (Linacre & Wright, 1994; Wright & Stone, 1979) and equated to common scales for the total test and for each of the eight disciplines (physiology, biochemistry, pathology, microbiology, pharmacology, behavioral sciences, gross anatomy, and histology). Nine item difficulty estimates were available for each item: a total test calibrated difficulty and eight within-discipline calibrated item difficulty estimates. The use of these item difficulties in the simulations is explained below. The Rasch model has been demonstrated to fit the Step 1 data quite well; thus, there was no compelling reason to use a more complicated model for the simulations. Practically speaking, it also would have been extremely difficult to use any of the existing multidimensional software described above to calibrate the response data for these examinations of 600+ items, much less attempt to link 2,458 items onto a common set of eight multidimensional scales.

Item Statistics and Data Generation for the Simulated CATs

To simulate true θs for the discipline subscores, eight multivariate normal deviates were generated for each of 2,000 simulated examinees, using an implementation of the Box-Muller algorithm (Aquinis, 1994). The total test trait composite was ignored during this data generation phase of the simulation. The means and SDs of the marginal distributions were input as .75 and .65, respectively, for all θs and corresponded closely to past performance of the Step 1 population, where the within-discipline item difficulties have a mean of 0 to fix those scales, in accordance with typical Rasch model calibration practices. Therefore, the SDs of the item difficulties and the means and SDs of the trait distribution were free to vary.

Table 2 shows the means and SDs for the sample of generated traits, as well as the correlations among the eight discipline scores. The lower triangle of the correlation matrix in Table 2 shows the input values assumed for the population. The upper triangle shows the sample-based product-moment correlations for the 2,000 vectors of generated discipline trait scores. The input values for the population correlations (lower triangle) are based on empirical interdiscipline correlations for recent Step 1 examinations, disattenuated for unreliability. As Table 2 shows, the population parameters and sample statistics for the 2,000 generated trait vectors were similar.

The Rasch difficulty estimates for the bank of 2,458 items were also treated as known parameters in generating the data and were subsequently used for estimation of the trait scores. To approximate a M2PLM (i.e., Equation 1) pseudodiscriminations or loadings were generated for each item as binary vectors; that is, $\mathbf{a}_{ik} \in \{0,1\}$, $i = 1, ..., I$, $k = 1, ..., K$, (for I items in the bank) such that $\mathbf{a}_{ik} = 1$ if the item loaded on trait θ_k or 0 otherwise. For example, an item that loaded only on the second and fourth traits would be assigned the vector, $\mathbf{a} = (0, 1, 0, 1, 0, 0, 0, 0)$. This approach is conceptually similar to using "imputed" discrimination weights under the one-parameter logistic model developed by Verhelst & Glas (1995), but in a multidimensional context (also see Glas, 1992); here, the weights were binary. Any variety of nonzero values could have been derived or estimated to use as the discrimination weights for items loading on particular discipline dimensions.

Finally, a minor reparameterization of the M2PLM was made to accommodate the use of the eight separate discipline-based difficulty estimates for each item in the simulations. That is, Equation 1 was reparameterized as

$$P_i = \frac{\exp\left[\mathbf{a}_i^{\mathrm{T}}(\boldsymbol{\theta} - \mathbf{b}_i)\right]}{1 + \exp\left[\mathbf{a}_i^{\mathrm{T}}(\boldsymbol{\theta}_i - \mathbf{b}_i)\right]} = \frac{1}{1 + \exp\left[\sum_{k=1}^{K} -\mathbf{a}_{ik}(\theta_k - \mathbf{b}_{ik})\right]}, \quad \mathrm{i} = 1, \ldots, I. \qquad (11)$$

Table 2
Sample-Based Means and SDs for the Eight Disciplines, Sample-Based Correlations for the 2,000 Vectors of Generated Discipline Trait Scores (Upper Triangle), and Input Values for the Population Correlations for the Eight Discipline Traits (Lower Triangle)

Discipline and Statistic	1	2	3	4	5	6	7	8
1. Physiology		.87	.83	.83	.83	.64	.78	.77
2. Biochemistry	.87		.79	.87	.84	.56	.83	.79
3. Pathology	.83	.78		.87	.85	.80	.89	.84
4. Microbiology	.83	.86	.87		.89	.65	.84	.79
5. Pharmacology	.83	.83	.86	.89		.69	.78	.78
6. Behavioral Sciences	.65	.58	.80	.66	.70		.64	.64
7. Gross Anatomy	.78	.83	.89	.84	.78	.64		.78
8. Histology	.76	.78	.84	.79	.78	.64	.78	
Mean	.76	.75	.75	.75	.75	.74	.76	.75
SD	.64	.65	.65	.65	.65	.65	.66	.64

Follow-up analyses with the simulated dataset indicated that unidimensional Rasch calibrations, within each discipline, were fairly accurate in recovering the item parameters used in the multidimensional data generation model. No claims are made, however, that Equation 11 would be a useful multidimensional model in practice (see, e.g., Glas, 1992). Segall (1996) used a similar approach, but used the three-parameter logistic model (Lord, 1980) with the discrimination weights fixed at their unidimensional estimates, such that each item loaded on a single trait. In this study, (1) items were individually allowed to load on multiple traits; and (2) the vectors of discriminations, $\mathbf{a}_i$, were specified rather than based on empirical estimates of multidimensional loadings. Had the items been forced to load uniquely on a single dimension, the entire maximum likelihood estimation process could be shown to simplify to solving eight independent likelihood equations.

Given the matrix of item difficulties (2,458 × 8) and the binary 2,458 × 8 matrix of loadings, **a**, Equation 11 was used to generate a 2,000 × 2,458 matrix of examinee × item P_{ji} (Equation 1) for the entire item bank ($j = 1, ..., 2{,}000$ examinees and $i = 1, ..., 2{,}458$ items in the bank). A corresponding uniform random probability, π_{ji} ($j = 1, ..., 2{,}000$ and $i = 1, ..., 2{,}458$) was also computed for each examinee × item interaction. A dichotomous score, $u_{ji} \in \{0, 1\}$, was produced such that $u_{ji} = 1$ if $P_{ji} \geq \pi_{ji}$ or $u_{ji} = 0$ otherwise. The 2,000 × 2,458 matrix of scored item responses, **u**, was used throughout the simulations.

Study Conditions, Item Selection, and Estimation of Examinee Scores

The simulation study involved two principal manipulated conditions: CAT length and item selection mechanism. Test lengths of 100, 150, 200, 250, and 300 items were used to simulate CATs ranging from approximately one-sixth to one-half the length of a typical paper-and-pencil Step 1. Two statistical optimization procedures were used as part of the item selection heuristics: (1) maximizing the item information for provisional estimates along the unidimensional total test trait composite (UTOTCAT); and (2) maximizing the determinant of the inverse information matrix under a multidimensional solution involving the eight disciplines (i.e., MIRTCAT). Maximum likelihood trait estimates were used throughout the simulations.

UTOTCAT requires some explanation. Although it should be clear that the Step 1 total test scores are used for the primary purpose of making a pass/fail decision and therefore ought to be as reliable as possible, there is a legitimate statistical rationale for optimizing the unidimensional total information that relates to improving the subscore reliabilities, as well, in a multidimensional context.

From a multidimensional perspective, maximizing the total test unidimensional information function is an indirect means of maximizing a "slice" through the information surface, in the direction of average maximum information along a "reference composite" (Wang, 1986; for a discussion of directional information and "reference composites" see also Green, 1988; Reckase, 1985; Reckase & McKinley, 1983). That is, the total test composite trait fit using a unidimensional IRT model essentially extracts a common factor (e.g., Baker, 1992; Lord, 1980). Maximizing the common information, in principle, should result in items being selected that are also in the general vicinity of the true parameters along the individual discipline trait composites, because those traits can be considered linear composites of the common and unique factors in which the unique variances might be assumed to be small for highly intercorrelated traits and ignorable for purposes of item selection.

Each CAT was constructed to meet precisely the Step 1 content specifications outlined in Table 1, proportionally reduced for each of the five test lengths (i.e., exact matches to frequencies in each of the 60 content cells described earlier), and also subject to the relaxed upper-bound constraints placed on the frequencies of items selected for each of the eight disciplines. The two item selection conditions × five test lengths × 2,000 examinees resulted in 20,000 simulated CATs.

Results

Because the simulations modeled a testing situation with two distinct purposes (i.e., obtaining accurate scores for the content balanced total test and improving the reliability of the eight discipline scores), different criteria were evaluated, taking into account the purpose of the outcome measures.

Total Test Results

A unique "true trait score" for the total test was never generated for the simulated examinees; instead a surrogate value was created that encompassed the multidimensionality induced in the data by the simulation model and removed as much estimation error as possible. The obvious choice was to use a total test MLE, based on the item difficulties from the bank and using all 2,458 item responses for each simulated examinee. The marginal reliability of the obtained total test MLEs was .994. Using these item bank-based MLEs as the approximate total test "true" scores (T), reliability coefficients were computed as the squared correlation between the CAT-produced MLEs, X, and T; that is, $R^2_{XT} = R_{XX'}$ (which follows from classical true score theory). Table 3 provides the reliability coefficients for each of the test lengths (100, 150, 200, 250, and 300 items).

The results were somewhat predictable. First, increasing the test length improved reliability. For example, the coefficient for UTOTCAT increased from .921 at 100 items to .978 at 300 items. A similar trend was evident for MIRTCAT. Second, because UTOTCAT specifically optimized information for the total test "reference composite" (Wang, 1986), it was expected that UTOTCAT would do a better job of selecting items than MIRTCAT for that same trait composite. That appears to be precisely what happened. For example, at 200 items the reliability coefficients for UTOTCAT and MIRTCAT were .963 and .958, respectively, where UTOTCAT showed a slight improvement in estimation accuracy. The same type of improvement in

Table 3
Total Test Score Reliability Coefficients (R^2_{XT})

Test Length	UTOTCAT	MIRTCAT
100	.921	.920
150	.947	.945
200	.963	.958
250	.972	.968
300	.978	.972

accuracy was shown at the other test lengths.

Similar results (not reported here) were obtained with respect to the accuracy of pass/fail decisions. The cutscore was set to correspond approximately to where the actual Step 1 standard would fall for this assumed population. The simulated examinees were classified as passing or failing the test in terms of their total bank MLEs and for the total test scores obtained from UTOTCAT and MIRTCAT. As might be expected, the false positive and false negative rates (the proportion of true failing examinees who passed and the proportion of true passing examinees who failed, respectively) dropped off for the increased test lengths. UTOTCAT produced nominally better results.

Thus, although UTOTCAT achieved better score reliability and decision accuracy than MIRTCAT, the magnitude of improvement was, from a practical perspective, quite small. The small differences seem insufficient to warrant rejecting the multidimensional approach. It may also be worth reiterating the fact that all CAT forms were forced to meet strict content requirements (i.e., the specifications derived from Table 1). Those requisite constraints may have actually caused the total test outcomes for UTOTCAT and MIRTCAT to be closer than if no such constraints had been imposed.

Discipline Subscore Results

The estimated discipline scores under UTOTCAT and MIRTCAT provided a somewhat different picture than at the total test level. First, consider the recovery of the true subscore trait parameters used to generate the data. Table 4 provides the squared product-moment correlations between the MLEs for UTOTCAT and MIRTCAT for each of the eight disciplines (i.e., R^2_{XT}). As expected, moving from a 100- to a 300-item test increased the score reliability. However, the reliabilities for the MIRTCATs were consistently higher than those for UTOTCAT. For example, in physiology, pathology, pharmacology, gross anatomy, and histology, a 150-item MIRTCAT had an effective test length of a 200-item UTOTCAT. In biochemistry, microbiology, and behavioral sciences, the 150-item MIRTCAT approached the reliability of a 250-item UTOTCAT.

Table 5 shows the mean bias as an average simple deviation between the CAT subscore estimates and the "true" discipline subscores. There were no unexpected outcomes. Generally, the mean bias decreased with increasing test length, and the more reliable subscores (see Table 4) tended to have less bias. For example, the mean bias for the physiology subscores was .054 at 100 items and .013 at 300 items under UTOTCAT. Similarly, for MIRTCAT, the bias for the physiology subscores was .035 at 100 items, but dropped to .018 at 300 items. No obvious patterns of bias were evident between UTOTCAT and MIRTCAT, when considered for the same disciplines.

One final finding involved the nature of the item selections made using UTOTCAT and MIRTCAT. Table 6 provides the range statistics (maximum – minimum, across examinees) of the number of items adaptively selected in each discipline category for the 100-item and 300-item tests. In UTOTCAT, the ranges were reasonably small for the 100-item test in physiology, pathology, behavioral sciences, gross anatomy, and histology (i.e., 6 to 9 items). However, areas like biochemistry (with a range of 20 items) exhibited marked fluctuations in the maximum versus minimum number of items selected for individual examinees. At 300 items, the ranges increased even further for UTOTCAT. However, under MIRTCAT, the ranges were small and fairly homogeneous. At 100 items, examinees differed only by 1 to 9 items administered in each of the disciplines. At 300 items, the largest range was only 14 items for pathology; the ranges for the remainder of the disciplines were from 3 to 7 items. From a validity perspective, MIRTCAT therefore appears to have produced more consistent allocations of items within the disciplines across examinees.

Discussion

The simulations presented here suggest that there may be some compelling advantages to using a multidimensional approach, subject to many practical considerations. This particular application of MIRT CAT

Table 4
R^2_{XT} Reliability Coefficients Resulting from UTOTCAT and MIRTCAT for the Eight Discipline Subscores at Test Lengths of 100, 150, 200, 250, 300

	Test Length				
Test	100	150	200	250	300
UTOTCAT					
Physiology	.612	.721	.796	.828	.848
Biochemistry	.490	.599	.659	.723	.764
Pathology	.762	.823	.870	.897	.916
Microbiology	.560	.667	.743	.797	.832
Pharmacology	.682	.773	.823	.854	.874
Behavioral Sciences	.513	.593	.658	.734	.773
Gross Anatomy	.446	.539	.601	.635	.686
Histology	.410	.487	.584	.645	.687
MIRTCAT					
Physiology	.719	.801	.830	.848	.865
Biochemistry	.656	.719	.755	.794	.821
Pathology	.832	.872	.899	.918	.925
Microbiology	.774	.808	.830	.859	.874
Pharmacology	.787	.824	.852	.872	.887
Behavioral Sciences	.618	.707	.745	.778	.803
Gross Anatomy	.476	.585	.664	.682	.714
Histology	.484	.573	.659	.728	.752

to a medical licensure context attempted to illustrate three of those considerations.

First, there needs to be a reason to use a multidimensional model, given the estimation and item selection complexities involved in MIRT CAT. It was suggested that the justification for implementing MIRT CAT

Table 5
Mean Bias Functions (Observed – True) From UTOTCAT and MIRTCAT for the Eight Discipline Subscores at Test Lengths of 100, 150, 200, 250, 300

	Test Length				
Test	100	150	200	250	300
UTOTCAT					
Physiology	.054	.022	.008	.013	.013
Biochemistry	.020	−.014	−.021	−.023	−.018
Pathology	−.011	−.008	−.002	−.002	−.003
Microbiology	.020	−.015	−.019	−.028	−.016
Pharmacology	−.005	−.017	.002	.002	.004
Behavioral Sciences	.014	.010	−.002	−.004	−.006
Gross Anatomy	.075	.041	.022	.013	.008
Histology	.045	.024	.027	.022	.025
MIRTCAT					
Physiology	.035	.032	.027	.020	.018
Biochemistry	.066	.031	.025	.012	.006
Pathology	.024	.008	.006	.009	.008
Microbiology	−.029	−.011	−.021	−.022	−.019
Pharmacology	−.011	−.003	.002	.003	.006
Behavioral Sciences	.045	−.007	−.028	−.024	−.021
Gross Anatomy	.064	.044	.040	.033	.025
Histology	.060	.049	.036	.018	.020

Table 6
Ranges of Adaptive Item Selections Within Discipline Categories From UTOTCAT and MIRTCAT at Test Lengths of 100 and 300 Items

	UTOTCAT		MIRTCAT	
Discipline	100	300	100	300
Physiology	7	29	9	6
Biochemistry	20	22	2	5
Pathology	9	28	5	14
Microbiology	15	28	1	5
Pharmacology	11	20	3	4
Behavioral Sciences	8	25	4	4
Gross Anatomy	7	20	5	7
Histology	6	15	2	3

might come from considerations of the test purpose and the scores that are reported. If a CAT is to be used to report a single score or pass/fail decision, there may be little practical advantage (and perhaps many disadvantages) to using MIRT. However, when the test serves a dual purpose (e.g., reporting total test outcomes and subscores, or performance profiles in specific categories), there may be sufficient justification for using a multidimensional model.

Second, any CAT can be viewed as a heuristic process that controls the sequence of item selections by local optimizations. It was demonstrated that the differences between unidimensional CAT and MIRT CAT heuristics are a rather simple reformulation of the coefficients in a particular objective function. There is no absolute requirement, however, of using the determinant of the inverse information matrix in the MIRT CAT objective function. In fact, when considering the computational loads associated with calculating the inverse and determinant for 1,000–3,000 unselected items in the bank at each step in the MIRT CAT—perhaps for 5–10 dimensions—computing time and numerical accuracy factors become important issues and it is tempting to look for alternatives. Any reasonable function that minimizes the asymptotic dispersion in some way (e.g., maximizing the trace of the information matrix) could be a likely candidate to replace the determinant in the objective function (e.g., Miller et al., in press).

Finally, the choice of a medical licensure application was intended to emphasize that unidimensional CAT and MIRT CAT need to consider content as far more than a set of simple frequency constraints that sum to the total test length and that might reduce reliability. In professional certification and licensure testing, content validity is usually non-negotiable; content experts, not psychometricians, often determine what the minimum content requirements will be. From this content-driven perspective, adaptive testing is merely another way of constructing tests; it is acceptable if and only if the test content specifications are met. In some cases, aspects of the multidimensional content may be used to optimize score reliability or other criteria. For situations in which that is possible, the simulations presented here suggest that MIRT CAT may have certain psychometric advantages related to reported scores.

There are numerous other issues to resolve and a great deal more research needed before any final conclusions can be made regarding MIRT CAT. Building on Segall's (1996) research, the present study illustrated a particular application and demonstrated some reasonable gains in the reliability of reported subscores under MIRT CAT, with no serious degradation of total test outcomes.

References

Aquinis, H. (1994). A QuickBasic program for generating correlated multivariate random normal scores. *Educational and Psychological Measurement, 54,* 687–690.

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick, *Statistical theories of mental test scores* (pp. 397–479). Reading MA: Addison-Wesley.

Bloxom, B., & Vale, C. D. (1987, June). *Multidimensional adaptive testing: A procedure for sequential estimation of the posterior centroid and dispersion of theta.* Paper presented at the annual meeting of the Psychometric Society, Montreal.

Bock, R. D. (1985). Contributions of empirical Bayes and marginal maximum likelihood methods to the measurement of individual differences. In E. E. Roskam (Ed.), *Measurement and personality assessment* (pp. 75–99). North Holland, The Netherlands: Elsevier.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 29–51.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement, 12,* 261–280.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431–444.

Doody-Bogan, E., & Yen, W. M. (1983, April). *Detecting multidimensionality and examining its effects on vertical equating with the three-parameter logistic model.* Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Federation of State Medical Boards & National Board of Medical Examiners. (1996a). *Step 1: General instructions, content description, and sample items.* Philadelphia PA: National Board of Medical Examiners.

Federation of State Medical Boards & National Board of Medical Examiners. (1996b). *Step 2: General instructions, content description, and sample items.* Philadelphia PA: National Board of Medical Examiners.

Fraser, C. (1986). *NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, New South Wales, Australia: The University of New England.

Glas, C. A. W. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (pp. 236–258). Norwood NJ: Ablex.

Green, B. F. (1988, October). *Notes on the item information function in the multidimensional compensatory IRT model* (Research Report 88-10). Baltimore MD: Psychometric Laboratory, The Johns Hopkins University.

Greenburg, A. G., Case, S. M., Golden, G. S., & Melnick, D. E. (1996, June). *Core content of Step 2 of the USMLE using surgery as an example.* Paper presented at the 7th Annual International Ottawa Conference, Maastricht, The Netherlands. (Proceedings in press).

Kendall, M. G., & Stuart, A. (1967). *The advanced theory of statistics* (Vol. 2). New York: Hafner.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257–283). New York: Academic Press.

Lehmann, E. L. (1983). *Theories of point estimation.* New York: Wiley.

Linacre, M., & Wright, B. D. (1994). *BigSteps* [Computer program]. Chicago: MESA Press.

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

Luecht, R. M. (in press). A generalized heuristic for test construction using item response theory. *Applied Psychological Measurement.*

Luecht, R. M., & Hirsch, T. M. (1992). Item selection using an average growth approximation of target information functions. *Applied Psychological Measurement, 16,* 41–51.

Luecht, R. M., & Miller, T. R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement, 16,* 279–294.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6,* 379–396.

Miller, T., Reckase, R., Spray, J., Luecht, R., & Davey, T. (in press). *Multidimensional item response theory.* Iowa City IA: ACT Publications.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359–381.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177–195.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351–356.

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

Reckase, M. R., & McKinley, R. L. (1983, April). *The definition of difficulty and discrimination for multidimensional item response theory models.* Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61,* 331–354.

Stocking, M. L. (1993). *Controlling item exposure rates in a realistic adaptive testing paradigm* (ETS Research Report 93-2). Princeton NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1995). *A new method of controlling item exposure in computerized adaptive testing* (ETS Research Report 95–25). Princeton NJ:

Educational Testing Service.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277–292.

Swanson, D. B., Case, S. M., Kelley, P. R., Lawley, J. L., Nungester, R. J., Powell, R. D., & Volle, R. L. (1991). Phase-in of the NBME Comprehensive Part I Examination. *Ideas for Medical Education, 66,* 443–444.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving large item selection problems. *Applied Psychological Measurement, 17,* 151–166.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item exposure rates in computerized adaptive testing. *Proceedings of the 27th Annual Meeting of the Military Testing Association* (pp. 973–977). San Diego CA: Navy Personnel Research and Development Center.

Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 215–237). New York: Springer-Verlag.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer.* Hillsdale NJ: Erlbaum.

Wang, M. M. (1986, April). *Fitting a unidimensional model to multidimensional item response data.* Paper presented at the Office of Naval Research Contractors' Meeting, Knoxville TN.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21,* 361–375.

Wilson, D., Wood, R. L., & Gibbons, R. (1984). *TESTFACT: Test scoring and item factor analysis* [Computer program]. Chicago: Scientific Software, Inc.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: MESA Press.

Acknowledgments

Thanks to Terry Ackerman and two anonymous reviewers for their extremely helpful comments on an earlier draft of this paper.

Author's Address

Send requests for reprints or further information to Richard M. Luecht, National Board of Medical Examiners, 3750 Market Street, Philadelphia PA 19104, U.S.A. Email: rluecht@mail.nbme.org.