Improving Measurement Precision of Test Batteries Using Multidimensional Item Response Models

Wen-Chung Wang
National Chung Cheng University

Po-Hsi Chen
National Yunlin University of Science and Technology

Ying-Yao Cheng
National Sun Yat-sen University

A conventional way to analyze item responses in multiple tests is to apply unidimensional item response models separately, one test at a time. This unidimensional approach, which ignores the correlations between latent traits, yields imprecise measures when tests are short. To resolve this problem, one can use multidimensional item response models that use correlations between latent traits to improve measurement precision of individual latent traits. The improvements are demonstrated using 2 empirical examples. It appears that the multidimensional approach improves measurement precision substantially, especially when tests are short and the number of tests is large. To achieve the same measurement precision, the multidimensional approach needs less than half of the comparable items required for the unidimensional approach.

Item response theory (IRT) has been widely used to analyze test data. A basic assumption in most item response models is unidimensionality, that is, the items in a test are assumed to measure the same latent trait (e.g., proficiency, attitude, or personality). If several unidimensional tests (e.g., a test battery) are to be calibrated, a typical way to do so is to apply unidimensional item response models separately, one test at a time. This unidimensional item response approach (hereafter referred to as the unidimensional approach for brevity), which ignores the correlations between latent traits, yields imprecise measures when tests are short. To take the correlations into account, one needs multidimensional item response models that simultaneously calibrate all the tests and thus use the correlations to increase measurement precision. This is referred to as the multidimensional approach. Figures 1a and 1b depict the unidimensional and multidimensional approaches, respectively. In reality, there are always nonzero correlations between latent traits, meaning that at least in theory Figure 1b is more appropriate than Figure 1a. If the correlations are ignored and the approach in Figure 1a is used, the estimation will be less efficient than when the approach in Figure 1b is used in which the correlations are taken into account. Only if the correlations are zero will the multidimensional approach be equivalent to the unidimensional approach. The greater the correlations, the greater the measurement efficiency using the multidimensional approach. This situation is analogous to seemingly unrelated regression equations models in econometrics (Srivastava & Giles, 1987) in which dependent variables are correlated. The estimation of model parameters is more efficient when the nonzero correlations between dependent variables are taken into account than when they are ignored (Srivastava & Giles, 1987, p. 30).

If tests are long enough, the measures of each individual latent trait obtained from the unidimensional
approach will be accurate, and thus the multidimensional approach will yield little improvement in measurement precision. On the other hand, if tests are too short for the unidimensional approach to yield precise measures, using the multidimensional approach will “squeeze” as much information as possible from the data to provide measures that are more precise. In addition to the advantage of measurement efficiency for the multidimensional approach, direct estimation of the correlation between latent traits is possible only for the multidimensional approach, not the unidimensional one.

To make the discussion of measurement efficiency of the multidimensional approach more concrete, consider that a student’s achievement in arithmetic is to be assessed. The most relevant tool for this is an arithmetic test. However, assume the test at hand is very short. Now suppose there are two pieces of collateral information about the student: IQ and language achievement. These two pieces of information can be used to increase the measurement precision of the student’s achievement in arithmetic if arithmetic achievement is correlated with IQ and language achievement. Also, the information about achievements in arithmetic and language can be used to increase the measurement precision of IQ, and the information about IQ and arithmetic achievement can be used to increase the measurement precision of language achievement, when the three corresponding tests are analyzed jointly. That is, any two can serve as collateral information for the other. Statistically speaking, collateral information refers to any variable that is correlated with the target variable. It can be information about test items (e.g., format, content, or cognitive process) or about examinees (e.g., educational background, demographic status, or performance in other tests).

It can be shown that, under the normality assumption on the likelihood and the prior distribution, the posterior precision (precision is defined as the inverse of the variance of an estimator) is equal to the datum precision plus the prior precision (Lee, 1989, p. 38). The more informative the collateral information about the prior distribution, the greater will be the improvement in the posterior precision.

Mislevy (1987); Mislevy and Sheehan (1989b); and Adams, Wilson, and Wu (1997) have shown that using collateral information about persons’ background variables, such as educational background or status on demographic variables, increases estimation precision of item parameters and person abilities. The present study, adopting the same idea of using collateral information to increase measurement precision, differs from theirs in that (a) item responses to other tests rather than values of a person’s background variables are treated as collateral information and (b) multidimensional item response models rather than unidimensional item response models are used.

Segall (1996) derived a Bayesian procedure for person estimation and item selection in the context of multidimensional adaptive testing (MAT) and found that for nine latent traits with moderate to high cor-
relations, the MAT method achieved greater or comparable precision about persons with one third fewer items than were required for the conventional unidimensional adaptive testing (UAT) method. Luecht (1996) examined the benefits of applying MAT over UAT in a licensing context in which mandatory complex content constraints were imposed. MAT with content constraints could achieve about the same measurement precision as UAT with 25% to 40% fewer items. The improvement in measurement precision in the MAT procedure occurred mainly because the correlations between the latent traits were taken into account when selecting items and estimating person abilities. The present study is similar to Segall’s and Luecht’s studies in that the correlations between latent traits are used, but it differs from theirs in that (a) both person parameters and item parameters are unknown and have to be jointly calibrated from data in the present study, whereas item parameters were known in theirs, and (b) the present study includes both dichotomous and polytomous items, whereas their studies were limited to dichotomous items.

The main purpose of the present study is to show how measurement precision can be improved through use of the correlations between latent traits. In the following, a multidimensional item response model is described that jointly calibrates all tests and thus uses the correlations between latent traits to increase measurement precision about persons. Two real data sets were analyzed, one a science test battery consisting of 50 multiple-choice items in five areas of science, and the other a personality inventory consisting of forty 4-point Likert-type items in 10 scales. The unidimensional and multidimensional approaches were compared in terms of measurement efficiency and accuracy in the estimation of the variance–covariance matrices of latent traits.

The Multidimensional Item Response Model

In the present study, test batteries are regarded as a collection of several unidimensional tests in which items in each unidimensional test are designed to measure the same latent trait, as shown in Figure 1b. This “simple” structure in which each item measures a single latent trait can be extended to the “complex” structure in which some items are designed to measure more than one latent trait simultaneously, as shown in Figure 1c. The complex structure is discussed later.

Consider the case in which item responses to a test battery are to be analyzed. When only unidimensional item response models are available, one has to either ignore individual latent traits and treat the test battery as a unidimensional test or separately analyze the tests in the battery, one test at a time. The former composite approach is not recommended because typically every test in the battery is designed to measure a distinct latent trait and the battery is designed to measure multiple latent traits. The latter unidimensional approach is often preferred because it preserves individual latent traits and it is consistent with the purpose and construction of the test battery. Unfortunately, the unidimensional approach will not yield accurate measures if tests are short. In such cases, one needs an item response model that not only preserves the individual latent traits in a test battery but also provides more accurate measures of each latent trait.

The multidimensional random coefficients multinomial logit (MRCML) model (Adams, Wilson, & Wang, 1997) is such a model. It assumes a set of D latent traits that determine test performances. Person $n$’s levels on the $D$ latent traits are denoted as $\theta_n = (\theta_{n1}, \ldots, \theta_{nD})$, which are considered to represent a random sample from a population with a multivariate density function $g(\theta_n; \alpha)$ where $\alpha$ indicates a vector of parameters that characterize the distribution. If $g$ is constrained to be normal, then $\alpha \equiv (\mu, \Sigma)$. Under the MRCML model, the probability of a response in category $k$ of item $i$ for person $n$ is expressed as

$$f(X_{nik} = 1; \xi | \theta_n) = \frac{\exp(b_{ik}^\prime \theta_n + a_{ik}^\prime \xi)}{\sum_{k=1}^{K_i} \exp(b_{ik}^\prime \theta_n + a_{ik}^\prime \xi)}$$

where $X_{nik} = 1$ if response to item $i$ for person $n$ is in category $k$ and 0 otherwise; $K_i$ is the number of categories in item $i$; $\xi$ is a vector of difficulty parameters that describe the items; $b_{ik}$ is a score vector given to category $k$ of item $i$ across the $D$ latent traits, which can be collected across items into a scoring matrix $B$; and $a_{ik}$ is a design vector given to category $k$ of item $i$ that describes the linear relationship among the elements of $\xi$, which can be collected across items into a design matrix $A$. Equation 1 can be expressed as

$$\ln \left[ \frac{f(X_{nik} = 1; \xi | \theta_n)}{f(X_{n,k-1} = 1; \xi | \theta_n)} \right] = (b_{ik} - b_{ik-1})^\prime \theta_n + (a_{ik} - a_{ik-1})^\prime \xi,$$

which is more consistent with the usual expressions of Rasch-type models. Notice that $b_{ik}$ and $a_{ik}$ are not parameters; rather, they are specified by test analysts.
to form customized item response models. Only $\Psi$ and $\alpha$ ($\theta$ and $\Sigma$ if normal) are estimated.

Using $a_n$ and $b_n$ to define the relationship between items and persons allows a general model to be written that includes most of the existing unidimensional Rasch models, such as the simple logistic model (Rasch, 1960), the linear logistic test model (Fischer, 1973), the rating scale model (Andrich, 1978), the partial credit model (Masters, 1982), the partial order model (M. R. Wilson, 1992), the facets model (Linacre, 1989), and the linear partial credit model (Fischer & Ponony, 1994). In addition, the definitions allow the specification of a range of multidimensional models by imposing linear constraints on the item parameters, such as multidimensional forms of the simple logistic model (to be used in Example 1), the rating scale model (to be used in Example 2), the partial credit model, and the linear partial credit model. For details on how to manipulate the scoring matrix and the design matrix to form the above models and other customized models, see Adams, Wilson, and Wang (1997) and the manual of the computer program ACER ConQuest (Wu, Adams, & Wilson, 1998).

ACER ConQuest is implemented with marginal maximum-likelihood (MML) estimation with Bock and Aitkin’s (1981) formulation of the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). Based on the assumption of conditional independence among items and persons, the probability of a response vector $x$ conditioned on the random quantities $\theta_n$ is

$$f(x = x; \xi(\theta_n)) = \frac{\exp[x'(B\theta_n + A\xi)]}{\Psi(\theta_n, \xi)},$$

where $\Psi(\theta_n, \xi) = \sum_{z \in \Omega} \exp[z'(B\theta_n + A\xi)].$

The likelihood equations for the item parameters are

$$\frac{\partial \ln \Lambda(\xi, \alpha|x)}{\partial \xi} = \prod_{n=1}^{N} \int_{0_n} \frac{\partial \ln f(x_n'; \xi; \alpha)}{\partial \xi} dH(\theta_n'; \xi; \alpha|x_n) = 0,$$

where $H(\theta_n'; \xi; \alpha|x_n)$ is the cumulative posterior marginal distribution of $\theta_n$ given $x_n$ with a density function

$$h(\theta_n'; \xi; \alpha|x_n) = \frac{f(x_n'; \xi|\alpha)}{f(x_n'; \xi)}.$$

Assuming the distribution of the latent traits is multivariate normal so that $\alpha \equiv (\mu, \Sigma)$, the likelihood equations for the mean and variance–covariance matrix are

$$\frac{\partial \ln \Lambda(\xi, \mu, \Sigma|x)}{\partial \mu} = \prod_{n=1}^{N} \int_{0_n} \frac{\partial \ln g(\theta_n; \mu, \Sigma)}{\partial \mu} dH(\theta_n; \xi, \mu, \Sigma|x_n) = 0,$$

and

$$\frac{\partial \ln \Lambda(\xi, \mu, \Sigma|x)}{\partial \Sigma} = \prod_{n=1}^{N} \int_{0_n} \frac{\partial \ln g(\theta_n; \mu, \Sigma)}{\partial \Sigma} dH(\theta_n; \xi, \mu, \Sigma|x_n) = 0.$$
(e.g., a test battery), and within-item multidimensional tests in which some items are designed to simultaneously measure more than a single latent trait. Figures 1b and 1c depict these two kinds of modeling: between-items multidimensionality and within-item multidimensionality, respectively. Note that the dimensionalities in Figures 1, a–c, are specified rather than discovered based on the data. Therefore, this approach is considered confirmatory rather than exploratory multidimensional item response modeling. A necessary and sufficient condition for identification of the MRCML model is provided in the ACER ConQuest manual (Wu et al., 1998, p. 132). The present study focuses on modeling between-items multidimensionality. The MRCML model does not require items to be administered together nor that they belong to a test battery. Different tests may be given at different time points, and the tests could measure quite diverse but correlated latent traits. The MRCML model, being an item response model, does require that item responses to different tests should be gathered for joint calibration.

The improvement in measurement precision with the multidimensional approach occurs because all the model parameters $\xi$, $\mu$, and $\Sigma$ are jointly estimated. Only if the variance–covariance matrix $\Sigma$ is a diagonal matrix, which means that the latent traits are independent, will measurement efficiency of the multidimensional approach be equivalent to that of the unidimensional approach. When the model is correctly specified, MML estimation will yield unbiased estimates for the model parameters $\xi$, $\mu$, and $\Sigma$ (Adams, Wilson, & Wang, 1997; Wang, 1999). In addition, the use of collateral information leads to smaller errors in the estimation of item parameters and reduces the mean-square errors of person estimates (Adams, Wilson, & Wu, 1997; Mislevy, 1987; Mislevy & Sheehan, 1989a, 1989b). Because the present study adopts the same idea of using collateral information to increase measurement precision, the same conclusions will hold. The implication of these findings is that using collateral information does not invalidate measurement; rather, it increases measurement precision by reducing measurement errors.

After the model parameters are calibrated, point estimates for individual persons can be obtained from either the mean vector of the marginal posterior distribution, Equation 8—the expected a posteriori (EAP) estimates—or the maximum point of the marginal posterior distribution—the maximum a posteriori (MAP) estimates. The EAP estimator is a shrinkage estimator, shrunken toward the mean of the prior distribution. Therefore, the sample variance of the EAP estimates does not accurately approach the population variance, nor does the sample covariance of the EAP estimates accurately approach the population covariance. It is thus inappropriate to use the sample variance and covariance of the EAP estimates to depict the population variance and covariance, respectively. Although the EAP estimator is a shrinkage estimator, the rank ordering of the EAP estimates across persons is not affected by the shrinkage (Wainer & Thissen, 1987). Given that an appropriate prior distribution has been specified, finite and plausible EAP estimates can be obtained even when the test taker responds in either the lowest or highest category for every item (which may happen for short tests), whereas maximum-likelihood (ML) estimates cannot be obtained. The EAP estimator has minimum mean-square error over the population of abilities and, in terms of average accuracy, cannot be improved upon. This property applies only when the prior is correct and when there is an infinite number of items (Bock & Mislevy, 1982).

To reduce the possibility of incorrectly specifying the mean vector and the variance–covariance matrix of the prior multivariate normal distribution, one may empirically estimate these parameters from item response data. This is referred to as the empirical Bayes method (Lee, 1989). This is exactly what ACER ConQuest does, because only the form of the prior distribution is specified (e.g., multivariate normal); the corresponding parameters (e.g., $\mu$ and $\Sigma$) are empirically estimated from data. Furthermore, item parameters $\xi$ do not appear to be strongly influenced by moderate departures from normality, so inappropriately specified person distributions may not be very important (Embreton & Reise, 2000, p. 214). In general, the EAP estimate is favored over the ML and MAP estimates and is the method of choice (Mislevy & Stocking, 1989). MML with the EM algorithm has been implemented in many popular IRT computer programs, such as BILOG (Mislevy & Bock, 1990), BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), MULTILOG (Thissen, 1991), PARSCALE (Muraki & Bock, 1993), TESTFACT (D. T. Wilson, Wood, & Gibbons, 1991), and XCALIBRE (Assessment Systems Corporation, 1996), as well as ACER ConQuest.

Measures are always imperfect and contain measurement error. When measurement error is considerable (even with test reliability augmented via incor-
porating collateral information), one should not overly rely on “point” estimates of person abilities (e.g., the EAP estimates). Rather, to deal with the imprecision in the EAP estimate, one may either report the EAP estimate together with a band of measurement error (e.g., the standard deviation of the marginal posterior distribution) or obtain several plausible values, which are random draws from the marginal posterior distribution. The National Assessment of Educational Progress (NAEP) program (Beaton & Zwick, 1992) reports five such plausible values for each person. (For details on the uses of plausible values, see Mislevy, 1991; Mislevy, Beaton, Kaplan, & Sheehan, 1992.) ACER ConQuest provides three kinds of person estimates for each person: the EAP estimate with a standard error, the ML estimate with a standard error, and five plausible values.

Ideally, tests should be long enough to provide accurate person estimates so that important decisions (e.g., admission to college) can be reliably made. If a test has already been administered and its test reliability found too low to yield reliable individual estimates, one may have to postpone decisions and collect new test data. Practical constraints might make it very difficult, if not impossible, to recall the test takers and to administer new test items. In such a case, incorporating collateral information might be helpful.

Although using collateral information improves measurement precision, it might damage “face validity,” because it is relatively difficult to explain to the public and test takers why a person’s score on a test depends partly on his or her background or performance on other tests. Statistically speaking, all possible pieces of collateral information could be taken into account to increase measurement precision. In practice, testers usually have to decide how divergent the tests contributing collateral information can reasonably be. For instance, should a mechanical aptitude test be used to augment information in a college admission test? Even if such a test were correlated moderately with the admission test, some people may object to its use because they think the mechanical content is too different to be incorporated with a test of verbal or mathematical skills. Mislevy (1987) offered a good solution: Collateral information should be very cautiously used in estimates of person abilities when tests are used in individual selection or important placement decisions, for example, determining who should be admitted to college or who should receive awards. In these situations the tester ought to gather enough data directly relevant to proficiency to make sufficiently precise decisions on that basis alone. That is, for ease of interpretation, it is recommended that tests providing collateral information not be so divergent that one might question the reasonableness of adding information from them to a test taker’s score. Consensus about the coverage of tests should be reached when very divergent tests are to be used as collateral information for important decisions.

Measurement Efficiency

In the present study, two indices are used to compare measurement efficiency of the unidimensional and multidimensional approaches: test reliability and numbers of items needed to achieve the same measurement precision. Within classical test theory (CTT), test reliability is defined as

\[ \rho_{CTT} = \frac{\sigma_T^2}{\sigma_X^2} \]  

where \(\rho_{CTT}\) is the test reliability within CTT, and \(\sigma_T^2\) and \(\sigma_X^2\) are the variances of true scores \(T\) and observed scores \(X\), respectively. Within IRT, measurement errors are no longer homogeneous; they depend on the levels of the latent trait \(\theta\). More specifically, each item provides a different degree of measurement precision at each \(\theta\) level, which is referred to as the item information, denoted as \(I(\theta)\). Summation of the item information over items in a test leads to the test information \(T(\theta)\),

\[ T(\theta) = \sum I(\theta), \]  

which also depends on the \(\theta\) level and is inversely related to the error variance at the \(\theta\) level (Lord, 1980):

\[ \sigma_{\theta}^2 = \frac{1}{T(\theta)}. \]  

The test reliability within IRT is no longer a constant but rather is a function of \(\theta\). For ease of comparison, a single quantity can be formed as the counterpart of the classical test reliability, shown in Equation 11. First, the test information is averaged over the \(\theta\) level to obtain \(\bar{T}\). This average test information \(\bar{T}\) describes an average degree of measurement precision the test provides for the sampled persons. Then, the IRT-based test reliability (also called the composite test reliability) is defined as

\[ \rho_{IRT} = \frac{\sigma_{\bar{T}}^{-1}}{\sigma_\theta^2}, \]  

where \(\sigma_\theta^2\) is the variance of the \(\theta\) distribution.
Sometimes, computing the IRT-based test reliability in this way can be tedious and time-consuming because each item information function has to be computed to form the test information and the test information has to be integrated over the \( \theta \) level to form the average test information. Mislevy et al. (1992) suggested a simpler solution when MML estimation is used,

\[
\rho_{\text{MML}} = \frac{\sigma_E^2}{\sigma_0^2},
\]

where \( \sigma_E^2 \) is the variance of the EAP estimates. Because IRT computer programs that use MML estimation usually report both the estimates of the variance \( \sigma_0^2 \) and the EAP estimate for every person, Equation 15 may be more practical than Equation 14 in real data analysis.

Combining Equations 14 and 15 leads to

\[
\bar{T} = \frac{1}{\sigma_0^2 - \sigma_E^2}.
\]

The relative efficiency \( RE \) of the multidimensional approach over the unidimensional approach is defined as the ratio of the test information from the multidimensional approach over that from the unidimensional approach,

\[
RE_{\text{MU}}(\theta) = \frac{T_M(\theta)}{T_U(\theta)}
\]

where the subscripts M and U denote the multidimensional and unidimensional approaches, respectively. The relative efficiency depends on the \( \theta \) level. If a single quantity is preferred, one may compute the average relative efficiency \( ARE \) as follows:

\[
ARE_{\text{MU}} = \frac{\bar{T}_M}{\bar{T}_U}.
\]

The average relative efficiency describes how many times more precise the multidimensional approach is on average relative to the unidimensional approach. Equivalently, this describes roughly how many times more comparable items are needed for the unidimensional approach to achieve the same measurement precision as required for the multidimensional approach. For example, if \( ARE \) is 2, the unidimensional approach needs twice as many comparable items as are required for the multidimensional approach to achieve the same measurement precision.

Direct Estimation of the Variance–Covariance Matrix of Latent Traits

In addition to having an interest in estimates for individual persons, test users may be interested in the correlations between latent traits. In the present study, the persons are assumed to be a random sample of a multivariate normal distribution with a mean vector \( \mu \) and a variance–covariance matrix \( \Sigma \). Using the multidimensional approach, the variance–covariance matrix is estimated directly. The estimates of the variance–covariance matrix (and thus the correlation matrix) yielded by ACER ConQuest are unbiased (Adams, Wilson, & Wang, 1997; Wang, 1999). When the unidimensional approach is applied, as shown in Figure 1a, direct estimation of the variance–covariance matrix or the correlations between latent traits is not possible.

To obtain the correlations between latent traits when the unidimensional approach is applied, a two-step procedure was proposed (Wang, 1999). In the first step, the Pearson product–moment correlation between the ML estimates for two latent traits is computed. For small tests, the measurement errors are often so large that this correlation substantially underestimates the correlation between latent traits. To adjust the attenuated correlation, Spearman (1904) provided the following disattenuation formula within CTT:

\[
\rho_{\text{XY}}' = \frac{\rho_{\text{XY}}}{\sqrt{\rho_{\text{XX}} \rho_{\text{YY}}}}
\]

where \( \rho_{\text{XY}}' \) is the (disattenuated) correlation between latent traits of \( X \) and \( Y \) (also called the true scores of \( X \) and \( Y \)), \( \rho_{\text{XY}} \) is the attenuated correlation, and \( \rho_{\text{XX}} \) and \( \rho_{\text{YY}} \) are the test reliabilities of \( X \) and \( Y \), respectively. In the second step, Equation 19 is applied to compute the disattenuated correlations. Applying this disattenuation formula does not mean that the test reliabilities are going to be raised to unity in practice. Instead, it reflects that the correlation between latent traits, rather than the correlation between measures, is the focus. In this regard, the disattenuation formula should be used when measurement error is not negligible.

Sometimes using this disattenuation formula can lead to a correlation exceeding unity. For instance, if the attenuated correlation is .8 and the reliabilities for both tests are .7, applying the disattenuation formula yields a disattenuated correlation of 1.14 (\( = .8/\sqrt{.7 \times .7} \)). Even so, using the disattenuation formula...
will generally yield a better approximation of the correlations between latent traits than not using it, especially when tests are short. However, the two-step disattenuation procedure does not yield unbiased estimates of the correlation between latent traits (Wang, 1999); after all, the disattenuation formula was developed within CTT rather than IRT.

Example 1: The Basic Science Proficiency Test

The experimental items of the Taiwan Basic Science Competency Test of Year 2000 for junior high school students in Taiwan (Chen & Wang, 2000) were analyzed. This test, based on a random sample from a set of item banks, contained 50 multiple-choice items in five science subjects, including nine items for Biology, nine for Health Education, eighteen for Physics and Chemistry, seven for Earth Science, and seven for Environmental Science. Five item banks, each for a science subject, were built through careful and thorough survey of domains of the school courses. All items were written by high school teachers of the relevant subject matter and then modified by other teachers of the same subject matter and psychometric experts. A total of 1,716 tenth graders took the test, including 1,108 male students and 608 female students.

Because the five subtests were short, the measures were expected to be imprecise. One way to deal with this problem is to ignore individual latent traits and treat all 50 items as measuring a single latent trait, say, global science proficiency, and report a composite score. The major justification for using this composite approach is greater measurement precision, because, other things being equal, the longer the test is, the more precise the measure. However, the composite approach has the following shortcomings. First, the measured subjects are taught and tested separately in school; thus, test takers would like to know how they performed in each subject-matter area. Individual score reports may also help test takers differentially adjust learning strategies and clarify potential aptitudes toward development of future studies and career plans. Second, teachers of a particular subject are likely to be more interested in knowing how their students perform in their own subjects than in science as a whole. Profile information about students’ score patterns could help teachers revise teaching strategies, redesign curricula when needed, and provide necessary remedial instruction for particular students. Third, parents might need the profile information to assess their children’s performance in the individual subjects, to formulate specific guidelines of learning and career development for their children, to adjust the home environment to stimulate efficient learning, and to consult teachers to identify learning problems and work together to build a better instructional environment. Finally, education policymakers would like to know how schools educate children on the individual subjects. As the subjects are often taught separately, education policies should be made on the basis of information about individual subjects.

An ideal way to resolve the problem of imprecise measures is to increase test length, prior to test administration. However, test analysts are usually requested to squeeze as much information as possible from the data to meet testers’ needs, after test administration. Practical considerations might make increasing test length impossible. In such cases, the multidimensional approach may be an alternative that can provide accurate measures while preserving individual latent traits. If measurement precision of each individual latent trait can be raised to a more satisfactory level via the multidimensional approach, then there is little benefit in using the composite approach.

Measurement Precision

This data set was analyzed with the simple logistic model (Rasch, 1960) using both the multidimensional and unidimensional approaches. Appendix A shows the scoring matrix \( B \) and design matrix \( \mathbf{A} \) and the corresponding ACER ConQuest commands for these two approaches. Although the formulations of these two matrices appear to be complicated, test analysts do not have to worry because they are automatically generated by the program.

Throughout the present study, the dimensionality and scoring functions (item weights) are specified rather than obtained by exploring the data, so that the multidimensional approach is confirmatory rather than exploratory. Items from which item scores were summed to form a raw score for test reports were treated as measuring the same latent trait. In this example, items in each subtest were designed to measure the general knowledge of the corresponding subject and thus presumably formed a unidimensional test. Standard item fit statistics in IRT analysis were computed to check whether the data fitted the model’s expectation. If data fit the model’s expectation, the weighted mean-square error (Wright & Masters, 1982; Wu et al., 1998) will be close to its expected value of unity. The mean-square errors for high-stakes
multiple-choice items for testing whether the data fitted the model’s expectation should be between 0.8 and 1.2 and the corresponding $z$ statistics within the alpha (.05 or .01) nominal levels (Linacre & Wright, 1994). For the composite approach in which all 50 items were assumed to measure the same latent trait, the weighted mean-square errors for the 50 items were between 0.59 and 1.57 ($M = 1.01$). The corresponding $z$ statistics were between −14.1 and 14.3, which were far beyond the critical range at the .01 nominal level. Therefore, there was a poor fit between the data and model so that these 50 items should not be treated as measuring the same latent trait.

For the unidimensional approach in which the items in each subtest were assumed to measure the same latent trait, the weighted mean-square errors for the 50 items were between 0.95 and 1.07 ($M = 1.002$). The corresponding $z$ statistics were between −2.3 and 2.4, indicating that the null hypotheses of model-data fit were not rejected at the .01 nominal level. As a result, the items in each subtest could be treated as measuring the same latent trait. When the multidimensional approach was applied, the weighted mean-square errors for the 50 items were almost identical to those found from the unidimensional approach. They were between 0.95 and 1.08 ($M = 0.998$), with the corresponding $z$ statistics between −2.2 and 2.3. Therefore, there was a good fit between the data and the model, which indicated that treating the items in each subtest as measuring the same latent trait was appropriate.

The upper part of Table 1 shows the IRT-based test reliabilities of the five subtests (Equation 15) obtained from both approaches. With the unidimensional approach, the test reliability of Physics and Chemistry was .80, which was rather high because the test was long, having 18 items. The other four test reliabilities were very low, between .43 and .55, as their tests were very short, having either 7 or 9 items. These four test reliabilities were too low to accurately depict individual differences on the latent traits. Obviously, the unidimensional approach, although preserving individual latent traits, yielded imprecise measures. When the multidimensional approach was used, the test reliability of Physics and Chemistry was .86, slightly higher than that obtained from the unidimensional approach. The improvement in measurement precision of Physics and Chemistry using the multidimensional approach was not huge because this subtest was long enough to yield relatively precise measures using the unidimensional approach. In contrast, the test reliabilities of the other four subtests were much higher using the multidimensional approach; they were between .76 and .83. The multidimensional approach substantially increased the test reliabilities of these four subtests by 51% to 77% as shown in Table 1.

The unidimensional approach requires 24, 21, 26, 18, and 15 comparable items on Biology, Health Education, Physics and Chemistry, Earth Science, and Environmental Science, respectively, to achieve the same measurement precision as the multidimensional approach for 9, 9, 18, 7, and 7 items, respectively. The multidimensional approach requires only 38% (i.e., 9/24), 42%, 70%, 40%, and 47% of the numbers of comparable items that are required for the unidimensional approach. The measurement precision that the unidimensional approach is expected to achieve with 104 (i.e., 24 + 21 + 26 + 18 + 15) comparable items can be achieved with only 50 items using the multidimensional approach. In other words, the multidimensional approach requires only 48% ($= 50/104$) of the number of items required for the unidimensional approach.

Incorporating collateral information is analogous to multiple regression analysis in that adding additional predictors generally increases prediction precision to some extent because the coefficient of determination $R^2$ is always nondecreasing. Generally, the larger the number of latent traits (subtests), the greater the collateral information that is available, and thus the greater the increment in measurement precision. In other words, as the number of latent traits increases, larger increments in measurement precision are expected using the multidimensional approach, compared with the unidimensional one.

To give the reader a more concrete impression, we next illustrate the effect of increasing the number of subtests on measurement efficiency by adding each additional subtest one at a time. First, suppose there are only two subtests, say, Subtests 1 (Biology) and 2 (Health Education), each with 9 items. The measurement precision of these 18 items using the multidimensional approach can be achieved with 21 ($= 10 + 11$) comparable items using the unidimensional approach. That is, the multidimensional approach needs 86% ($= 18/21$) of the number of items required for the unidimensional approach as shown in the lower part of Table 1. When Subtest 3 (Physics and Chemistry) is added to these two subtests, the measurement precision of these 36 items using the multidimensional approach can be achieved with 54 ($= 16 + 15 + 23$) comparable items using the unidimensional approach. Equivalently, the multidimen-
sional approach requires 67% of the number of items required for the unidimensional approach. When Subtest 4 (Earth Science) is added to the three subtests, the multidimensional approach requires only 54% of the number of comparable items required for the unidimensional approach. Finally, when Subtest 5 (Environmental Science) is added to the four subtests, the multidimensional approach requires only 48% of the number of items required for the unidimensional approach. Obviously, the percentage of items needed decreases as the number of subtests increases, which indicates that adding additional subtests generally leads to higher measurement efficiency for the multidimensional approach, compared with the unidimensional approach. Note that the ordering of the subtests is not a critical issue in this illustration. If the ordering were different, the main result—that the larger the number of subtests is, the greater the measurement efficiency of the multidimensional approach—would remain unchanged.

Correlation Matrices of the Latent Traits

Table 2 shows the correlation matrices obtained from the unidimensional and multidimensional approaches. When the multidimensional approach was used, the correlations were between .85 and .94, indicating that the five latent traits were highly correlated. When the unidimensional approach was used, the attenuated correlations were between .40 and .64, suggesting that the five latent traits were only moderately correlated. After disattenuation, these five latent traits appeared highly correlated, with correlations between .85 and 1.03 (which should be treated as unity). The disattenuated correlations were much closer to the correlations obtained from the multidimensional approach than the attenuated correlations.

Summary and Discussion

The multidimensional approach improves measurement precision dramatically. The magnitude of improvements depends on the correlations between latent traits, the number of latent traits (subtests), and the test lengths. The higher the correlations, the greater the number of latent traits, and the shorter the subtests, the more significant the improvement will be. The collateral information about item responses to other tests is most useful in settings in which few responses to each test are solicited from each person. When tests are long, the improvement shows a ceiling effect. In other words, there is less to be gained using collateral information when the target latent
trait is already well determined by its test score alone. Greater benefit accumulates from using collateral information as it relates more strongly to the target latent trait, and when less information is available from the observed target test score itself. This is evident from the fact that the test reliability of the longest subtest, Physics and Chemistry, increased from .80 with the unidimensional approach to only .86 with the multidimensional approach, whereas the test reliabilities of the four other shorter subtests increased substantially. The direct estimation of the variance–covariance matrix (and the correlation matrix) of latent traits is another advantage of the multidimensional approach.

Although the improvement in measurement precision via the multidimensional approach is impressive, it does not mean that the multidimensional approach should be applied to every test containing multiple domains. For instance, a science test about weather may contain several domains: air pressure, wind, temperature, humidity and precipitation, and clouds. Because in practice a single total score is reported and the domain scores lack substantial educational and psychological meanings, the test as a whole would be better treated as unidimensional and analyzed as such. The Multilevel Edition of the Cognitive Abilities Test (Lohman & Hagen, 2002) is another example. The test contains nine subtests, which provide Verbal (Verbal Classification, Sentence Completion, and Verbal Analogies), Quantitative (Quantitative Relations, Number Series, and Equation Building), and Nonverbal (Figure Classification, Figure Analogies, and Figure Analysis) scores. Because in practice only three scores (Verbal, Quantitative, and Nonverbal) are reported, the test can be viewed as containing three subtests, that is, the test is three-dimensional. Applying the multidimensional approach, three measures for the three corresponding latent traits are obtained. The test should not be treated as nine-dimensional because it contradicts the usual way of reporting scores and because the resulting construct interpretability and test utility will be very weak. The test should not be treated as unidimensional, either, because that certainly disagrees with the purpose and construction of the test. Test constructs and existing score reporting methods should be taken into consideration when specifying dimensionality. More important, the specification of dimensionality has to be checked empirically, such as by checking model–data fit as was done in this example.

One might question whether in this example the five subtests were so short and so highly correlated that the test as a whole should be treated as measuring a single latent trait labeled science. This composite approach was not appropriate for this particular example for several reasons.

1. The subtests were designed to measure knowledge of the specified subjects, not to measure science as a whole. Score reports for each individual subtest were more meaningful and useful than those for the whole test alone. The composite approach made the individual subjects invisible, which was against the purpose and construction of the test.

2. When the composite approach was applied, the responses to the 50 items did not fit the model’s expectation. In contrast, treating the items in each subtest as measuring the same latent trait and each subtest as measuring a distinct latent trait yielded a better model–data fit.

3. Despite the poor model–data fit for the composite approach, the test reliability obtained from the composite approach was .89, only slightly larger than

<table>
<thead>
<tr>
<th>Subtest</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Biology</td>
<td>—</td>
<td>.93</td>
<td>.89</td>
<td>.97</td>
<td>.94</td>
</tr>
<tr>
<td>2. Health education</td>
<td>.45</td>
<td>—</td>
<td>.87</td>
<td>.88</td>
<td>—</td>
</tr>
<tr>
<td>3. Physics and chemistry</td>
<td>.64</td>
<td>.51</td>
<td>—</td>
<td>—</td>
<td>.93</td>
</tr>
<tr>
<td>4. Earth science</td>
<td>.54</td>
<td>.43</td>
<td>.59</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>5. Environmental science</td>
<td>.48</td>
<td>.40</td>
<td>.55</td>
<td>.45</td>
<td>—</td>
</tr>
</tbody>
</table>

Note. The correlations in the lower triangle are the attenuated correlations obtained from the unidimensional approach. The first and second correlations in each cell of the upper triangle are the disattenuated correlations obtained from the unidimensional approach and the direct estimates obtained from the multidimensional approach, respectively. The disattenuated correlation may exceed unity using the disattenuation formula. In this case, it should be treated as unity.
the reliabilities obtained from the multidimensional approach, which were between .76 and .86. The improvement in measurement precision was limited, compared with the expense that the individual latent traits became invisible.

4. Finally, high correlations between latent traits do not necessarily mean that they are the same construct. For instance, even if height and weight are almost perfectly correlated, they are still quite different concepts (though the information in them is mostly redundant for some purposes).

For short tests, estimates of individual persons are often so imprecise that the testers usually do not present score reports for individual persons but focus instead on score reports for subgroups (e.g., a class, a school, or a school district). In this example, the test reliabilities of those four subtests with seven or nine items were between .43 and .55 when the unidimensional approach was applied. It appeared that these subtests were too unreliable to assess individuals’ performances. When the multidimensional approach was used, these test reliabilities increased to between .76 and .83, which made the score reports for both individuals and subgroups more appropriate.

In the NAEP program, examinees are given different but overlapping tests to avoid fatigue and reduce testing time. The tests that the examinees take are so short that the measure of each examinee is very imprecise. Therefore, individual-based data analysis is not recommended (the NAEP focuses on measuring differences between groups). To partially address this problem, the NAEP adopts a procedure to take collateral information into account to improve measurement precision (Mislevy, 1987, 1991). Also, the NAEP presents five plausible values for each student to represent uncertainty of the measures. Basically, the present study deals with a similar problem of short tests. The NAEP uses background variables to increase measurement precision, whereas the present study uses multiple subtests. As the correlations between latent traits are usually higher than those between background variables and the target latent trait, the improvements in measurement precision should be greater when intertest correlations are taken into account than when collateral information from background variables is used.

Example 2: The Teacher Personality Inventory

In Example 1 we focused on ability tests with dichotomous items. In Example 2 we examine a personality inventory with Likert-type items. The Teacher Personality Inventory (Wang, 1997) contains 10 scales: Nurturance, Affiliation, Inquiry, Change, Democracy, Thoroughness, Emotional Stability, Playfulness, Sensitivity, and Endurance. Each scale has four 4-point Likert-type items. The 10 scales are short because the inventory was designed primarily for self-evaluation of those college students who plan to enroll in teacher-training programs to determine whether they have the personality traits important to becoming a successful teacher. Participants were 297 secondary school teachers and 476 college students. These 10 scales should not be treated as measuring a single latent trait because each scale taps a distinct aspect of personality. Therefore, the composite approach was not appropriate. This data set was analyzed with the rating scale model (Andrich, 1978) using both the unidimensional and multidimensional approaches. The weighted mean-square errors for these 40 items ranged from 0.8 to 1.2 (M = 1.001). The corresponding z statistics were between -2.1 and 2.4. Basically, the data fitted the model’s expectation fairly well. Appendix B shows the scoring matrix B and design matrix A and the corresponding ACER ConQuest commands.

Measurement Efficiency

Table 3 shows the test reliabilities of the 10 scales obtained from both approaches and the numbers of comparable items that are needed for the unidimensional approach to achieve the same measurement precision produced via the multidimensional approach. With the unidimensional approach, the test reliabilities of the 10 scales were between .59 and .71. With the multidimensional approach, the reliabilities were much higher, between .76 and .86. The test reliabilities of the 10 scales were 17% to 39% higher when the multidimensional approach was used.

When all 10 scales are analyzed, the multidimensional approach achieves the same measurement precision for the 4-item scales as the unidimensional approach can achieve with 7 to 10 comparable items in each scale. In other words, to achieve the same accuracy, the multidimensional approach requires only 40% to 57% of the numbers of comparable items required for the unidimensional approach. The measurement precision that is achieved with 40 items using the multidimensional approach can be achieved with 82 comparable items using the unidimensional approach.

As in Example 1, the effect of increasing the number of scales on measurement efficiency of the mul-
Table 3
*Test Reliability and Number of Comparable Items Needed for the Unidimensional Approach to Achieve the Same Measurement Precision as the Multidimensional Approach in Example 2*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of items</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Test reliability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unidimensional</td>
<td>.64</td>
<td>.66</td>
<td>.67</td>
<td>.61</td>
<td>.60</td>
<td>.62</td>
<td>.71</td>
<td>.66</td>
<td>.66</td>
<td>.59</td>
</tr>
<tr>
<td>Multidimensional</td>
<td>.84</td>
<td>.80</td>
<td>.86</td>
<td>.85</td>
<td>.79</td>
<td>.85</td>
<td>.83</td>
<td>.82</td>
<td>.81</td>
<td>.76</td>
</tr>
<tr>
<td>% of increment in reliability</td>
<td>31</td>
<td>21</td>
<td>28</td>
<td>39</td>
<td>32</td>
<td>37</td>
<td>17</td>
<td>24</td>
<td>23</td>
<td>29</td>
</tr>
<tr>
<td>Items needed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First two scales</td>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First three scales</td>
<td>7</td>
<td>5</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First four scales</td>
<td>7</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First five scales</td>
<td>7</td>
<td>5</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First six scales</td>
<td>7</td>
<td>5</td>
<td>8</td>
<td>8</td>
<td>6</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First seven scales</td>
<td>7</td>
<td>6</td>
<td>9</td>
<td>10</td>
<td>6</td>
<td>9</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First eight scales</td>
<td>8</td>
<td>6</td>
<td>7</td>
<td>10</td>
<td>7</td>
<td>10</td>
<td>6</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First nine scales</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>10</td>
<td>8</td>
<td>10</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>All scales</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>10</td>
<td>8</td>
<td>10</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>
tidimensional approach over the unidimensional one is illustrated by adding each additional scale one at a time. First, suppose there are only two scales, say, Scales 1 and 2. The measurement precision of these 8 items using the multidimensional approach can be achieved with 10 (i.e., 5 + 5) comparable items using the unidimensional approach. That is, the multidimensional approach requires 80% (i.e., 8/10) of the number of items required for the unidimensional approach as shown in the lower part of Table 3. When Scale 3 is added to these two scales, the multidimensional approach requires 67% of the number of items required for the unidimensional approach. When Scale 4 is added, the multidimensional approach again requires 67% of the number of items required for the unidimensional approach. Likewise, when Scales 5, 6, 7, 8, 9, and 10 are added sequentially, the multidimensional approach requires 65%, 59%, 53%, 53%, 49%, and 49%, respectively, of the numbers of items required for the unidimensional approach. As found in Example 1, the percentage of items needed decreases as the number of scales increases, indicating that the multidimensional approach achieves greater measurement efficiency when the number of scales is large.

**Correlation Matrix of the Latent Traits**

Table 4 shows the correlation matrices obtained from the unidimensional and multidimensional approaches. The attenuated correlations obtained from the unidimensional approach were between .30 and .61, indicating that these 10 latent traits were moderately correlated. After disattenuation, they appeared moderately to highly correlated, with correlations between .46 and .95, which were very close to the correlations obtained from the multidimensional approach, which were between .47 and .90.

**General Discussion and Conclusion**

The multidimensional approach provides more precise measures than the unidimensional approach by using the correlations between latent traits, that is, the off-diagonal elements of the variance–covariance matrix of latent traits. As demonstrated in Examples 1 and 2, the improvement in measurement precision varies as a function of the number of latent traits and test length. The greater the number of latent traits and the shorter the target tests, the more significant the improvements are. Although the effect of the magnitude of correlations between latent traits on the improvement in measurement precision is not directly

---

**Table 4**

<table>
<thead>
<tr>
<th>Scale</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Nurturance</td>
<td>.82</td>
<td>.80</td>
<td>.38</td>
<td>.78</td>
<td>.81</td>
<td>.79</td>
<td>.64</td>
<td>.41</td>
<td>.34</td>
<td>.30</td>
</tr>
<tr>
<td>2. Affiliation</td>
<td>.38</td>
<td>.81</td>
<td>.79</td>
<td>.65</td>
<td>.48</td>
<td>.34</td>
<td>.55</td>
<td>.38</td>
<td>.31</td>
<td>.31</td>
</tr>
<tr>
<td>3. Inquiry</td>
<td>.38</td>
<td>.81</td>
<td>.79</td>
<td>.65</td>
<td>.48</td>
<td>.34</td>
<td>.55</td>
<td>.38</td>
<td>.31</td>
<td>.31</td>
</tr>
<tr>
<td>4. Change</td>
<td>.38</td>
<td>.81</td>
<td>.79</td>
<td>.65</td>
<td>.48</td>
<td>.34</td>
<td>.55</td>
<td>.38</td>
<td>.31</td>
<td>.31</td>
</tr>
<tr>
<td>5. Democracy</td>
<td>.38</td>
<td>.81</td>
<td>.79</td>
<td>.65</td>
<td>.48</td>
<td>.34</td>
<td>.55</td>
<td>.38</td>
<td>.31</td>
<td>.31</td>
</tr>
<tr>
<td>6. Thoroughness</td>
<td>.38</td>
<td>.81</td>
<td>.79</td>
<td>.65</td>
<td>.48</td>
<td>.34</td>
<td>.55</td>
<td>.38</td>
<td>.31</td>
<td>.31</td>
</tr>
<tr>
<td>7. Emotional stability</td>
<td>.38</td>
<td>.81</td>
<td>.79</td>
<td>.65</td>
<td>.48</td>
<td>.34</td>
<td>.55</td>
<td>.38</td>
<td>.31</td>
<td>.31</td>
</tr>
<tr>
<td>8. Playfulness</td>
<td>.38</td>
<td>.81</td>
<td>.79</td>
<td>.65</td>
<td>.48</td>
<td>.34</td>
<td>.55</td>
<td>.38</td>
<td>.31</td>
<td>.31</td>
</tr>
<tr>
<td>9. Sensitivity</td>
<td>.38</td>
<td>.81</td>
<td>.79</td>
<td>.65</td>
<td>.48</td>
<td>.34</td>
<td>.55</td>
<td>.38</td>
<td>.31</td>
<td>.31</td>
</tr>
<tr>
<td>10. Endurance</td>
<td>.38</td>
<td>.81</td>
<td>.79</td>
<td>.65</td>
<td>.48</td>
<td>.34</td>
<td>.55</td>
<td>.38</td>
<td>.31</td>
<td>.31</td>
</tr>
</tbody>
</table>

**Note.** The correlations in the lower triangle are the attenuated correlations obtained from the unidimensional approach. The first and second correlations in each cell of the upper triangle are the disattenuated correlations obtained from the unidimensional approach and the direct estimates obtained from the multidimensional approach, respectively.
demonstrated in these two examples, it can be inferred from Bayesian statistics that the higher the correlation between latent traits, the more collateral information they contain, and thus the greater the improvement will be.

When using the unidimensional approach, the correlations between latent traits cannot be estimated directly. Even though a two-step procedure can be used to estimate the attenuated correlations and then to adjust them for the attenuation caused by imperfect reliability, such estimation of the correlation matrix has been found to be biased (Wang, 1999). In contrast, direct estimation of the variance–covariance matrix of latent traits via the multidimensional approach obviates the problem of the bias introduced by the two-step procedure.

Although test batteries are taken as examples to depict the measurement efficiency of the multidimensional item response model in the present study, the multidimensional approach does not require items to be administered together, nor that they belong to a test battery. Items may be administered at different time points, and the tests may measure quite diverse but correlated latent traits. However, for ease of interpretation, tests providing collateral information should not be so divergent that one might question the face validity of adding information from them to a test taker’s score on some other scale.

If other kinds of collateral information, such as persons’ educational background, status on demographic variables, in-class test grades, or homework grades, are accessible, they could be incorporated into the multidimensional approach to further improve measurement efficiency. Similarly, if scaled scores rather than original item responses to some tests are accessible (e.g., the tests were administered a long time ago), they can also be treated as background variables. More studies are needed to investigate implications and applications when both kinds of information, background variables and additional tests, are used.

In both Examples 1 and 2, all examinees were administered the same tests. In some testing contexts, multiple test forms can be generated and administered to examinees. Linking across multiple forms can be carried out by any common-item or common-person linking procedure (Kolen & Brennan, 1995). This raises no further problem for the multidimensional approach. If some tests were administered long ago but the original item responses are still accessible, all the tests including the recently administrated tests can be jointly calibrated. If tests are administered repeatedly (as in longitudinal studies), all the item responses can be calibrated jointly. When this is done, all the test items will be automatically set on a common metric. The item parameters should be checked empirically to determine whether they remain unchanged over time (i.e., no item parameter drift; Bock, Muraki, & Pfeiffenberger, 1988). In general, there are no extra sample size requirements for the multidimensional approach in addition to those for the unidimensional approach. For Rasch-type models, a sample size of 200 is usually sufficient (Wright & Stone, 1979).

When tests contain items measuring more than one latent trait simultaneously, the unidimensional approach (see Figure 1a) and the between-items multidimensional approach (see Figure 1b) used throughout this study are no longer appropriate; instead, the within-item multidimensional approach (see Figure 1c) can be applied. For more applications of the within-item multidimensional approach, see Adams, Wilson, and Wang (1997); Hoijtink, Rooks, and Wilmink (1999); Hoskens and De Boeck (2001); Wang (1999); and Wang, Wilson, and Adams (1997, 2000). Within the family of Rasch models, several multidimensional Rasch models have been proposed, in addition to the MRCML model. For instance, Kelderman and Rijkes (1994) and Kelderman (1996, 1997) have proposed item-scoring functions to describe various kinds of within-item multidimensionality. Rost and Carstensen (2002) presented a multidimensional Rasch model and applied it to an interest questionnaire. Although the developments of multidimensional Rasch models seem to be promising, applications of the within-item multidimensional approach in routine test practice remain unexplored.

In the present study, multidimensionality is specified rather than determined from the data so that the multidimensional approach is confirmatory rather than exploratory. The term exploratory means that the degree to which each item is indicative of each of the dimensions of interest (called the item weight, which is analogous to the factor loading in factor analysis) is estimated from the data. In contrast, the term confirmatory means that the item weight is specified a priori. Exploratory models enable researchers to get an impression of the number of latent traits and of which items are indicative of which latent traits. They are valuable tools to detect dimensionality and determine item weights when little prior knowledge about them is available. Confirmatory models allow researchers to formulate a number of theories about dimensionality and item weights (e.g., as shown in Fig-
ures 1b and 1c), based on the researcher’s theoretical knowledge and previous research findings. Multidimensional Rasch models are confirmatory per se because dimensionality and item weights are specified. If little knowledge is available to specify the dimensionality and item weights, confirmatory models should not be used (Hojitjink et al., 1999). Exploratory and confirmatory analyses of the same data do not necessarily lead to the same results, even if the dimensionality and item weights specified for the confirmatory analysis are correct. For example, if five dimensions (e.g., the five subtests in Example 1) are highly correlated, an exploratory analysis might result in a one-dimensional representation of the persons. A confirmatory analysis will probably support the five separate hypothesized dimensions. An advantage of the confirmatory analysis is that each person can be located in a five-dimensional space. In addition, as demonstrated in Examples 1 and 2, it uses the correlations to improve measurement precision.

The underlying item response models used in this study are Rasch-type models (i.e., one-parameter item response models) because they have good measurement properties (e.g., specific objectivity and sufficient statistics). Unidimensional multiparameter item response models (Birnbauam, 1968; Muraki, 1992; Samejima, 1969) are also widely used to fit educational and psychological test data. The multidimensional approach in this study, using MML estimation to take collateral information into account, is an idea that can be generalized to multidimensional multiparameter item response (MMIR) models (e.g., Béguin & Glas, 2001; Bock, Gibbons, & Muraki, 1988; Fraser, 1988; McDonald, 1982; McKinley & Reckase, 1983; Reckase, 1985, 1997; D. T. Wilson et al., 1991) if the problem of model identification can be resolved properly. MMIR models are analogous to linear factor models. The major difference between these two kinds of models is that the relationship between latent trait and indicator (item) is nonlinear in the former models but linear in the latter. As the orthogonal factor structure (i.e., factors are constrained to be independent) is imposed in exploratory factor analysis for model identification, the dimensions in MMIR models are usually constrained to be orthogonal. Under this constraint, the derived dimensions are independent but they are also not readily interpretable because they are usually not consistent with the original test constructs that the test items were designed to measure. Once the dimensions are constrained to be independent, no collateral information from the interdimensional correlations is available so that the multidimensional approach outlined in this study is no longer useful. Only if the orthogonality constraint is released will the multidimensional approach provide better measurement precision. To release this constraint, one should specify the dimensionality of at least part of the items, and then one can allow the other items to measure all the dimensions (subject to model identification constraint), which is analogous to confirmatory factor analysis in which the factor structure is specified rather than discovered from the data. In particular, the simple structure shown in Figure 1b is identifiable even for MMIR models. To our knowledge, no commercial computer programs for MMIR models allow users to easily specify the dimensionality of certain items so as to release the orthogonality constraint.

Also, most MMIR models are limited to dichotomous items. Several researchers have extended MMIR models to polytomous items, for example, Muraki and Carlson (1995) described estimation procedures, and Muraki (1993) developed a beta version of the computer program POLYFACT, for multidimensional polytomous items. As for the other MMIR models for dichotomous items, the same orthogonality constraint on dimensions is also made for model identification. Future studies may investigate how to implement the multidimensional approach on within-item multidimensional Rasch-type models (i.e., see Figure 1c) and how to release the orthogonality constraint in MMIR models so that collateral information from the interdimensional correlations can be taken into account.

References


Mislevy, R. J., & Sheehan, K. M. (1989b). The role of col-


(Appendices follow)
Appendix A

Formulation of the Scoring and Design Matrices and the ACER ConQuest Commands for Example 1

For the sake of simplicity, consider two subtests, each with four dichotomous items. The responses to these eight items are located in column 1 through column 8 in the data file. The first four items belong to the first subtest and the other four items to the second subtest. Assume the items in each subtest follow the Rasch model. With the multidimensional approach, the scoring and design matrices \([B|A]\) look like this:

<table>
<thead>
<tr>
<th>Subtest</th>
<th>Item</th>
<th>Score</th>
<th>(b_{ni})</th>
<th>(a_{di})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
</tbody>
</table>

\[f(X_{ndi} = 1) = \frac{\exp(\theta_{nd} - \xi_{di})}{1 + \exp(\theta_{nd} - \xi_{di})}\]  \(A1\)

where \(X_{ndi}\) denotes the response to Category 2 (correct answer) of item \(i\) in dimension \(d\) for person \(n\), \(\theta_{nd}\) denotes person \(n\)’s latent trait level on subset \(d\), and \(\xi_{di}\) denotes the item difficulty of item \(i\) in subtest \(d\).

With the unidimensional approach, the scoring and design matrices \([B|A]\) for each subtest look like this:

<table>
<thead>
<tr>
<th>Item</th>
<th>Score</th>
<th>(b_{ni})</th>
<th>(a_{di})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
</tbody>
</table>

The commands for the multidimensional approach look like this:

SET CONSTRAINTS=CASES; /* Constrain the means of the person parameters to be zero for model identification. */
DATAFILE RASCH.DAT;
FORMAT RESPONSE 1–8;
SCORE (0,1) (0,1) () !ITEMS (1–4);
SCORE (0,1) () (0,1) !ITEMS (5–8); /* Items 1 through 4 belong to Subtest 1, and items
5 through 8 belong to Subtest 2. */
MODEL ITEM; /* the Rasch modeling */
ESTIMATE;
SHOW >> RASCH.SHW; /* Set results to the output file. */
SHOW CASES! ESTIMATES=EAP >> RASCH.EAP; /* Set EAP estimates to the output file. */
QUIT;

The commands for the unidimensional approach look like this:

SET CONSTRAINTS=CASES;
DATAFILE RASCH.DAT;
FORMAT RESPONSE 1–4; /* Run again for
the second subtest, replacing this line
with: FORMAT RESPONSE 5–8; */
MODEL ITEM;
ESTIMATE;
SHOW >> RASCH1.SHW;
SHOW CASES! ESTIMATES=EAP >> RASCH1.EAP;
QUIT;

Appendix B

Formulation of the Scoring and Design Matrices and the ACER ConQuest Commands for Example 2

Consider 10 scales, each with four 4-point Likert-type items. The responses to these 40 items are located in column 1 through column 40 in the data file. The first 4 items belong to the first scale, the second 4 to the second scale, and so on. Assume the items in each scale follow the rating scale model. With the multidimensional approach, the commands look like this:

DATAFILE RATING.DAT;
FORMAT RESPONSE 1–40;
SCORE (0,1,2,3) (0,1,2,3) () () () () () () () !ITEMS (1–4);
SCORE (0,1,2,3) (0,1,2,3) () () () () () () () !ITEMS (5–8);
SCORE (0,1,2,3) () () (0,1,2,3) () () () () () !ITEMS (9–12);
SCORE (0,1,2,3) () () () (0,1,2,3) () () () () !ITEMS (13–16);
SCORE (0,1,2,3) () () () () (0,1,2,3) () () () !ITEMS (17–20);
SCORE (0,1,2,3) () () () () () (0,1,2,3) () () !ITEMS (21–24);
SCORE (0,1,2,3) () () () () () () (0,1,2,3) () () !ITEMS (25–28);
SCORE (0,1,2,3) () () () () () () () (0,1,2,3) () !ITEMS (29–32);
SCORE (0,1,2,3) () () () () () () () () (0,1,2,3) !ITEMS (33–36);
/* Each set of four items measures the same latent trait. */
MODEL ITEM+STEP: /* the rating scale modeling */
ESTIMATE ! METHOD=MONTECARLO, NODES= 5000;
SHOW >> RATING.SHW; /* Set results to the output file. */
SHOW CASES! ESTIMATES=EAP >> RATING.EAP; /* Set EAP estimates to the output file. */
QUIT;

The commands for the unidimensional approach look like this:

SET CONSTRAINTS=CASES;
DATAFILE RATING.DAT;
FORMAT RESPONSE 1–4; /* Run again for the second scale, replacing this line with: FORMAT
RESPONSE 5–8; and so on until the last scale. */
MODEL ITEM+STEP;
ESTIMATE;
SHOW >> RATING1.SHW;
SHOW CASES! ESTIMATES=EAP >> RATING1.EAP;
QUIT;

Received January 11, 2001
Revision received August 26, 2003
Accepted September 24, 2003