JOURNAL OF APPLIED MEASUREMENT, *4*(1), 87–100
Copyright© 2003

Briggs, D. C. & Wilson, M. (2003). An Introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*(1), 87-100. **Copyright 2003 by the *Journal of Applied Measurement*. Posted with the permission of the publisher.**

An Introduction to Multidimensional Measurement using Rasch Models

Derek C. Briggs

Mark Wilson
University of California, Berkeley

The act of constructing a measure requires a number of important assumptions. Principle among these assumptions is that the construct is unidimensional. In practice there are many instances when the assumption of unidimensionality does not hold, and where the application of a multidimensional measurement model is both technically appropriate and substantively advantageous. In this paper we illustrate the usefulness of a multidimensional approach to measurement with the Multidimensional Random Coefficient Multinomial Logit (MRCML) model, an extension of the unidimensional Rasch model. An empirical example is taken from a collection of embedded assessments administered to 541 students enrolled in middle school science classes with a hands-on science curriculum. Student achievement on these assessments are multidimensional in nature, but can also be treated as consecutive unidimensional estimates, or as is most common, as a composite unidimensional estimate. Structural parameters are estimated for each model using ConQuest, and model fit is compared. Student achievement in science is also compared across models. The multidimensional approach has the best fit to the data, and provides more reliable estimates of student achievement than under the consecutive unidimensional approach. Finally, at an interpretational level, the multidimensional approach may well provide richer information to the classroom teacher about the nature of student achievement.

Measuring is a combination of art and science—the art gives us the momentum, and the science keeps us on track. Wright and Masters (1982, p. 3) have identified four basic requirements for measuring:

1. The reduction of experience to a *one dimensional* abstraction,

2. more or less comparisons among persons and items,

3. the idea of linear magnitude inherent in positioning objects along a line, and

4. a unit determined by a process which can be repeated without modification over the range of the variable.

These provide us useful ground rules for the science of measuring, but unfortunately, the art of measuring often hands us something that doesn't quite conform to these fundamental rules.

In general, a latent domain can be deconstructed into subcomponents, and these subcomponents can in turn be deconstructed (see Figure 1), and so on until the number of latent domains requiring estimation may well equal the number of items being administered! In such a scenario when items are allowed to contribute to more than one domain, the number of dimensions are no longer identifiable parameters. The art of assessing dimensionality is to find the smallest number of latent ability domains such that they are both statistically well-defined and substantively meaningful. In the context of classroom assessment, for example, one would want to use dimensions that were sufficiently fine (i.e., so many) that they were instructionally useful, yet not so many that they overwhelmed the teacher (and/or student). Another factor that limits the number of dimensions is the practical issue of administering and scoring items.

The focus of this paper is on maintaining the advantages of these requirements in cases where the instrumentation involves complexities beyond *simple unidimensionality*. By "simple unidimensionality", we mean the case when instrument creators intend that every item measures the same single dimension (note that this is an intent, and would always need to be both theoretically and empirically justified). There are two rather common cases where simple unidimensionality may be a problematic assumption.

First, there is the case when an instrument is designed to be unidimensional, but results in scores that are interpreted multidimensionally. Many instruments are designed with what is termed an

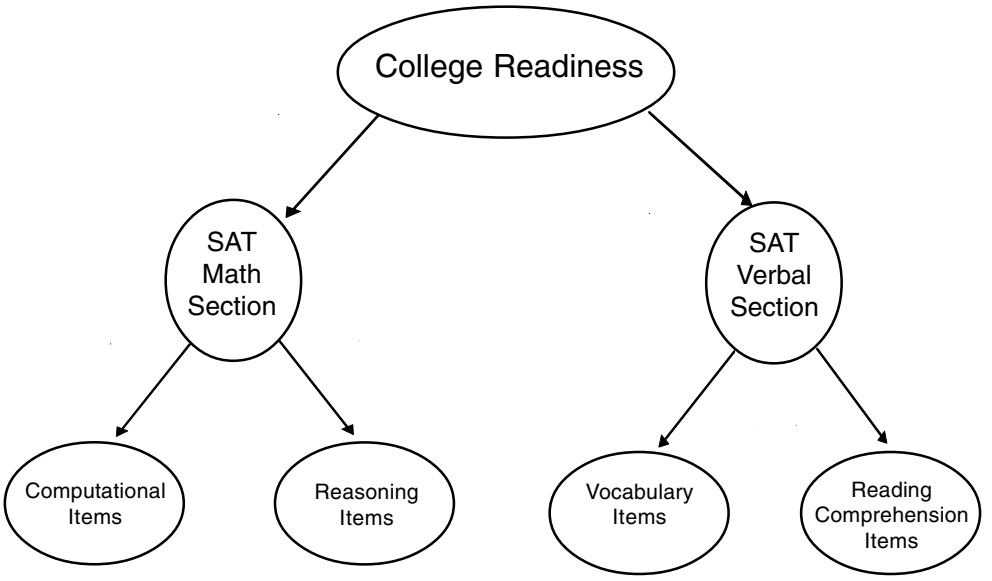


Figure 1. Example of a Multidimensional Taxonomy

"instrument blueprint" that indicates the underlying structure of the domain being measured. These underlying structures are almost invariably some sort of set of sub-dimensions, or even a cross-referencing of sub-dimensions. For example, in achievement testing, it is very common to find a skills-by-content analysis of the target domain, in which case each of the types of skills and content is a potential dimension. It is sometimes very hard to resist the temptation to report scores on these subscales. For example, the Stanford 9 (Harcourt Educational Measurement) reports a single score for its mathematics test, but also subscale scores for "mathematical procedures" and "mathematical problem solving". Also, some technical procedures require using subsets of the items (e.g., linking subsets). In these cases there is a potential for problems if a multidimensional perspective has been overlooked. Ackerman (1992, p. 67) points out that if a test is truly multidimensional, it becomes impossible to rank order test-takers without implicitly or explicitly weighting the dimensions.

Second, some instruments are explicitly designed with items meant to measure multiple domains of ability. The SAT I (The College Board), for example, is administered and reported with two independent sections—math and verbal. Nonetheless the summed score from the two sections is often reported as a single measure of performance. When performance on an instrument has a multidimensional interpretation, then the proper modeling of these as separate, though not necessarily unrelated, dimensions is a prerequisite before a measure can be properly constructed.[1]

The purpose of this paper is to illustrate the usefulness of a multidimensional approach to measurement. We will use the example of a teacher working within a classroom as the central context, but the points we will make generalize far beyond this context. The next section briefly introduces and describes a multidimensional item response model known as the multidimensional random coefficients multinomial logit (MRCML) model. The model is then illustrated with an example from a classroom setting in which students are being taught from a structured science curriculum. The heart of this paper takes a set of assessments administered to students, and analyzes how the interpretation of student performance might change in the context of a multidimensional model.

A Multidimensional Item Response Model

The potential usefulness of multidimensional item response models has been recognized for many years and there has been considerable recent work on the development of multidimensional item response models and, in particular, on the consequences of applying unidimensional models to multidimensional data, both real and simulated (Ackerman, 1992; Ackerman, 1994; Camilli, 1992; Embretson, 1991; Folk and Green, 1989; Kelderman and Rijkes, 1994; Kupermintz, Ennis, Hamilton, Talbert, and Snow, 1995; Luecht and Miller, 1992; Reckase, 1985; Reckase and McKinley, 1991; Walker and Beretvas, 2000). Despite this activity it appears that the application of multidimensional item response models in practical testing situations has been limited. This has probably been due to (a) the statistical problems that have been involved in developing and fitting such models and (b) the difficulty associated with the interpretation of the parameters of existing multidimensional item response models. We address (a) directly below, and then turn to (b) for the remainder of the paper.

The MRCML model and its applications have been previously and more extensively introduced in other settings (Adams, 1997; Wang, 1999; Wang, Wilson, and Adams, 1997). Much of the notation used here is borrowed directly from these sources. The MRCML model is an extension of the Rasch family of item response models, and is built up from a basic conceptual building-block. To illustrate this building block, assume first that for an item i with ordered categories of response indexed by k there corresponds a unique dimension among a larger set of possible dimensions[2] denoted by d ($d = 1,\ldots, D$). The persons responding to a given item are indexed by $P(P,=1,...,P)$. Then, we model the log odds of the probability a person's response in category k of item i (P_{ik}) compared to category k-1

($P_{ik\text{-}1}$) as a linear function of latent ability on that dimension (θ_d), and the relative difficulty of category k (δ_{ik}):

$$\log\left(\frac{P_{ik}}{P_{ik\text{-}1}}\right) = \theta_d - \delta_{ik} \qquad (1)$$

Moreover, each person is measured by a profile of estimates $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_D)$, where the dimensions are allowed to be non-orthogonal. For category k of item i, the associated difficulty, δ_{ik}, indicates the relative difficulty of being in category k as opposed to category k-1, commonly called a "step difficulty." The θ_d in equation 1 represents the latent ability of the person as a function of the dimension of ability mapped onto item i. Thus, for example, in an achievement testing context, the dimensions might be components of the curriculum. The mapping then would indicate that item i was related to only component d, and the value of δ_{ik} would indicate whether it was relatively easier or harder for a student to be classified as achieving category k-1 compared to k.

For a slightly more formal presentation of the MRCML model, we borrow further from the notation developed in Adams, Wilson and Wang (1997). Let items be indexed i = 1, ..., N with each item having $K_i + 1$ possible response categories (k = 0, 1, …, K_i). The random variable X_{ik} is introduced such that

$$X_{ik} = \begin{cases} 1 & \text{if response to item } i \text{ is in category } k, \\ 0 & \text{otherwise.} \end{cases}$$

Then the MRCML model can be written at the item category level as

$$P(X_{ik} = 1; \mathbf{A}, \mathbf{B}, \xi \mid \boldsymbol{\theta}) = \frac{\exp(\mathbf{b}_{ik}\boldsymbol{\theta} + \mathbf{a}'_{ik}\xi)}{\sum_{k=1}^{K_i} \exp(\mathbf{b}_{ik}\boldsymbol{\theta} + \mathbf{a}'_{ik}\xi)}. \qquad (2)$$

where q has been collected into a D by 1 column vector with D corresponding to the number of hypothesized dimensions in a given instrument. Item and category parameters represented by δ_{ik} in (1) have been gathered into the vector $\boldsymbol{\xi}$. The matrices **A** and **B** are known as the scoring and design matrices respectively, and are used to specify the functional form of the model relative to the hypothesized mapping of items to dimension (**B**) and difficulty parameters (**A**). The design and scoring matrices give the MRCML model the flexibility to represent a wide range of Rasch family models, including multidimensional forms of the Dichotomous Rasch Model, the Rating Scale Model (Andrich, 1978), the Partial Credit Model (Masters, 1982), the Facets Model (Linacre, 1989), the Linear Logistic Test Model (Fischer, 1973), the Item Bundle Model (Wilson and Adams, 1995), and others designed for more complex situations involving raters and other measurement features. The MRCML model is a direct extension of the unidimensional random coefficient multinomial logit (RCML) model (Adams and Wilson 1996), which takes the same form except that the scoring vector $\mathbf{b}_{ik}$, and ability vector $\boldsymbol{\theta}$ in (2) are modeled as scalar values.

The item parameters and the population means and variances of the new theta parameters in the MRCML model are estimated by the marginal maximum likelihood technique. Log response probabilities are summed up over items and persons into a likelihood function. Maximum likelihood estimates and asymptotic standard errors are found iteratively using the EM algorithm (Bock and Aitken, 1981; Dempster, Laird, and Rubin, 1977). A detailed discussion of parameter estimation in the MRCML can be found in Adams, Wilson and Wang (1997). We now describe a context which we will use to illustrate certain features of multidimensional measurement.

Methods

Instrumentation and Sample

The Science Education for Public Understanding Program (SEPUP) is based at the Lawrence Hall of Sciences on the campus of the University of California at Berkeley. Starting in the early 1990s, SEPUP staff began development of a science curriculum for middle school students called "Issues, Evidence and You" (IEY). A focus of IEY is on making middle school science a more hands-on, issue oriented experience. A prominent feature of IEY is that unlike most secondary school curricula, an embedded assess-

ment system has been designed concurrently with the curriculum itself. Assessment developers from the Berkeley Evaluation and Assessment Research (BEAR) Center in UC Berkeley's Graduate School of Education have worked with SEPUP curriculum developers and science teachers using the curriculum to design what Roberts, Wilson and Draney (1997) describe as a "comprehensive, integrated system for assessing, interpreting and monitoring student performance." SEPUP assessments are criterion referenced, hence student performance is rated within a taxonomy of performance categories. For more detail on the IEY curriculum, see SEPUP, 1995. For a deeper look at the principles behind this assessment system, see Wilson and Sloane (2000).

The IEY curriculum has four topic areas: Water Usage and Safety, Materials Science, Energy, and Environmental Impact. SEPUP assessments include embedded assessments throughout, a pre-test at the outset of the IEY curriculum, a post-test at the end, and three "link tests" in between the four topic areas. The link tests are useful to teachers as a means of assessing student performance during the course of a school year. From a technical standpoint, the link tests help to disentangle the effects of students' changing proficiencies from the difficulty of assessment activities. From the standpoint of our discussion of multidimensionality, the IEY link tests are important because all items on the test can be linked to one of four SEPUP "variables"[3]: *Designing and Conducting Investigations* (DCI); *Examining Evidence and Tradeoffs* (ET); *Understanding Concepts* (UC); and *Communicating Scientific Information* (CSI). These four descriptors are variables in the sense that performance along each of them should change as students demonstrate more or less achievement on each. Figure 2 describes the SEPUP variables in greater detail. Here we think of science ability as the higher order latent variable, and the four SEPUP variables as the disaggregated components of this composite variable, much as the example illustrated in Figure 1. This structure suggests the desirability of a theory-driven confirmatory analysis, with each SEPUP variable treated as a unique, but non-orthogonal dimension of science ability.

Students taking the IEY curriculum are given three link tests during the course of a school year, usually at a transition point between topic areas. On each link test there are five open-ended items. As an example, the first item from the first IEY link test reads:

Scientific Process:

Designing and Conducting Investigations (DCI)—designing a scientific experiment, performing laboratory procedures to collect data, recording and organizing data, and analyzing and interpreting the results of an experiment.

Evidence and Tradeoffs (ET)—identifying objective scientific evidence as well as evaluating the advantages and disadvantages of different possible solutions to a problem based on the available evidence.

*Scientific Concept*s:

Understanding Concepts (UC)—understanding scientific concepts (such as properties and interactions of materials energy, or thresholds) in order to apply the relevant scientific concepts to the solution of problems.

Scientific Skills:

Communicating Scientific Information (CSI)—organizing and presenting results of an experiment, or explaining the process of gathering evidence and weighing tradeoffs in selecting a solution to a problem effectively, and free of technical errors.

Figure 2. The SEPUP Variables

> You must decide which energy source to use to power a car: gasoline, electricity, or natural gas. To do this you must first design an experiment to identify the most efficient source of energy. In the space below, first describe your experimental procedures, including a data table that you would use to organize your data. Then indicate how you would use the data to show you which form of energy is most efficient for running your car.

The answers given by students to such items are scored along one or more of the four SEPUP variables. In this example, the item was scored along the DCI and ET variables. One link test item might yield as many as three different variable scores for a particular student. Student responses are scored with separate scoring guides for each variable, though each scoring guide follows the same general framework (Biggs and Collis, 1982). In the general framework, there are five score categories, ranging from a "0", indicating an "off-task or irrelevant response", to a "4", indicating that a student has "gone way beyond what was expected as a correct answer in some significant way." It is important to note that these scores are hierarchical. That is, students cannot receive a score of a "3" if their responses do not meet all the criteria for a score of a "2". The scoring guide for the SEPUP variable DCI is given in Figure 3. Each item is also accompanied by exemplars for each score level to guide application of the scoring guide.

In what follows, each link test prompt response scored along one of the four SEPUP variables is treated as a distinct item.[4] During the 1994-95 school year, data from three link tests were collected, encompassing responses from 541 students taking the IEY curriculum. Responses from 14 prompts yielded 34 items. Twelve prompts were scored along the DCI variable, 11 along the ET variable, seven along the UC variable, and four along the CSI variable.

Procedures

Two fairly common approaches might be taken in the attempt to assess student performance on the SEPUP link test items: the *unidimensional* approach and the *consecutive* approach. Both of these approaches are illustrated graphically in Figure 4. In the unidimensional approach, the sum of scores received on the 34 items, ranging between

Score	***Designing Investigation:*** Response states problem and general approach for the investigation.	***Selecting and Recording Procedures:*** Response reflects recognition and recording of relevant procedures performed completely, accurately, and safely.	***Organizing Data:*** Response accurately records and logically displays data.	***Analyzing and Interpreting Data:*** Response accurately summarizes data; detects patterns and trends; and draws valid conclusions based on the data used.
4	Accomplishes Level 3 AND goes beyond in significant way, e.g. describing limitations of approach or design, or describing relevant controls and variables.	Accomplishes Level 3 AND goes beyond in significant way, e.g. identifying alternative procedures to effectively carry out test.	Accomplishes Level 3 AND goes beyond in significant way, e.g. innovation in the organization or display of data.	Accomplishes Level 3 AND goes beyond in significant way, e.g. explaining unexpected results, judging the value of investigation, suggesting additional relevance investigation, etc.
3	Includes complete statement of the problem and/or design which demonstrates complete understanding of the problem and the design.	Reflects choice and recording of appropriate procedures completely and accurately.	Logically reflects complete and accurate data; minor errors in data may exist.	Analyzes and interprets data correctly and completely; conclusion is compatible with data analysis.
2	Incompletely states a problem or the design of an experiment.	Reflects appropriate choice; some steps are not fully described OR are omitted.	Reports data logically and may contain minor errors BUT records are incomplete.	Notes patterns or trends but does so incompletely.
1	States incorrect problem that demonstrates lack of understanding of problem or design of investigation.	Indicates incorrect or inappropriate choice and/or recording of procedures.	Reports data BUT records are illogical and/or contain major errors in the data.	Attempts an interpretation, but ideas are illogical OR show a lack of understanding.
0	Does not include problem or design of investigation.	Missing, illegible, or no record of relevant procedures.	Missing, illegible, or no record of data included.	Missing, illegible, or no analysis or interpretation of data included.
X	Student had no opportunity to respond.			

Figure 3. Scoring Guide: Designing and Conducting Investigations (DCI) Variable

0 and 136, could be treated as the sufficient statistic for a single estimate of student ability in science. The probability of a response in each of the five categories of the 34 link test items is

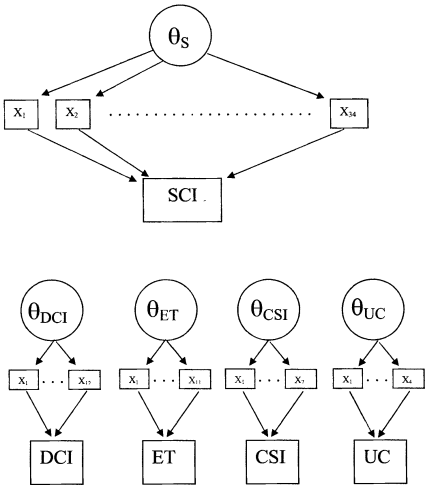
$$\begin{aligned} P(0) &= 1/\gamma \\ P(1) &= \exp(\theta_p - \delta_i - \tau_1)/\gamma \\ P(2) &= \exp(2\theta_p - 2\delta_i - \tau_1 - \tau_2)/\gamma \\ P(3) &= \exp(3\theta_p - 3\delta_i - \tau_1 - \tau_2 - \tau_3)/\gamma \\ P(4) &= \exp(4\theta_p - 4\delta_i)/\gamma \end{aligned} \quad (3)$$

where γ = the sum of the numerators in the five equations. Item difficulty has been divided into two components, the average difficulty of the item (δ_i), and the incremental, or "step" difficulty of a response in a higher category of item i (τ_1, τ_2, τ_3, τ_4). The sum of the step difficulties for each SEPUP variable are constrained to equal zero to allow for parameter identification, hence τ_4 is set equal to the negative sum of τ_1, τ_2, τ_3 and does not need to be estimated directly.

Modeling the RCML equations with the software package Conquest (Wu, 1998) produces estimates for a total of 47 parameters—45 item (δ) and step (τ) difficulties, one population mean, and one population variance. Three different step difficulties are estimated for each of the four SEPUP variables because each variable was scored under differing scoring guide criteria.[5] As we noted in equation 3, each item has one parameter, δ_i, and each variable has three parameters, τ_1, τ_2, τ_3. Note that the sum of the item difficulties has also been constrained to zero for parameter identification, resulting in one fewer free item difficulty parameters. An equally acceptable identification approach would have been to constrain the population mean of student ability (θ) to equal zero.

The unidimensional approach has the advantage of parsimony in modeling student performance, summarizing student achievement with a single number and its associated standard error. Furthermore, the reliability of student ability estimates is quite high at .90. Yet a clear disadvantage is that the differential information about student achievement relative to the profile of four SEPUP variables is lost.

As an alternative, one might model science ability using the consecutive approach. In this approach, the sum of scores on items associated with each of the four SEPUP variables are treated as four different sufficient statistics and modeled independently as unidimensional constructs. So in essence, the consecutive approach is just the unidimensional approach repeated for each hypothesized dimension, and the RCML model above can again be applied to estimate item and person parameters. In such an approach 50 parameters would be estimated: four separate means and variances for student ability plus a total of 30 item



Unidimensional Approach

SCI = Sum of raw scores on all 34 link test items for student p
X_i = Individual link test items
θ_S = Single estimate of latent science ability

Consecutive Approach

DCI = Sum of raw scores on 13 DCI items for student p
ET = Sum of raw scores on 12 ET items for student p
CSI = Sum of raw scores on 4 CM items for student p
UC = Sum of raw scores on 7 UC items for student p
θ = Four *independent* estimates of latent science ability

Figure 4. Modeling Science Ability

parameters and 12 step parameters. Note that the parameter estimates from the unidimensional approach could also be used in a similar way. This would effectively be modelling the projection of each subscale into a single dimension.

The consecutive approach has the advantage of producing ability estimates and standard errors for each of the SEPUP variables. Yet the consecutive approach ignores the possibility that performance across the SEPUP variables might be interrelated. When the number of items defining each dimension is small, the standard errors of the consecutive estimates are substantially larger than a combined unidimensional estimate. As Table 1 indicates, the upshot of this is a reduction in reliability[6] for each SEPUP variable relative to the unidimensional reliability estimate. The reduction in reliability is substantial: .19 for the DCI variable, .16 for the ET variable, .12 for the CSI variable, and .21 for the UC variable.

The *multidimensional* approach can be viewed as a compromise between the unidimensional and consecutive approaches, one that incorporates the best of both approaches; the scores on each variable are treated as distinct information about each student, yet by incorporating the correlation between the latent variables, the loss in reliability for each of the four SEPUP variable ability estimates is small relative to the unidimensional composite estimate. Figure 5 illustrates the multidimensional approach. Note how there is a direct influence of the DCI latent variable on the DCI observed responses through the straight arrows, but that there is also an influence from each of the other latent variables through the correlations represented by the curved lines. This approach can be modeled using MRCML to estimate latent abilities across the four SEPUP variables simultaneously. Applying (2) yields the following set of equations at the item level.

$$
\begin{aligned}
P(0) &= 1/\gamma \\
P(1) &= \exp(\theta_{pd} - \delta_i - \tau_{1d})/\gamma \\
P(2) &= \exp(2\theta_{pd} - 2\delta_i - \tau_{1d} - \tau_{2d})/\gamma \\
P(3) &= \exp(3\theta_{pd} - 3\delta_i - \tau_{1d} - \tau_{2d} - \tau_{3d})/\gamma \\
P(4) &= \exp(4\theta_{pd} - 4\delta_i)/\gamma
\end{aligned}
\tag{4}
$$

where γ is again the sum of the numerators in all five equations.

Here the probability of a response in any of the five item categories is a function of a student's latent ability (θ_{pd}) on the SEPUP variable associated with item i, and item and step difficulty ($\delta_i + \tau_{kd}$). Note that the MRCML item equations are identical to the RCML equations, except that both θ and τ are now a function of the SEPUP variable (i.e.

Table 1

SEPUP Dimension	Consecutive Reliability
DCI	.71
ET	.74
CSI	.78
UC	.69

Unidimensional Reliability = .90

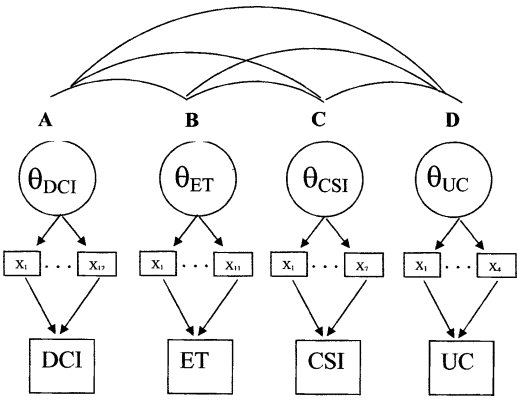


Figure 5. Modeling Science Ability

Multidimensional Approach

DCI = Sum of raw scores on 13 DCI items for student p
ET = Sum of raw scores on 12 ET items for student p
CSI = Sum of raw scores on 4 CM items for student p
UC = Sum of raw scores on 7 UC items for student p
θ = Four *correlated* estimates of latent science ability
AB, AC, AD, BC, BD, CD = Dimensional correlations

dimension) upon which the item has been scored. There are now four populations means for each SEPUP variable, four variance estimates, and six covariance estimates. The four sets of item difficulties are all constrained to sum to zero. Altogether, a total of 56 parameters would be estimated using ConQuest: 42 item and step parameters, and 14 population (person) parameters.[7]

Results

Table 2

MODEL	AIC	G^2	# of parameters
Unidimensional	33470	33564	47
Consecutive (4)	33442	na	51
Multidimensional	33087	33207	60

Because the multidimensional approach is hierarchically related to the unidimensional approach, the model fit can be compared relative to the change in the deviance (G^2) value, where the difference in deviance between the two models is approximately distributed as a chi-square with 13 degrees (60-47) of freedom. As Table 2 indicates, the difference in deviance between the two models is 357. This suggests that the multidimensional model fits the data significantly better than the unidimensional model. On the basis of a comparison of Akaike's Information Criterion (Akaike 1981), the multidimensional model fits the data better than both the unidimensional and the consecutive model. These indicators of statistical significance lead one to expect that there will also be differences between the models at the interpretational level. We illustrate several of these below with respect to a) reliability, b) estimated correlations among dimensions, and c) student ability estimates under the different models.

Earlier the claim was made that an advantage of the multidimensional approach is an improvement in reliability relative to the consecutive approach, and Table 3 supports this claim. Under the multidimensional approach, the reliability for each SEPUP dimension comes closer to the unidimensional reliability estimate. In the case of the ET variable, the reliability of the multidimensional ability estimate is actually equal to the unidimensional estimate. For the other three SEPUP variables, the reliability of the multidimensional ability estimates fall between those estimated using the consecutive and unidimensional approaches. This result is consistent with a previous investigation by Wang, Wilson and Adams (1997).

Table 3

SEPUP Dimension	Consecutive Reliability	Multidimensional Reliability
DCI	.71	.83
ET	.74	.90
CSI	.78	.80
UC	.69	.79

Unidimensional Reliability = .90

The multidimensional and consecutive approaches can also be compared with respect to estimated correlations between the SEPUP variables. The numbers below the diagonal of the four by four matrix in Table 4 shows the correlation between each of the four SEPUP variables under the multidimensional model. The numbers above the diagonal of Table 4 show the variable correlations calculated from consecutive model estimates. Unlike the variance-covariance matrix produced by ConQuest using the multidimensional model, this latter set of correlations is attenuated due to measurement error. Hence in the consecutive approach the true correlations between the SEPUP variables are underestimated[8]. With respect to the disattenuated correlations on the lower diagonal of Table 4, all four variables enjoy moderate to strong correlations, ranging from about .6 to .8. The lowest correlations tend to involve the CSI variable.

Restricting our attention to the multidimensional model, Figure 6 plots the standardized ability estimates of the SEPUP students in logit

Table 4

Correlations Between SEPUP Variables

	DCI	ET	CSI	UC
DCI	1.00	.55	.45	.60
ET	.73	1.00	.66	.55
CSI	.59	.83	1.00	.43
UC	.81	.79	.64	1.00

Multidimensional Correlations below diagonal
Consecutive Correlations above diagonal

values (positive values indicate higher estimated ability) for the DCI dimension against the standardized estimates of the ET dimension. In the SEPUP example it seems that in general, a high estimated ability in one dimension implies a high estimate of ability in another dimension. Nonetheless, it is worth noting that while the two dimensions exhibit a fairly strong correlation at .73, there are a significant number of students who differ by a full standard deviation when their ability estimates are compared across dimensions. In four dimensions, the number of "discrepant" cases might be larger still. To capture this, standardized ability estimates on all four dimensions can be compared for the full sample of 541 students using a sum of squares indicator, DI for each student.

$$\mathrm{DI}_p = \sum_{d=1}^{4} \left(\bar{\theta} - \theta_d\right)^2 \qquad (5)$$

If we arbitrarily set the threshold for a discrepant case at $\mathrm{DI}_p = .5$, there are 162 students (30%) in the sample for whom dimensional estimates might reveal differing stories about underlying student ability. To make the point more concrete, consider an example. Table 5 shows ten students from the SEPUP sample who had DI values greater than .5. Each of the four columns of the table give the students' corresponding standardized dimensional ability estimates in logit values. The values in each of the four columns represent related aspects of the overall domain

Table 5

Discrepant Cases: Multidimensional Ability Estimates

$\hat{\theta}_{DCI}$	$\hat{\theta}_{ET}$	$\hat{\theta}_{CSI}$	$\hat{\theta}_{UC}$
.76	-.71	-1.23	.83
.69	-.96	-1.33	-.11
.17	-1.01	-1.06	.43
.78	1.98	2.17	.96
-.15	.97	.93	-.22
1.32	2.18	2.38	1.26
-.43	.21	.14	-.83
.14	-.56	-.98	-.40
.71	1.00	1.16	.13
.95	1.19	.55	1.52

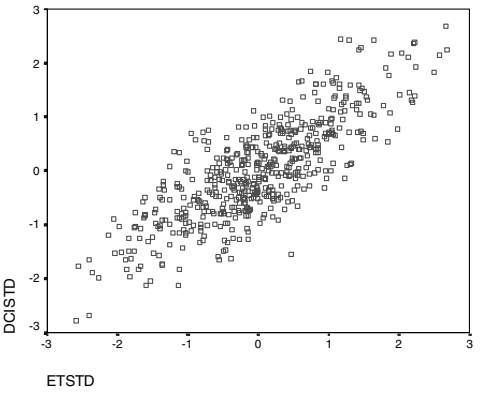


Figure 6. Standardized Student abilitly Estimates

of science achievement, and as such they are generated from a multidimensional model. It should be clear upon inspection that interpreting any one of these estimates as a general, unidimensional measure of science ability is likely to either underestimate or overestimate student ability on the excluded dimensions.

Analyzing examples of discrepant cases is an indication that there might be important educational ramifications of ignoring multidimensionality when assessing student performance. For many students in the SEPUP sample, a teacher might do well to represent performance as unidimensional. Because the four SEPUP variables are positively correlated, most students who perform well on one variable also perform well on the others. However, in this example we suggest such an analysis might be misleading for roughly 1/3 of the students in the SEPUP sample. This might seem reasonable in a statistical sense, but will be problematic for teachers hoping to respond to the needs of all their students.

Consider the simulated case of two such students, "John" and "Shelly." The response vectors of these students to the SEPUP link items are specified a priori such that John and Shelly have DI values of .54 and 1.32 respectively. As shown in Table 6, the DI value of .54 for John corresponds to raw scores of 30 points on the 13 DCI items, 24 on the 12 ET items, 10 on the four CSI items, and 14 on the seven UC items. Applying the consecutive model, one could use these

four numbers as the basis for four separate ability estimates:

$$\hat{\theta}_{DCI_j},\ \hat{\theta}_{ET_j},\ \hat{\theta}_{CSI_j}, \text{ and } \hat{\theta}_{UC_j}$$

Now consider John's classmate, Shelly, with a DI value of 1.32. Shelly receives raw scores of 30, 34, 12 and 22 on the DCI, ET, CSI and UC items. Four separate ability estimates are again generated under the consecutive model:

$$\hat{\theta}_{DCI_s},\ \hat{\theta}_{ET_s},\ \hat{\theta}_{CSI_s}, \text{ and } \hat{\theta}_{UC_s}$$

A comparison of ability estimates along the four variables shows that while

$$\hat{\theta}_{ET_s} > \hat{\theta}_{ET_j},\ \hat{\theta}_{CSI_s} > \hat{\theta}_{CSI_j}, \text{ and } \hat{\theta}_{UC_s} > \hat{\theta}_{UC_j},$$

note that $\hat{\theta}_{DCI_s} = \hat{\theta}_{DCI_j}$ (.244).

In words, Shelly's achievement in science is greater than John's when compared relative to the ET, CSI and UC variables, but equal to John's on the DCI variable. This last relationship changes if a multidimensional approach is adopted because the correlation between dimensions affects the dimensional ability estimates. Now, since the four variables are positively correlated,

$$\hat{\theta}_{DCI_s} > \hat{\theta}_{DCI_j},$$

where $\hat{\theta}_{DCI_s} = .487$ and $\hat{\theta}_{DCI_j} = -.079$.

The estimate of Shelly's science achievement is now greater than that on John's even though they have identical raw scores on the DCI items, because the MRCML model takes into account the higher ability estimates for Shelly along the other three variables. Such simulated results are only suggestive because the dimensional estimates using the SEPUP data have not been made with great precision. The standard error of the estimates is shown in parentheses for Table 6, and for each student it is approximately .3 logits. The two θ_{DCI} estimates are not statistically distinct at the conventional .05 significance level.

Discussion

In many assessment situations there is a desire to either disaggregate the scores from a test into subcomponents and report them as separate dimensions of performance; or to aggregate the scores of test subcomponents and report this as a single dimension of performance. Both these scenarios are a departure from the measurement ideal of simple unidimensionality. The RCML and MRCML models are useful tools when the objective is to measure more than one latent domain. In this paper, the RCML model has been used to illustrate what we term the consecutive approach, while the MRCML model has been used to illustrate the multidimensional approach. Our example taken from the SEPUP assessment system highlights some important differences between the two approaches. The essential difference is that the consecutive approach is simply a unidimensional model repeated a number of times using subsets of the full range of items on a given instrument. Because there are fewer items defining each latent domain, the standard error of measurement for person estimates is necessarily larger, and the reliability of the estimates is smaller than the full unidimensional model. We present the multidimensional approach as an improvement over the consecutive approach. The approach provides distinct estimates for multiple latent domains, yet by modeling the domains as

Table 6

Comparing Ability Estimates of Two Hypothetical Students

	"JOHN"			"SHELLY"		
SEPUP Dimension	Raw Score on Items	Consecutive Abilitly Estimate	Multidimensional Ability Estimate	Raw Score on Items	Consecutive Ability Estimate	Multidimensional Ability Estimate
DCI	30	.244 (.24)	-.079 (.33)	30	.244 (.24)	.487 (.33)
ET	24	-.865 (.37)	-.871 (.36)	34	.497 (.35)	.971 (.36)
CSI	10	-.059 (.70)	-.009 (.54)	12	1.535 (.78)	2.02 (.51)
UC	14	-.418 (.46)	.014 (.38)	22	.848 (.45)	1.51 (.34)

Note: Consecutive and multidimensional ability estimates in logit values, standard errors in parentheses.

interrelated, the reliability of the estimates comes closer to that found under the full unidimensional model.

Beyond the statistical rationale for a multidimensional approach, we believe that there are important interpretational differences as well, particularly in classroom settings. Treating student performance that is multidimensional in nature as unidimensional can have the effect of misrepresenting student ability. This is less of a danger when the dimensions in question have a moderately high and positive correlation, as in the case of the SEPUP data from our example. Nonetheless, we introduce the notion of discrepant cases to show that even when dimensions are well correlated, a relatively large number of students could still have their performance misrepresented. In a high stakes setting, this could have troubling consequences.

The data example used in this paper is probably a simplistic approach to modeling both unidimensional and multidimensional student ability. We ignore what is a likely violation of local item independence in the "bundling" of items within a common prompt. In addition, we ignore the fact that student responses were scored by different teachers, and that these teachers may have scored students with different levels of severity. Finally, student responses are treated as if they had occurred at one point in time when in fact the responses occurred over the course of a school year. This analysis is not an attempt to draw inferences in any absolute sense about student performance in SEPUP. Rather, it has been meant as an illustration of the *relative* merits of a multidimensional approach to a unidimensional approach in the context of classroom measurement.

Footnotes

[1] There is a third case, where individual items are designed to measure two or more dimensions. This has been termed "within-item multidimensionality" (Adams, Wilson and Wang, 1997), and is beyond the scope of this paper.

[2] Note that the MRCML approach allows a broader interpretation than this—the case we are assuming is termed a "between-item" model (Adams, et al., 1997).

[3] There are actually five SEPUP variables. The variable "Group Interaction" has been excluded from this discussion because it was not calibrated within the scope of the SEPUP evaluation project.

[4] This ignores the measurement issue of local item dependence. The methods we discuss can be extended to deal with this, but for the sake of clarity, we will ignore this issue here. See Wilson and Adams (1995) for a way to do this within the MRCML framework.

[5] In general, step difficulty may be modeled for each item separately. In this case, the same scoring guide was used across all items within each SEPUP variable, so a model that keeps the pattern constant within each variable makes sense, and has been found to be empirically accurate (Wilson and Sloane, (2000)).

[6] The reliability coefficient is calculated using an approach described in Mislevy, Beaton, Kaplan and Sheehan (1992) as the ratio of the variance in expected a priori ability estimates from the sample over the estimated variance of the population. The resulting coefficient is conceptually analagous to person separation reliability, and more convenient to calculate given the emprical Bayes, MMLE underpinnings of the MRCML model.

[7] For the example used in this paper, all item and step parameters are anchored to values previously estimated using ConQuest by Draney and Peres (1998). When item anchor values are used, there is no need for model identification constraints, hence in what follows we consider the multidimensional model with respect to a total of 60 parameters rather than 56.

[8] This result is illustrated by simulation studies and theoretical arguments in Adams, Wilson and Wang 1997.

References

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

Ackerman, T. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7*, 255-278.

Adams, R. J., Wilson, Mark, Wang, Wen-chung. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement, 21*(1), 1-23.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Biggs, J. B., and Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Bock, R. D., and Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46*, 443-459.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16*, 129-147.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39* (Series B), 1-38.

Draney, K., and Peres, D. (1998). *Multidimensional modeling of complex science assessment data* (BEAR Research Report). Berkeley: University of California, Berkeley.

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

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

Folk, V. G., and Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement, 13*, 373-389.

Harcourt Educational Measurement. (2002). http://www.hemweb.com/trophy/achvtest/achvindx.htm.

Kelderman, H., and Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59*(2), 149-176.

Kupermintz, H., Ennis, M. M., Hamilton, L. S., Talbert, J. E., and Snow, R. E. (1995). Enhancing the Validity and Usefulness of Large-Scale Educational Assessments .1. Nels-88 Mathematics Achievement. *American Educational Research Journal, 32*(3), 525-554.

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

Luecht, R. M., and Miller, R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement, 16*, 279-293.

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

Mislevy, R. J., Beaton, A. E., Kaplan, B., and Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133-161.

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

Reckase, M. D., and McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15*, 361-373.

Roberts, L., Wilson, M., and Draney, K. (1997). *The SEPUP Assessment System: An overview* (SA-97-1). Berkeley: University of California, Berkeley.

The College Board. (2002). http://www.collegeboard.com/sat/html/students/indx001.html.

Walker, C. M., and Beretvas, S. N. (2000). *Using Multidimensional Versus Unidimensional Ability Estimates to Determine Student Proficiency in Mathematics*. Paper presented at the 2000 Anual Meeting of the American Educational Research Association, New Orleans, LA.

Wang, W.-C. (1999). Direct estimation of correlations among latent traits within IRT framework. *Methods of Psychological Research Online, 4*(2), 47-70.

Wang, W.-C., Wilson, M., and Adams, R. (1997). Rasch models for multidimensionality between items and within items. In G. Englehard, Wilson, Mark (Ed.), *Objective Measurement* (Vol. 4,): Greenwich, CN: Ablex Publishing.

Wilson, M., and Adams, R. (1995). Rasch models for item bundles. *Psychometrika, 60*, 181-198.

Wilson, M., and Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 12*(2), 181-208.

Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis* . Chicago: MESA Press.

Wu, M. L., Adams, R., and Wilson, M. (1998). ACER ConQuest (Version 1.0). Melbourne, Australia: Australian Council for Educational Research.