

Multidimensional modeling of complex science assessment data

Karen Draney and Deborah Peres

University of California, Berkeley

December, 1998

Multidimensional modeling

Introduction

Recent advances in computing technology have allowed for the development and estimation of a wide variety of new item response models, some quite complex. Of particular interest to many measurement and psychometric researchers are multidimensional models. Such models allow for the more realistic exploration of many real-world testing situations, in which person proficiency is not and cannot be assumed to be a unidimensional construct. One area in which multidimensional models may prove particularly useful is that of performance assessment. Students are asked to produce a relatively small number of quite complex performances, as opposed to a large number of relatively simple multiple-choice responses. This presents both researchers and users of such tests with a new set of problems. If performance assessments are scored with a single holistic score, the small number of items can result in low reliability. If a single item is scored on multiple dimensions, this dimensionality must be taken into account when analyzing the resulting data.

The Berkeley Evaluation and Assessment Research (BEAR) Assessment System is an example of an assessment system that is based in large part on performance assessment. This assessment system has been adopted as an integral part of a year-long middle school science curriculum, *Issues, Evidence and You* (IEY), developed by the Science Education for Public Understanding Program (SEPUP). The assessments serve as teaching and learning activities and are embedded throughout the course. These assessments allow teachers and students to track their progress through the year. In addition, they served as part of the

Multidimensional modeling

evaluation procedure used during the field test of the curriculum during the 1994-95 school year (Roberts, Wilson & Draney, 1997).

This report describes some of the issues involved in the use of the Multidimensional Random Coefficients Multinomial Logit (MRCML) model (Adams, Wilson & Wang, 1997) to investigate data collected during the field test of this curriculum and the accompanying assessment system. It also compares results with the unidimensional findings described by Roberts, Wilson and Draney (1997) and Wilson and Draney (1997).

The BEAR Assessment System

The BEAR assessment system is based on the idea of developmental assessment (Wilson, 1998; Wolf, Bixby, Glenn, & Gardner, 1991). Central to developmental assessment is the notion of variables, which are a major focus of instructional and assessment activities. A variable is an achievement continuum defined operationally by the assessment tasks in which student participate, and that can be used to chart student progress over time (Masters, Adams, & Wilson, 1990).

Components of the BEAR assessment system include:

1. **SEPUP Variables:** Five variables that represent student learning in terms of the core concepts of IEY.
2. **Assessment Tasks:** Activities that are an integral part of regular instruction.
3. **Scoring Guides:** Rubrics that establish baseline criteria for assessing levels of student performance.

Multidimensional modeling

4. **Assessment Blueprints:** Sequential list of IEY activities and opportune points in instruction for assessing student learning on one or more of the variables.
5. **Exemplars:** Examples of student work that have been scored and moderated for each variable and score level.
6. **Assessment Moderation:** A process by which teachers discuss and reach consensus on local standards for scoring student work.
7. **Performance Maps:** Graphical representations of student development on the SEPUP Variables.
8. **Link Tests:** Additional assessment activities for teachers' use at major course transitions that are also based on the SEPUP Variables.
9. **Mapped Portfolios:** A collection of student work throughout the year that demonstrates the student's progress through the course, and that includes the performance maps.

A detailed description of all of these components can be found in Roberts, Wilson, & Draney (1997); the most important components for the work described in this report are the variables and their realization via the scoring guides; the link tests; and the implications of this research for the performance maps.

The variables that are central to IEY are the following:

- **Evidence and Tradeoffs (ET):** Identifying objective, relevant scientific evidence, and evaluating the advantages and disadvantages of different possible solutions to a problem based on the evidence available.
- **Designing and Conducting Investigations (DCI):** Designing a scientific experiment to answer a question or solve a problem, selecting appropriate

Multidimensional modeling

laboratory procedures to collect data, accurately recording and logically displaying data (e.g. in graphs and tables), and analyzing and interpreting results of an experiment.

- **Understanding Scientific Concepts (UC):** Recognizing and applying relevant scientific concepts (e.g. threshold, measurement, properties of matter) to an investigation or problem solution.
- **Communicating Scientific Information (CSI):** Organizing and presenting results, arguments, and conclusions in a way that is free of technical errors and effectively communicates with the chosen audience.
- **Group Interaction (GI):** Developing time management skills, the ability to work together with teammates to complete a task (such as a lab experiment) and to share the work of an activity.

Each of these variables is composed of two to four sub-parts known as elements. For example, the Evidence and Tradeoffs variable is composed of two elements: Using Evidence, and Using Evidence to Make Tradeoffs.

The definitions of the variables are contained in the scoring guides, which describe the kind of achievement needed to reach the various scoring levels on the elements of the variables. Teachers use the scoring guides to rate student performance into 5 ordered, qualitatively different categories, labeled 0 through 4. The scoring guides describe the kind of performance that can be expected from students at each of the performance levels. Although each scoring guide is specific to the variable for which it was developed, there is a common structure. A score of 0 indicates an off-task or missing response; a score of 1 indicates performance that is

Multidimensional modeling

incorrect; a score of 2 indicates performance that is generally correct but missing something important; a score of 3 is complete and correct performance; and a score of 4 indicates performance that goes above and beyond what is asked of the student.

The IEY course is divided into four sections, each dealing with somewhat different subject matter: Part 1 is “Water,” Part 2 is “Materials Science,” Part 3 is “Energy,” and Part 4, designed to function as a review section in which students tie together the things that they have learned, is “Environmental Impact.”

Evaluation of student progress through the course requires that increases in task difficulty that are a natural part of the curriculum be disentangled from increases in student proficiency. Therefore a set of 14 unique assessment activities, similar in form to the embedded assessment activities but requiring less time and less directly tied to the curriculum, was developed. Each activity is an open-ended question calling for a paragraph-type response, and each can be scored on multiple elements of the variables. These activities were divided into five overlapping tests, called link tests. Each of these link tests shares at least one activity with the previous and with the next test, to help establish data links. These link tests were used during the field test of the curriculum and assessment system, to help establish baseline difficulties for all activities. The link items are also available to teachers who wish to use them, perhaps along with activities they develop in their classrooms, to structure more formal end-of-unit tests in addition to ongoing embedded assessment activities.

Variable maps are based on the concepts of item difficulty and person proficiency as defined in item response theory (IRT). Because the proficiencies of

Multidimensional modeling

persons and the difficulties of items are on the same scale, they can be directly compared. This allows us to make statements about the kind of work the student has mastered, the kind of work that is above the proficiency of the student, and the level at which the student is currently working. These maps can be used by teachers in classrooms to communicate with students, parents, and administrators, to detect trends in student development, as well as departures from those trends, and to clearly illustrate the current performance of both individual students as well as classes of students.

One of the questions that naturally arises, given the structure of the activities, variables, and scoring guides, is whether it is necessary to take the multidimensional nature of the data into account when constructing these maps. Implications of our analysis for map development, and for the understanding of student progress, will be discussed at the end of the report.

Data collection

Data from this implementation of the BEAR assessment was collected during a field test of IEY during the 1994-95 school year. The field test of the full curriculum, including the assessment system, took place at 6 centers around the country known as Assessment Development Centers (ADCs). Each ADC consisted of four to six teachers, each of whom used the IEY curriculum and the embedded assessment system with at least one middle-school science class. The ADC teachers were required to score link tests and activities, and to meet in monthly moderation meetings. In each ADC classroom, data were collected from both assessment

Multidimensional modeling

activities and link tests. In addition, a separate set of activities was administered in the fall as a pretest, and again in the spring as a post-test. This resulted in at least partial data from approximately 700 students.

In addition, there were seven Professional Development Centers (PDCs) involved in this study, structured similarly to the ADCs, where teachers taught the IEY curriculum and were provided with the same assessment materials as the ADC group, but were not required to score activities or link tests. The PDC teachers met monthly, but their focus was on the curriculum rather than on the assessment system. Each Center, both ADC and PDC, was also asked to choose a comparison teacher, who was to be as similar to the other teachers in the center, but who taught the regular middle-school science curriculum. The pretest and posttest were administered to the PDC and comparison groups, but the link tests were not, nor was data from the assessment tasks collected in these centers (since in the PDCs, the teachers were not required to use them, and in the comparison classes a different curriculum was used). The assessment tasks and link tests were scored on the appropriate variables by the ADC teachers during the field test year. The pretest and posttest were scored by BEAR staff.

It should be noted that during the field test year, data were collected for only four of the five SEPUP variables. Due to the nature of the Group Interaction variable, satisfactory methods for collecting data on this variable were not developed during the field test. Research is currently being undertaken to develop methods of reliably scoring students on their Group Interaction proficiency, and to enable

Multidimensional modeling

teachers to collect data on the performances of their students on this variable. Thus, the analyses in this report will include only data for ET, DCI, UC, and CSI.

Data analysis

Because of the structure of the SEPUP data, most analyses have been done using item response models. The models allow the estimation of student achievement on the same scale throughout the year, even though the sets of items differ between times and were not matched to have identical difficulties. This is done by anchoring the item parameters based on their difficulties at one time, in this case the beginning of the school year. A Rasch-family measurement model was chosen both because it allows this anchoring process and because it allows placing both students and items on this same scale (Wilson & Draney, 1997; Wright & Masters, 1982). Specifically, a modified rating scale model was used. An ordinary rating scale model has the same steps between scores for all items. In the SEPUP system, there are separate scoring guides for each variable, so the modified rating scale model allows different steps for items from different variables (but the same steps for all items in each variable). Although all SEPUP analyses have also included separate step parameters for each variable, earlier analyses all modeled one general latent achievement variable, general SEPUP achievement; general SEPUP achievement was simply achievement estimated by combining items from the four SEPUP variables (DCI, ET, UC, CSI) in whatever proportions they appear in the tests. The research described here tries to examine achievement in each variable separately.

Multidimensional modeling

The analyses to be discussed in this report are as follows: First, data from the link tests only will be analyzed, using first a unidimensional and then a multidimensional model. Discussion of these analyses will include complications that occurred naturally in the data, including: changes in the severity of teacher ratings at different times of the year, changes in student proficiency as the year progressed, and different rates of progress through the curriculum by different classes of students.

Once the analysis of the link tests is complete, data from the embedded assessment activities will be added, and a multidimensional model will be used with this expanded dataset.

Note that, due to the complicated nature of performance assessments, terms like “item” and “task” have a somewhat arbitrary meaning. In the analyses that follow, the term “item” will be used to refer to that which receives a single numerical score on the previously mentioned scale of 0 to 4: a single element of a single variable in an assessment activity or link test. The term “task” or “assessment task” will be used to refer to a single unit of performance that a student is asked to do: an essay or letter to write, a presentation in the classroom, or the answer to a single link test question. There are quite often numerous items attached to a single task, because teachers frequently score a single piece of student work on multiple elements of a variable, and/or on multiple variables. This is no doubt a violation of the local independence assumption associated with item response models; this is a topic that should be addressed in future research. The term

Multidimensional modeling

“activity” will be used as it is in the IEY curriculum: to refer to a single curricular unit, which may encompass one or several assessment tasks, or none at all.

Because students are expected to change over the course of the school year, item difficulties must be anchored at some fixed time. It was decided to obtain estimates of item difficulty for the link test items by using a population of middle school students not yet exposed to the curriculum. A subset of 14 SEPUP tasks, comprising 30 items, including both embedded activities (4 tasks/7 items) and link test questions (10 tasks/20 items), was chosen to use as a pretest with a new sample of students in September of 1995. The 1995 pretest also included 3 new items that were not scored during the 1994-95 field test, bringing the total to 30 items. The 30 items were divided into four overlapping forms, and in the fall of 1995 they were given to 720 seventh, eighth and ninth graders from a subset of the SEPUP schools. All of these pretests were scored by SEPUP Assessment project staff. This method should be roughly equivalent (except for sampling bias/error) to using the 1994-95 students at the beginning of the year.

The teachers who were involved in the ADC groups, and who were administering and scoring the embedded assessment activities during the field test year, were involved in a process called Local Assessment Moderation (see the list of components of the BEAR Assessment System). For a detailed discussion of assessment moderation, its procedures, purposes, and effects see Roberts & Wilson, (Insert appropriate references here); for the purpose of this report it is sufficient to describe it as a form of scorer calibration. Teachers within an ADC met in sessions during which they discussed and came to decisions about standards of performance

Multidimensional modeling

and methods for reliably judging student work (Roberts, Wilson & Draney, 1997). Using samples of responses from their own classes, they discussed what was really meant by the element of the variable being scored and reached consensus about where each sample response fell on the scoring rubric. During the field trial, all the teachers in the SEPUP group participated in these moderation sessions.

Because this process was designed to have an effect on teachers' scoring of student work, the SEPUP team felt that an investigation of changes in rater severity over time was necessary. Four SEPUP development staff members conducted a rescoring study during the summer of 1995. A random sample of 98 of the pretest/posttest pairs, and a random sample of 82 link test series, were rescored. These two samples were drawn independently, and thus may (indeed, were quite likely to) have overlapped, but were not specifically designed to do so. Student responses to items were photocopied, and the resulting copies from all link tests combined, so that the staff scorers could be blinded to the specific time point when the item had been administered. Each of the two sets of responses was rescored during a period of approximately two weeks.

Since the rescoring was done during a short time period, and since the staff did not know the date the response was written, comparing the staff scores to the teachers' scores can show how much of the variation in teacher scores was due to teacher severity changes over the school year, as well as to overall harshness differences between teachers and staff. (This method could also be used to examine differences in severity between teachers, although the current study does not address

Multidimensional modeling

this issue, due to the limited number of rescored tests per teacher. Data is currently being gathered and scored to examine this issue).

An additional complication arose when the progress of the various classes through the course was examined. The ADC classrooms fell into two major groups, varying in how much of the curriculum they covered. The first group consists of nine teachers who finished the second curricular unit (Materials Science) and gave link test II on or before May 1. This gave them about a month of class time before the end of the school year (and the collection of post-curriculum data); during this time they could begin the third curricular unit (Energy). The second group of eight teachers gave link test II at the same time (or within 10 days) of the post-curriculum tests. For this group, link test II is considered together with the end-of-year tests as evidence about post-curriculum achievement, since no material was covered between the two tests, and the two tests were taken at approximately the same time. In addition to these two groups, two classes fell between the fast and slow groups and three classes were dropped because the teachers stopped participating early in the year. Since a major focus of the analyses to be described involves estimating the change in student proficiencies between testing times, it is quite possible that these two groups would perform differently. Therefore, it was decided to use only the first group, which had covered more of the curriculum, in analyzing the link tests. Since all of the students in this group had similar amounts of instructional time between each link test, it seems reasonable to estimate the average change in student proficiency between the two times using only this group. Therefore, all analysis of the link test data will include only this group, consisting of 9 classes with a total of

Multidimensional modeling

267 students. Once the difficulties of the link items are established and can be used as anchors, the entire group of ADC students will be used to estimate the difficulty of the items associated with embedded activities.

Since none of the teachers taught the fourth (Environmental Impact) section, link test III, intended for use between Sections 3 and 4, was not analyzed as a separate point in time for any of the classes. In addition, in classes where it was administered, it was always given in the same week as the posttest. For classes who took link test III, it was considered, together with the end-of-year tests, as evidence of post-curriculum achievement.

Data from the PDC and comparison groups will not be used to estimate any parameters in these analyses. These two groups took only the pretest and posttest, each of which contained a total of 6 items. Either one or two of these items were associated with each of the variables; thus person locations in a multidimensional analysis cannot be reliably estimated using this data; nor is the data particularly useful in estimating parameters for a multidimensional model.

The ConQuest program (Wu, Adams, & Wilson, 1997) was used to fit all of the appropriate MRCML models.

Analysis of the 1995 pretest data

The first model fit to the link test data was unidimensional, a modified rating scale model for which the probability of a student n scoring at level k on an item ($_{nik}$) given that the student obtained level $k-1$ or level k , can be expressed as:

Multidimensional modeling

$$\frac{p_{nik}}{p_{ni(k)} - p_{nik}} = \frac{\exp(\beta_n - \delta_i - \tau_{vk})}{1 + \exp(\beta_n - \delta_i - \tau_{vk})}, \quad (1)$$

with β_n representing SEPUP proficiency of student n , δ_i representing the difficulty of item i (at time 0, the beginning of the school year), τ_{vk} representing step k ($k=1,2,3,4$) for variable v , which is measured by item i , where variable $v \in \{CSI, DCI, ET, UC\}$.

The τ_{vk} are constrained such that $\tau_{v4} = -1 * \sum_{t=1}^3 \tau_{vt}$.

The first step in parameter estimation for the link tests was to fit the above model to the 1995 pretest data. Parameter estimates for this data are given in Table 1. It is worthwhile to note that the step parameters all have relatively large positive t-values (ranging from 2.28 to 8.67). This was also true in earlier analyses (see Wilson & Draney, 1997) and may indicate that the rating scales were not used consistently across all the items in the variable. A partial credit model would describe the inconsistencies more accurately, but has two shortcomings. First, using a different scale for each item does not align with the SEPUP goal of helping teachers use the scoring guide for each variable such that levels of the variable have a consistent meaning across activities and throughout the year. Second, the partial credit model has too many parameters to yield estimates of reasonable precision with the current data.

 Insert Table 1 about here

Multidimensional modeling

The next step in the analysis was to repeat the analysis of the 1995 pretest data, this time using a four-dimensional model, with one dimension (and hence one student proficiency) assigned to each of the four SEPUP variables. The main addition in the multidimensional model over the unidimensional model described above, is that in the multidimensional model each student has four ability levels, one for each variable: DCI_n , ET_n , UC_n , and CSI_n (Adams, Wilson & Wang, 1997). In this analysis, means and standard deviations of the person proficiencies on the four variables and correlations between the four variables were estimated, in addition to item and step parameters. The Monte Carlo method (Volodin & Adams, 1995) as implemented in the ConQuest software (Wu, Adams, and Wilson, 1997) was used to obtain parameter estimates for the multidimensional model. Parameter estimates are given in Table 2. A comparison of the relative fit of the unidimensional and multidimensional models may be made by using the deviance statistics for the two models (found at the bottom of Tables 1 and 2). The difference in the deviance between the two models is distributed as approximately chi-square, with degrees of freedom equal to the number of extra parameters in the multidimensional model. For these two models, the difference in the deviance is 443.07, with degrees of freedom given by $52 - 43 = 9$. This is statistically significant at the .01 level, indicating that the multidimensional model shows better fit to the data. In addition, the misfit statistics for the step parameters are not as large for the multidimensional model as they were for the unidimensional model, which also suggests that the multidimensional model gives a better picture of what is happening in the data.

Multidimensional modeling

Insert Table 2 about here

Analysis of the field test data

The item difficulties and step parameter values estimated from the 1995 pretest data were then used as anchor values for the corresponding items and steps in the 1994-95 field test data. This analysis included the pretest and posttest, the preanchor and postanchor test, and all four of the link tests. As would be expected, not every student completed all seven tests. Of the 267 students in the group analyzed, 23 were missing all information and 29 others had information from only one of the four testing times. Since they contributed little to the estimation, the 23 cases with no student information were dropped. This left a total of 244 cases with at least partial data for this analysis.

We now describe the unidimensional analyses of the link test portion of the 1994-95 field test data. Since the 1994-95 field test data contained longitudinal information from four points in the school year, changes in teachers' scoring severity between different testing times during the school year were estimated. This is similar to the rater parameters used by Wilson and others (Wilson & Case, in press). The teacher rating parameters, δ_{tj} represent the difference between teacher scores at time t and the staff scores of student work from time t , which are presumably not biased by time. Thus the complete unidimensional model can be expressed as:

Multidimensional modeling

$$\frac{p_{nik}}{p_{ni(k)} + p_{nik}} = \frac{\exp(\delta_n - \tau_i - \rho_{vk} - \tau_t)}{1 + \exp(\delta_n - \tau_i - \rho_{vk} - \tau_t)}. \quad (2)$$

The primary parameters of interest are the teacher rating severities τ_t . Although the goal is to estimate these in a multidimensional context, the unidimensional analysis provides comparisons to earlier SEPUP analyses. We also needed to estimate 18 item difficulties δ_n , because not all linking items appeared on the 1995 pretest. Unfortunately, when we used the 1995 pretest anchoring values, and did the estimation with the field test data alone, the solution did not converge. This was probably due to the relatively large proportion of missing data in this dataset. Expanding the data in the following manner enabled us to achieve convergence.

To expand the data, all the cases from the 1995 pretest were added. However, these cases included only the 20 items on the 1995 pretest that also appeared on one of the anchor or link tests. If a 1995 pretest item was given at time 0 and rescored by staff in the field test, the item appeared as the same data column for both 1995 and field test cases. If the item was not given at time 0 or was not rescored in the field test, it appeared in a new data column; these new columns were blank for field test cases. Since the 1995 pretest cases were rated by staff, they have no data for any of teacher ratings. They also have no data for any testing time besides September. Therefore, the 1995 pretest cases should not effect the estimation of τ_t parameters.

With the expanded dataset, we used the EM algorithm (Bock & Aitkin, 1981) in the ConQuest program to get parameter estimates. Two of the anchored difficulty estimates and one of the CSI step estimates fit the model quite poorly (t-values for

Multidimensional modeling

weighted fit statistics greater than 15). The model was re-estimated without anchoring those two difficulty parameters and without anchoring any of the CSI step parameters. When these parameters were unanchored, the estimated values were indeed different from the anchors obtained with the 1995 pretest data. The two item difficulty parameters were for items that appeared only in the time 2 and time 3 tests; this suggests that perhaps the items changed difficulty, relative to other items, between the pretest time and midyear. One of the items, an ET item, was relatively harder in 94-95 than in the 95 pretest; the other, a UC item, was relatively easier.

Parameter estimates for this model are given in Tables 3a and 3b. Figure 1 shows the teacher rating severity estimates for the unidimensional model, surrounded by bands of ± 2 standard errors. The teacher rating estimate is negative for September; this indicates that teachers' scores from that time were less severe than the staff ratings of the same papers. The direction of this effect is the same as that found in earlier RCML analyses (see Wilson & Draney, 1997). One explanation for this is at the beginning of the year teachers did not give low scores, even for poor responses, because they did not want to penalize students for not knowing material that had not yet been taught. For January, the teacher rating estimate is slightly negative, indicating that teachers were still slightly less severe than the staff; though the effect is statistically significant, it is small. The teacher rating estimate for April is not statistically different from zero. The teacher rating estimate for June is slightly positive, indicating that teachers were slightly more severe than staff; again, while this estimate is statistically significantly different from zero, it is fairly small.

Multidimensional modeling

Insert Tables 3a and 3b and Figure 1 about here

To look at student change, we estimated the mean and variance of student achievement at each time by doing a separate analysis at each time, with all difficulty (), step () and teacher rating () parameters anchored at the values obtained from the overall unidimensional model. The results of this analysis are shown in Figure 2

Insert Figure 2 about here

The heavy line in the figure indicates overall achievement, for all four variables combined. In addition, student proficiencies were estimated for each variable at each time separately, by using only those items associated with that variable at that time, anchoring the item difficulties, and estimating person proficiencies.

Note that the overall achievement rises from September to January, and rises slightly more steeply from January to April, but drops from April to June. This drop in achievement during the last month agrees with the findings of earlier analyses (Wilson & Draney, 1997). One possible explanation for this phenomenon is that most of the post-curriculum tests were given during the final week of school, which tends to be a stressful time for both teachers and students. It may be that this interfered with the concentration necessary for students to perform as well as they

Multidimensional modeling

could. Also, the post-anchor test was designed to be given after all four sections of the curriculum, including the review section, were taught. However, the curriculum appeared to be somewhat too long for these teachers and students, at least during the field test year. While classes spend on average from September through December (four months) on the Water section, the amount of time spent on the other sections decreased progressively. Most classes in the current group spent January through April (3 months) on the Materials Science section, at most one month on Energy, and no time on the review section. Thus, one might imagine that the amount of knowledge students were able to construct about each subject of study would decrease proportionally to the amount of time they were able to spend on it. It is, of course, impossible to tell if either of these explanations is correct.

In general, the patterns of achievement for the other variables roughly follow that for the overall achievement variable. UC shows the largest increase overall until time 3, and then drops to the level of the other variables. DCI shows a notable departure from the general pattern at time 0, beginning quite high and then dropping between time 0 and time 1. However, there were very few usable DCI items given at time 0, and thus this departure is not particularly interpretable.

Reliability of person estimates can be calculated using a combination of the estimated population variance, and the variance of the EAP (expected a posteriori) estimates of person ability (Mislevy, Beaton, Kaplan, & Sheehan, 1992). The formula for the reliability is given by:

$$\rho = \frac{\text{var}(EAP)}{\sigma^2}, \quad (3)$$

Multidimensional modeling

where σ^2 is the estimated variance of the population.

Reliabilities were calculated for each variable at each time using the above formula. This was done by estimating person abilities at each time using only the items associated with a particular variable. The reliability of a composite of all four variables at each time was also calculated by using all of the items associated with each time. These reliabilities are given in Table 4. In order to correctly calculate reliabilities at a given time, persons with significant amounts of missing data at that time had to be dropped from the analysis (even if they had no missing data at other times); hence the population variances for the composite of all variables in Table 4 may not quite match the standard deviations given in Table 3b. (There is no reliability estimate for DCI at time 0 because there was not enough data to estimate it).

Insert Table 4 about here

Note that the reliabilities of each variable individually tend to be much lower than that of the combined proficiency variable. The individual reliabilities range from .07 to .28 lower than the overall reliability, with a mean of .16 lower.

The next step in the analysis was to fit a multidimensional model to the link test data. Again, each person was regarded as having four proficiencies, one corresponding to each of the four SEPUP variables. In addition, we allowed the teacher rating parameters to be different for each variable (in the same way the rating scale step parameters were in the unidimensional model).

Multidimensional modeling

The model is multidimensional between items, meaning that each item (teacher score) measured only one variable (Adams, Wilson & Wang, 1997; Wang, Wilson & Adams, 1997). Thus, for any item, the probabilities depend only on the parameters of one variable. For example, for a DCI item, the probability of a student n scoring at level k given that the student obtained level $k-1$ or level k , can be expressed as

$$\frac{\pi_{DCInk}}{\pi_{DCIn(k-1)} - \pi_{DCInk}} = \frac{\exp(\theta_{DCIn} - \delta_i - \rho_{DCIn})}{1 + \exp(\theta_{DCIn} - \delta_i - \rho_{DCIn})}. \quad (4)$$

In order to estimate this model, we used anchor values from the multidimensional analysis of the 1995 pretest. As with the multidimensional analysis of the 1995 pretest data, the multi-dimensional analysis of the 1994-95 field test data was done using the Monte Carlo method (Volodin & Adams, 1995) in the ConQuest program (Wu, Adams, Wilson, 1997). Item difficulty estimates are given in Table 5a; estimates for step parameters, teacher rating parameters, and population parameters (means, standard deviations, and correlations) are given in Table 5b, and a map of the four dimensions is given in Figure 3. The difference in deviance between the unidimensional and multidimensional models is 7894.5, with degrees of freedom $62 - 48 = 14$. This is statistically significant at the .01 level, again indicating that the multidimensional model shows better fit to the data.

Insert Table 5a and 5b and Figure 3 about here

Multidimensional modeling

It is difficult to compare the teacher rating effects between variables and between models (unidimensional versus multidimensional) because the students and the items have a different distribution in each analysis, and in each dimension within the multidimensional analysis (i.e. every dimension has been separately constrained such that the person proficiencies on the 95 pretest were centered at 0). Figure 4 shows all the teacher rating estimates from the unidimensional and multidimensional models. The estimates in Figure 4 are shown as a percent of the range of items, with error bars of two standard errors.

As with the unidimensional model, the teacher rating estimates for September are negative; this is true for the three variables for which sufficient items existed (there was only one usable DCI item during this period)--although only the estimate for the ET variable is large. This indicates that teacher scores from September for the ET variable were quite a bit more lenient than that of staff. SEPUP emphasizes that ET is a variable unique to their curricula, while tasks similar to DCI and UC are more common in other curricula; thus, teachers may have felt unwilling to rate students harshly on tasks with which they were completely unfamiliar.

Unlike in the unidimensional analysis, the teacher rating estimates at other times are not all close to zero, particularly in April and June. In April, teachers appear to have been more severe than staff on the CSI variable, and quite lenient on the DCI variable, while in June it appears that teachers were less severe on all variables but UC, and particularly lenient on CSI. All rating parameters are statistically significantly different from zero, although some are rather small.

Multidimensional modeling

Insert Figure 4 about here

To look at student change, we estimated the mean and variance of student achievement at each time by doing a separate analysis at each time, with all difficulty (), step () and teacher rating () parameters anchored at the values obtained from the overall multidimensional model. The means are shown at the bottom of Table 5b. These means indicate that student change differed across the dimensions. The differences between dimensions can be seen clearly in Figure 5. Between September and January, achievement in DCI and UC rose by about the same amount, while ET dropped. Between January and April the differences are even more striking; UC rose dramatically while DCI, CSI and ET did not change much. UC and to a lesser extent ET showed the drop at the end of the year mentioned in the unidimensional analysis. CSI was essentially flat across the entire year. The different patterns between the variables suggest that reporting progress to teachers and students by variable might allow them to concentrate their efforts more effectively.

Insert Figure 5 about here

Reliabilities for each variable at each time for the multidimensional model were calculated using the formula in Equation 5. These reliabilities are given in Table 6. In general, the reliability for a given variable at a given time is higher using

Multidimensional modeling

the multidimensional model than it was for the corresponding variable and time in the unidimensional model (see Table 4). The largest increase in reliability was .18, the smallest was zero, and the average .09. Many of the multidimensional model reliabilities are approximately as high as the composite reliability using the unidimensional model. This indicates that using a multidimensional model has improved the reliability of person proficiency estimates for individual variables.

Insert Table 6 about here

The final step in the analysis of the 1994-95 field test data was to include the embedded assessment activity data along with the data from the link tests. For this data, only a multidimensional analysis was done. Again, several complications arose. First, there were some embedded assessment activities for which we had no data, most notably, the entire set of activities in the Environmental Impact section (although there were others as well). Second, although the link tests were administered at distinct, and distinctly separate, points in time, with chunks of the curriculum between them, the activities were of course interspersed throughout the curriculum. The assigning of activities to time periods is therefore problematic. As each of the first three sections of the course (Water, Materials Science, and Energy) was divided into subsections, it was decided to analyze each subsection along with the link test to which it was nearest. For example, the Water section was divided into four subsections; the first two were analyzed along with time 0 (along with the pretest and preanchor), and the second two were analyzed along with link test I. For

Multidimensional modeling

the analysis of the embedded assessment data, each time period was analyzed separately. Item difficulties for the link test items, teacher rating parameters, and step parameters were anchored to the values obtained from the multidimensional analysis of the link test data. The difficulties for the embedded activity items were then estimated, along with the means and variances of the four dimensions, and the correlations between them.

Table 7a contains the estimated difficulties and some fit information for the embedded assessment activities (from the analysis including the link tests; link test item parameters are not given in this table because they can be found in Table 6a). Table 7b contains fit information for the step parameters for the three time periods, along with means, standard deviations, and correlations for person proficiency estimates for the three time periods. (Note that some fit information is not available due to difficulties with the current version of ConQuest). Note that the DCI variable at time 0 is problematic; the step parameters all show high (i.e. > 20.0) positive misfit, and the correlation of this variable with most of the other variables is low: .19 with UC and .36 with ET.

Insert Table 8 about here

Figure 6 shows the student proficiencies for the four variables at the four time periods, using the embedded assessment data. It can be seen that the patterns of student achievement change somewhat when the embedded assessment data are included. DCI, UC, and ET show approximately the same pattern when the

Multidimensional modeling

embedded assessment data are included than when only the link tests were analyzed: DCI and UC increase from August to January while ET drops, and UC shows a large increase between January and April, while ET remains essentially flat. In this case, however, DCI shows a slight increase between January and April (the slope is approximately the same as between August and January), rather than remaining flat as in the unidimensional analysis.

CSI, however, shows a quite different pattern. When only the link test data were considered, CSI remained essentially flat throughout the year. However, when the link test data are added, CSI shows a large increase from September to January, and a correspondingly large decrease from January to April. It should be noted, however, that in both cases the CSI estimates are based on a relatively small amount of data; CSI is the variable to which the fewest activities are targeted. Thus, it is probably not possible to read a great deal of meaning into these differing patterns.

Insert Figure 6 about here

Table 8 contains the reliabilities of the person estimates from the four time periods when including all of the embedded assessment activities for which we have data. These reliabilities are not much different than the reliabilities of the corresponding variables and time periods using only the items on the link tests. The largest change in reliabilities (link & embedded minus link only) between these two analyses was .06, the smallest -.05, and the average change was .01. This is surprising, given that when all else is held constant, adding items will improve

Multidimensional modeling

reliability, and the analyses that included the embedded assessments had many more items than those including only the link tests.

Discussion

Multidimensional analysis makes it clear that the pattern of development seen in the unidimensional analysis hides some important characteristics of student growth. Most notably, when looking at the multidimensional analysis, it becomes clear that students do not progress in all of the variables in the same way. The Understanding Concepts variable shows the most growth. This is perhaps not surprising, since as students are being exposed to concepts that they likely have never seen before, they would be expected to change considerably in their understanding of these concepts. However, anecdotal evidence from teachers had led us to expect that students would also show noticeable growth in other variables, and most particularly in the Evidence and Tradeoffs variable. This is considered a unique aspect of the SEPUP IEY curriculum, and in some respects is the most important of the variables. It is not clear at this point how much of the lack of growth we saw has to do with issues such as this being a field test year, when many of the teachers were relatively unfamiliar with the curriculum and the assessment system, or perhaps with some confounding inherent in the way the variable was measured (in the link tests especially, with relatively few tasks per test focused on each variable).

In addition, the use of a multidimensional model improves the reliability with which we can estimate person proficiencies on individual variables, by using

Multidimensional modeling

the collateral information that can be found in correlated variables to improve the precision of the estimates.

Another surprising result of these analyses was that adding the embedded assessment items to the link items did not improve the reliability of the person proficiency estimates. This could indicate one of several things. A positive interpretation would be that, by using multidimensional models such as those discussed in this paper, we have reached a ceiling for reliability--variability inherent in person performance and in teacher scoring may prevent us from achieving higher reliability, even when more items are added. A more troubling interpretation is that student performances on the link tests and on the embedded assessments in fact constitute different variables that are not strongly related. This could be problematic, since the link tests are used to anchor the scales of the variables, to separate student change throughout the year from item difficulty. If in fact these two types of performance do not show a strong relationship, this method may not be a satisfactory way of disentangling changes in person proficiency from changes in item difficulty—and hence, these may remain confounded in the analysis of the embedded tasks.

Limitations of the current research and suggestions for the future

One of the major limitations of the models discussed in this report is the potential for violations of the assumption of local independence. As mentioned before, two or three items may relate to one SEPUP link test question or embedded assessment activity. Teachers scored many of the responses on more than one

Multidimensional modeling

element of a variable or on more than one variable, and each individual score was considered an item. For example, if a response was scored on the two elements of Evidence and Tradeoffs and also on two elements of Communicating Scientific Information, there would be four items associated with that response. This issue has not been addressed in the current set of analyses. In order to adequately address such an issue, a multidimensional version of item bundles (Wilson & Adams, 1995; Rosenbaum, 1988) would be necessary. Item bundles involve combining the responses of items suspected to be dependent for some reason (e.g. they depend upon common stimulus materials, or, as in this case, they are multiple scores for the same response) to create one “mega-item”, with a total number of response categories corresponding to all possible combinations of the various scores. For example, a response scored on one element of ET and one element of CSI would have 25 (or 5 times 5) possible response categories corresponding to every possible ET score matched with every possible CSI score (although not all these categories would be necessary if some combinations were not used by scorers). A scoring matrix would then define how each response category on the mega-item would be scored on the separate variables. Each score associated with a single link test question or assessment activity would be associated with this single mega-item; however, the different score combinations would contribute to (potentially) different dimensions. Such a model would be multidimensional within-item (Adams, Wilson & Wang, 1997). Such an analysis is an important next step. Violation of local independence can lead to a number of difficulties. These include effects on parameter estimation (Chen & Thissen, 1996), accuracy of person

Multidimensional modeling

proficiency estimation (Wilson, 1988), and the related issue of test reliability (Wainer & Thissen, 1996; Sireci, Thissen, & Wainer, 1991). Yen (1993) describes a number of possible effects of such violation, including the reporting of subscale information, an issue clearly of importance here, since we are estimating person proficiencies on different variables, but which are often measured by the same tasks.

Another limitation of the current research is that no analysis was done of variation in individual teacher rating severity. Only the average of all rating effects per variable across all teachers was used at a given time period. This was because only a small number of papers (approximately 10) for each teacher were rescored; this does not provide enough information to accurately estimate rating severities for an individual teacher. This question is being at least partially addressed in a current research project. A series of new link items have been developed, divided into forms, and administered as a pretest in a number of classes that are similar in structure to the ADC, PDC, and comparison classes in the 1994-95 field test study. These link test items were designed to be scored on multiple elements of at least two, and in most cases three, of the SEPUP variables, in order to elicit the maximum information possible from a set of student responses. Each ADC teacher will administer this set of pretests to at least one of their classes, and score the result. The entire set of pretests will then be rescored by at least one, and possibly two, staff members. This will allow us to investigate, among other things, variation in teacher rating severity, at least at a single point in time.

Multidimensional modeling

References

- Adams, R. J., Wilson, M. & Wang, W. C. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement, 21*, 1-23.
- Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46*, 443-459.
- Chen, W. H. & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289.
- Masters, G. N., Adams, R. J., & Wilson, M. (1990) Charting of student progress. In T. Husen & T. N. Postlethwaite (Eds.), *International Encyclopedia of Education: Research and Studies, Supplementary Volume 2*. Oxford: Pergamon Press. (pp. 628-634).
- Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133-161.
- Roberts, L., Wilson, M., & Draney, K. (1997). *The SEPUP Assessment System: An overview*. University of California, Berkeley: BEAR Report Series, SA-97-1.
- Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 53*, 349-359.
- Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237-247.

Multidimensional modeling

- Volodin, N. A. & Adams, R. J. (1995, April). *Identifying and estimating a D-dimensional item response model*. Paper presented at the Eighth International Objective Measurement Workshop, University of California, Berkeley.
- Wainer, H. & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, 15 (1), 22-29.
- Wang, W. C., Wilson, M. & Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. In M. Wilson, G. Engelhard and K. Draney (Eds.), *Objective Measurement: Theory into Practice (Vol. 4)*. Greenwich, CT: Ablex.
- Wilson, M. (1998, July). *Embedding developmental assessment into an instructional curriculum: The case of SEPUP's Issues, Evidence, and You*. International Developmental Assessment Seminar, Royal Melbourne Institute of Technology, Melbourne, Australia.
- Wilson, M. (1988). Detecting and interpreting local item dependence using a family of Rasch models. *Applied Psychological Measurement*, 12, 353-364.
- Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika*, 60, 181-198.
- Wilson, M., & Case, H. (in press). An examination of variation in rater severity over time: A study of rater drift. In M. Wilson & G Engelhard (Eds.) *Objective Measurement: Theory into Practice (Vol. 5)*. Stamford, CT: Ablex.
- Wilson, M. & Draney, K. (1997). *Developing maps for student progress in the SEPUP assessment system*. University of California, Berkeley: BEAR Report Series, SA-97-2.

Multidimensional modeling

Wolf, D., Bixby, J., Glenn, John, III, & Gardner, H. (1991). To use their minds well:

Investigating new forms of student assessment. *Review of Research in Education*, 17, 31-74.

Wright, B. D. & Masters, G, N. (1982). *Rating Scale Analysis*. Chicago, MESA Press.

Wu, M. L., Adams, R. J., & Wilson, M. (1997). *ConQuest: Generalised item response modelling software, Draft Release 2*. Hawthorn, Australia: ACER.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187-213.

Multidimensional modeling

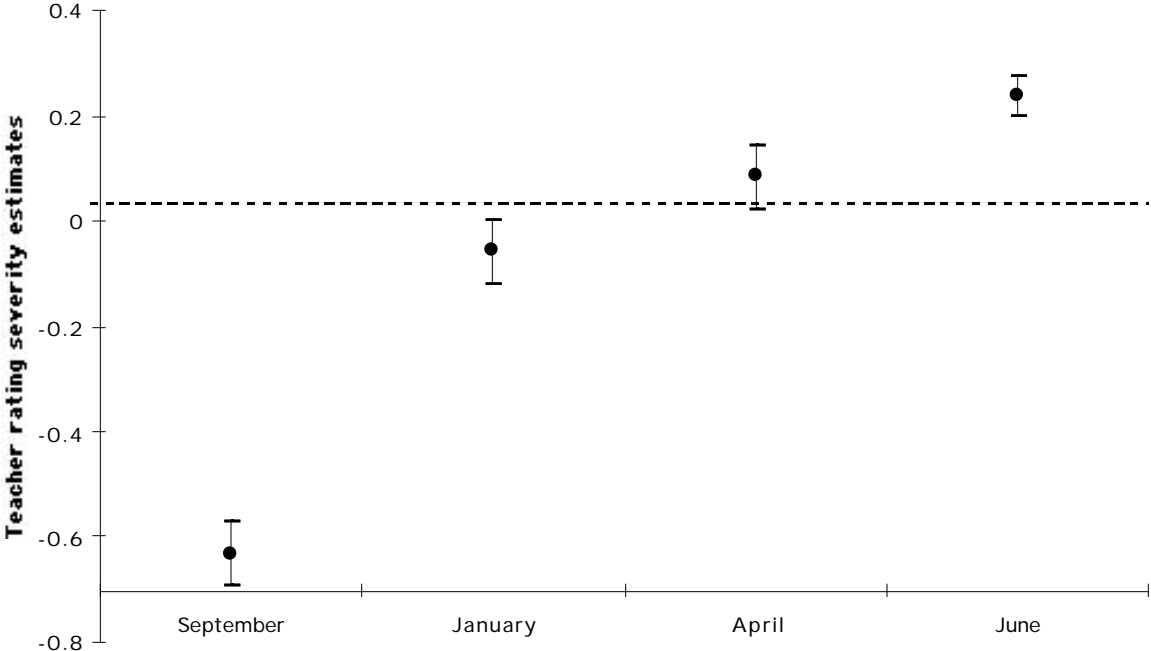


Figure 1: Teacher rating severity estimates for unidimensional model.

Multidimensional modeling

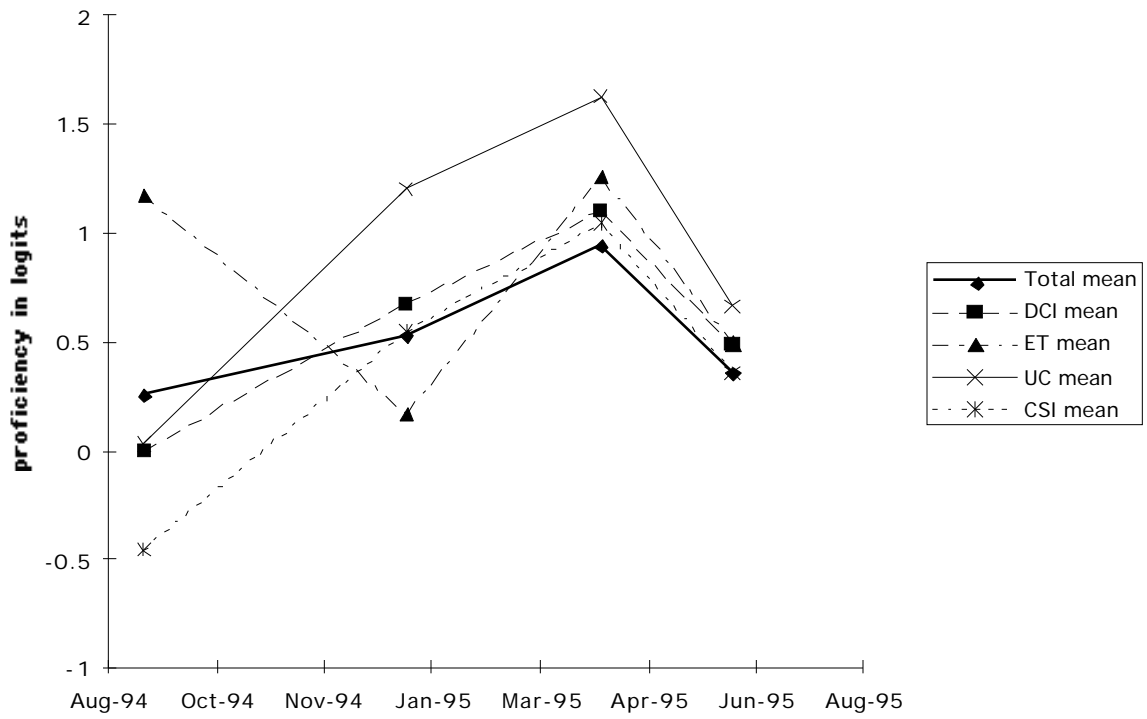


Figure 2: Student means by variable and overall for the unidimensional model.

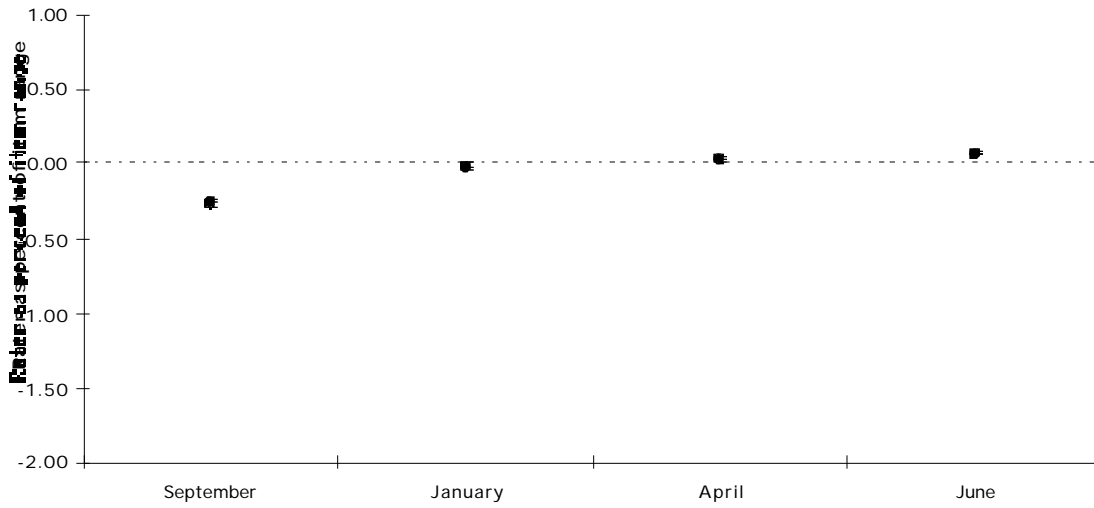
Multidimensional modeling

Logits	DCI	μ	ET	μ	UC	μ	CSI	μ
5.50								
5.25								
5.00					lk2_5			
4.75								
4.50			lk3_1.4					
4.25	psta_3.2				lk2_2			
4.00	psta_3.3		prea_1.2		lk3_1.2			
3.75	lk1_3.3 lk3_3.3				2			
3.50	lk1_3.1 lk1_3.2 lk3_3.1				lk2_1.2			
3.25	lk3_1.1							
3.00	lk2_1.1		lk1_4.2					
2.75			prea_1.1 prea_3.1		prea_4.1 prea_4.2			
2.50	pret_2 lk3_3.2		prea_3.2 lk2_1.4					
2.25	lk1_1.1		pret_4 0					
2.00			lk1_4.1		pret_1			
1.75							prea_5.1	
1.50		2			1 3			
1.25		1		3				
1.00	lk1_5.1 lk1_5.2			1				
0.75		3		2	pret_3		prea_5.1 lk1_4.3	
0.50	prea_2					2		2
0.25				2	0	3	prea_5.2	
0.00							pret_5.2 3	
-0.25		1 3		1		0 1	lk1_4.4 0 2	0 1
-0.50							1	
-0.75		2						3
-1.00								
-1.25				3				
-1.50								
-1.75								
-2.00								
-2.25								
-2.50				0				

Figure 3: Map of multidimensional model

Multidimensional modeling

Unidimensional model:



Multidimensional model:

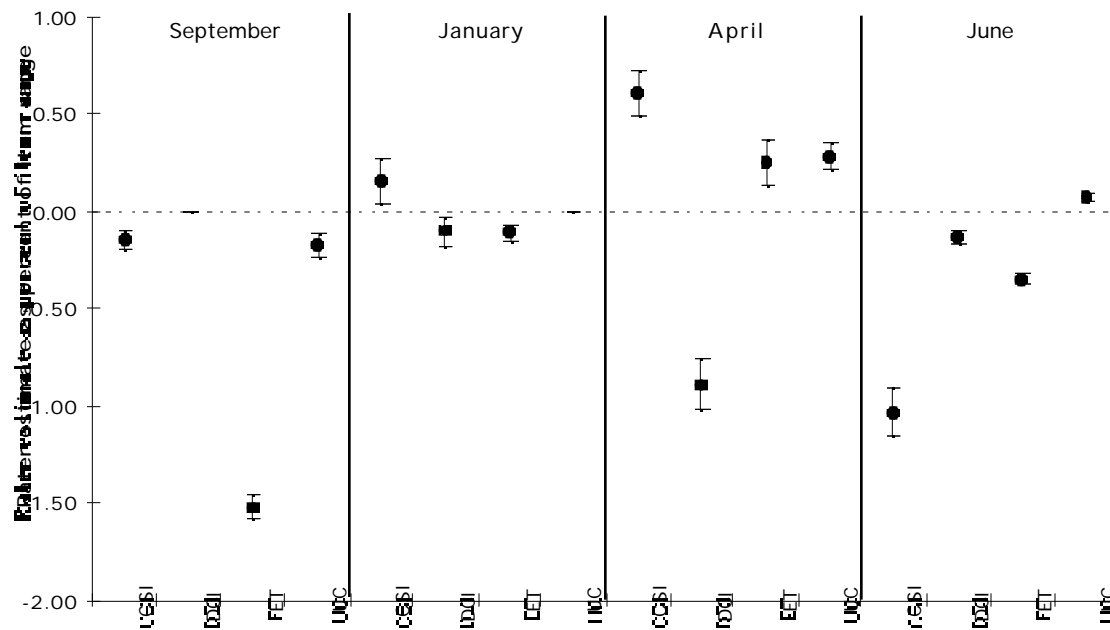


Figure 4: Teacher rating severity estimates expressed as percentages of item range for unidimensional and multidimensional models

Multidimensional modeling

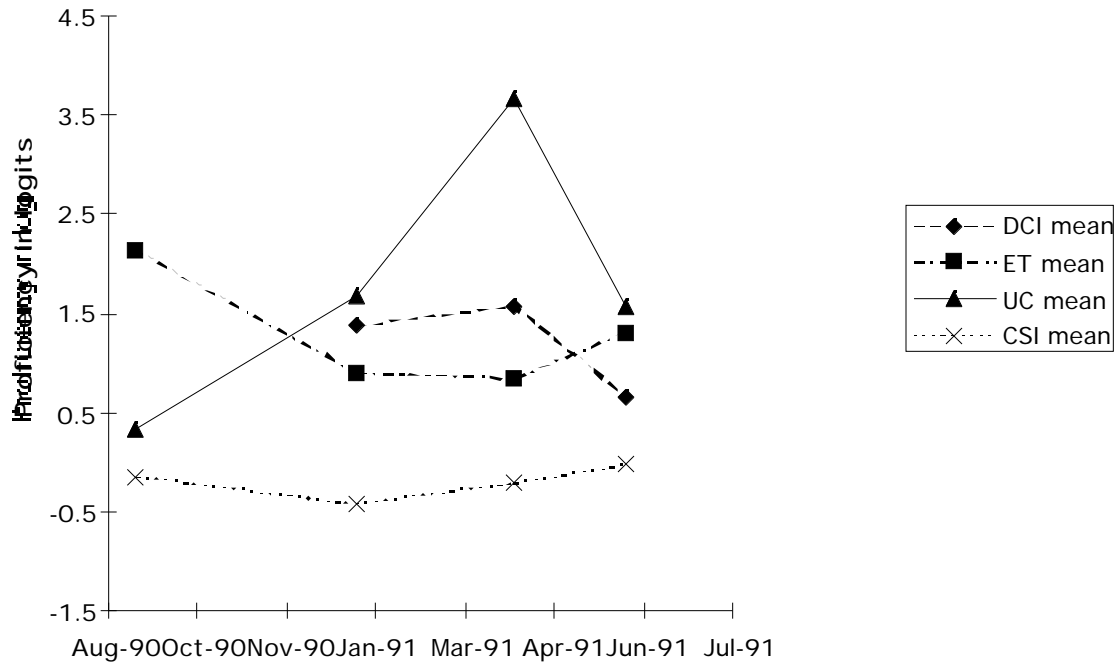


Figure 5: Proficiency estimates for multidimensional model, link tests only

Multidimensional modeling

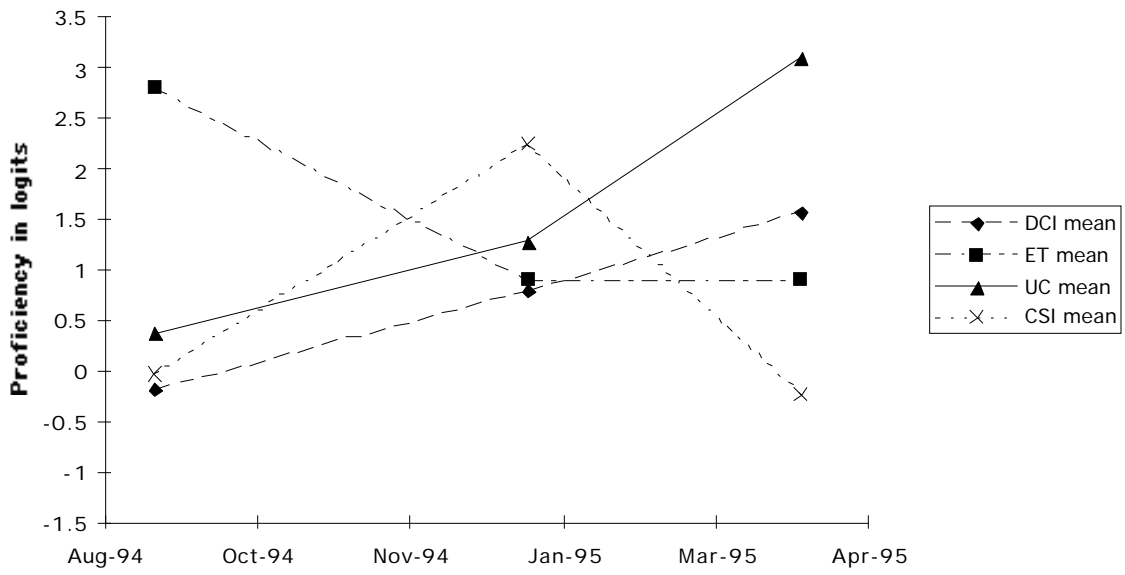


Figure 6: Proficiency estimates for multidimensional model, link tests and embedded assessments

Multidimensional modeling

Table 1: Unidimensional parameter estimates and item information for 1995 pretest

Par #	Estimate	SE	Unweighted Fit		Weighted Fit		SEPU P Var	94-95 Location(s)	Item 1995 form/item
			Mean Square	t	Mean Square	t			
1	0.31	0.07	1.18	1.67	1.07	0.99	CI	Link3-3.1	A-1
2	0.70	0.08	1.14	1.32	1.08	1.01	CI	Link3-3.2	A-1
3	0.54	0.08	1.09	0.87	1.02	0.27	CI	Link3-3.3	A-1
4	0.10	0.08	0.86	###	0.80	###	ET	not used in 94-95	A-2
5	0.23	0.08	0.96	###	0.91	###	ET	PreTest 4	A-2
6	-0.31	0.07	0.88	###	0.79	###	CM	not used in 94-95	A-2
7	-1.49	0.08	1.12	1.14	1.11	1.36	CM	not used in 94-95	A-2
8	0.51	0.06	1.06	0.71	1.07	1.18	CI	Activity 18	A-3; B-4
9	0.48	0.06	1.20	2.13	1.26	3.75	UC	Activity 18	A-3; B-4
10	-0.11	0.06	0.95	###	0.93	###	UC	(PreAnch-3.2); (Link1-1.3)	A-4; B-5
11	0.65	0.06	1.02	0.25	1.01	0.23	ET	PreAnch-3.1; Link1-1.2	A-4; B-5
12	1.70	0.09	1.35	3.01	1.23	2.48	UC	Link2-2; PostAnch-2	A-5
13	-1.18	0.06	0.97	###	0.95	###	ET	Link1-5.1; 4.1	Link3- B-1; C-4
14	-1.25	0.06	1.11	1.29	1.09	1.50	ET	Link1-5.2; 4.2	Link3- B-1; C-4
15	0.10	0.06	1.00	0.08	0.96	###	CM	PreAnch-5.1	B-2; C-5
16	-1.53	0.06	1.40	4.10	1.43	5.92	CM	PreAnch-5.2	B-2; C-5
17	-0.95	0.07	1.25	2.23	1.26	3.21	UC	Activity 16	B-3

Multidimensional modeling

18	-0.22	0.0	0.95	###	1.00	###	UC	PreAnch-4.1; Link-1 2.1	C-1; D-2
19	-0.09	0.0	0.91	###	0.93	###	UC	(PreAnch-4.2); (Link-2 2.2)	C-1; D-2
20	-0.25	0.0	0.92	###	0.93	###	CI	Link2-1; PostAnch-1	C-2; D-1
21	0.12	0.0	1.16	1.8	1.17	2.1	ET	Link2-1; PostAnch-1	C-2; D-1
22	0.64	0.0	1.07	0.6	1.14	1.7	CI	Activity 18	C-3
23	0.58	0.0	0.98	###	1.01	0.1	CI	Activity 18	C-3
24	1.37	0.0	1.18	1.6	1.19	1.6	ET	Activity 12	D-3
25	1.52	0.0	1.27	2.3	1.27	2.3	ET	Activity 12	D-3
26	0.15	0.0	1.22	2.0	1.24	2.2	CI	Link1-3.1; Link2- 3.1	D-4
27	0.19	0.0	1.10	0.9	1.12	1.1	CI	Link1-3.2; Link2- 3.2	D-4
28	0.40	0.0	0.94	###	0.95	###	CI	Link1-3.3; Link2- 3.3	D-4
29	-1.00	0.0	0.86	###	0.89	###	CM	PostAnch-5, pre/posttest 5	D-5
	-1.89						CM	PostAnch-5, pre/posttest 5	D-5
30	-1.72	0.0	1.50	4.5	1.56	8.6	CI	Step 1	
31	-1.44	0.0	1.46	4.1	1.63	8.6	CI	Step 2	
32	0.53	0.1	0.86	###	1.55	4.4	CI	Step 3	
33	-2.69	0.0	1.17	1.8	1.25	4.0	ET	Step 1	
34	-1.35	0.0	1.28	2.9	1.31	5.5	ET	Step 2	
35	0.73	0.0	1.79	7.4	1.59	6.7	ET	Step 3	
36	-2.10	0.0	1.66	6.0	1.60	7.8	CM	Step 1	
37	-1.70	0.0	1.24	2.4	1.24	4.1	CM	Step 2	
38	-0.12	0.0	1.42	4.0	1.17	3.3	CM	Step 3	
39	-2.15	0.0	1.24	2.7	1.30	5.4	UC	Step 1	

Multidimensional modeling

		6		1		5		
40	-0.68	0.0	1.44	4.7	1.37	6.1	UC	Step 2
		6		6		4		
41	0.34	0.1	1.35	3.8	1.21	2.2	UC	Step 3
		0		5		8		

Mean and SE of Person Distribution (720 cases)

	-1.64	0		
Variance and SE of Person Distribution				
	1.14	0.1		
Deviance	165		#	43
	31.		Paramete	
	90		rs	

Multidimensional modeling

Table 2: Four dimensional parameter estimates and item information for the 1995 pretest

P ar #	value	se	unw- mnsq	unw -t	w- mns q	w-t	Varia ble	95 pretest	Operational location
1	-0.11	0.08	0.81	-	0.87	-	DCI	A-1i1	Link3-3
				1.85		1.73			
2	0.43	0.08	0.83	-	0.90	-	DCI	A-1i2	Link3-3
				1.65		1.25			
3	0.20	0.08	0.80	-	0.81	-	DCI	A-1i3	Link3-3
				1.97		2.58			
4	0.11	0.07	1.17	1.87	1.23	3.32	DCI	A-3; B-4i1	Activity 18
5	-0.84	0.06	1.39	4.05	1.38	4.31	DCI	C-2; D-1i1	Link2-1; PostAnchor-1
6	0.35	0.08	0.82	-	0.90	-	DCI	C-3i1	Activity 18
				1.74		1.28			
7	0.27	0.08	0.76	-	0.83	-	DCI	C-3i2	Activity 18
				2.49		2.31			
8	-0.27	0.08	0.35	-	0.34	-	DCI	D-4i1	Link1-3; Link2-3
				8.22		8.04			
9	0.07	0.08	0.54	-	0.47	-	DCI	D-4i3	Link1-3; Link2-3
				5.26		5.94			
	-0.22						DCI	D-4i2	Link1-3; Link2-3
10	0.81	0.06	0.70	-	0.69	-	CSI	A-2i3	PreTest A-4
				3.78		5.86			
11	-0.64	0.06	0.81	-	0.81	-	CSI	A-2i4	PreTest A-4
				2.26		3.43			
12	-0.61	0.05	0.91	-	0.92	-	CSI	B-2; C-5i2	PreAnchor-5
				1.21		1.06			
13	0.05	0.06	0.88	-	0.87	-	CSI	D-5i1	PostAnchor-5
				1.41		2.34			
14	-1.04	0.06	0.84	-	0.84	-	CSI	D-5i2	PostAnchor-5
				1.86		2.78			
	1.42						CSI	B-2; C-5i1	PreAnchor-5
15	-0.12	0.08	0.30	-	0.31	-	ET	A-2i1	PreTest A-4
				9.43		9.30			
16	0.60	0.07	1.13	1.41	1.18	2.15	ET	A-4; B-5i2	PreAnchor-3; Link1-1
17	-1.89	0.07	0.81	-	0.81	-	ET	B-1; C-4i1	Link1-5; Link3-4
				2.34		3.36			
18	-1.98	0.07	0.90	-	0.91	-	ET	B-1; C-4i2	Link1-5; Link3-4
				1.17		1.56			

Multidimensional modeling

19	-0.11	0.07	1.24	2.64	1.18	2.90	ET	C-2; D-Link2-1; PostAnchor-1 1i2
20	1.63	0.09	0.74	-	0.81	-	ET	D-3i1 Activity 12
				2.70		2.60		
21	1.82	0.09	0.82	-	0.90	-	ET	D-3i2 Activity 12
				1.82		1.27		
	0.05						ET	A-2i2 PreTest A-4
22	0.41	0.06	0.89	-	0.92	-	UC	A-3; B-Activity 18
				1.47		1.52		4i2
23	-0.32	0.06	0.86	-	0.86	-	UC	A-4; B-PreAnchor-3; Link1-1
				1.93		2.66		5i1
24	1.96	0.07	0.93	-	0.82	-	UC	A-5 Link2-2; PostAnchor-2
				0.85		2.83		
25	-1.44	0.06	0.90	-	0.92	-	UC	B-3 Activity 16
				1.17		1.37		
26	-0.39	0.06	0.12	###	0.13	###	UC	C-1; D-PreAnchor-4; Link-1 2.2
				#		#		2i1
	-0.23						UC	C-1; D-PreAnchor-4; Link-1 2.2
								2i2
27	-2.59	0.06	1.26	2.48	1.24	3.62	DCI	
28	-1.74	0.06	1.19	1.86	1.39	5.05	DCI	
29	0.79	0.11	1.99	7.98	1.43	3.36	DCI	
30	-2.63	0.09	1.15	1.59	1.33	4.43	CSI	
31	-1.88	0.08	0.97	-	1.04	0.72	CSI	
				0.25				
32	0.06	0.07	1.39	3.77	1.00	0.07	CSI	
33	-3.64	0.07	1.19	2.10	1.18	2.76	ET	
34	-1.65	0.06	1.29	3.14	1.17	2.88	ET	
35	1.05	0.09	1.31	3.26	1.39	4.39	ET	
36	-2.78	0.06	1.06	0.77	1.13	2.26	UC	
37	-0.87	0.07	1.37	4.04	1.25	3.97	UC	
38	0.52	0.10	0.80	-	1.09	0.98	UC	
				2.52				

52

parameters

Deviance 16089

Correlations

DCI CSI ET

C 0.51

SI

Means and standard deviations of person distribution (N=720)

-2.71 1.72

-0.62 1.44

Multidimensional modeling

E	0.72	0.64		-2.36	1.73
T					
U	0.75	0.55	0.69	-2.20	1.52
C					

Multidimensional modeling

Table 3a: Unidimensional parameter estimators and item information for 1994-95 field test data

Parameter	Va r	value	se	Time 0 Fit				Time 1 Fit				Time 2 Fit						
				u-mnsq	unw-t	w-mnsq	w-t	se	u-mnsq	unw-t	w-mnsq	w-t	se	u-mnsq	unw-t	w-mnsq		
pret	1	UC	1.03	0.00	1.04	0.35	0.98	-0.22										
pret	2	DC	1.44	0.00	0.88	-0.44	0.91	-0.38										
		I																
pret	3	UC	0.23	0.00	2.16	6.77	2.44	10.63										
pret	4	ET	1.76	0.00	1.86	5.28	1.83	6.87										
pret	5.1	CSI	0.80	0.00	1.61	3.94	1.51	4.38										
pret	5.2	CSI	-0.08	0.00	1.67	4.25	1.70	5.17										
pra	1.1	ET	0.30	0.10	0.93	-0.65	0.94	-0.59										
pra	1.2	ET	1.46	0.10	1.20	1.94	1.20	2.02										
pra	2	DC	0.79	0.09	1.26	2.42	1.24	2.27										
		I																
pra	3.1	ET	2.31	0.00	3.88	13.88	3.60	16.16	0.00	1.75	5.05	1.69	5.72					
pra	3.2	UC	1.55	0.00	0.98	-0.06	0.95	-0.21	0.00	0.99	0.01	0.90	-0.55					
pra	4.1	UC	1.77	0.00	1.17	1.33	1.22	2.08	0.00	1.15	1.19	1.20	1.89					
pra	4.2	UC	1.84	0.00	1.04	0.28	1.03	0.21	0.00	1.24	1.83	1.28	2.54					
pra	5.1	CSI	1.74	0.00	0.91	-0.91	0.91	-0.98										
pra	5.2	CSI	0.06	0.00	1.83	6.64	1.93	6.73										
lk1	1.1	DC	1.02						0.10	1.47	4.11	1.45	3.95					
		I																
lk1	3.1	DC	2.11						0.00	1.53	3.72	1.45	3.99	0.00	1.55	3.83	1.53	
		I																
lk1	3.2	DC	2.15						0.00	1.80	5.27	1.50	4.34	0.00	1.37	2.70	1.34	
		I																
lk1	3.3	DC	2.37						0.00	1.69	4.69	1.56	4.84	0.00	1.34	2.53	1.45	
		I																
lk1	4.1	ET	1.76						0.09	1.07	0.57	1.10	0.97	0.00	1.34	2.54	1.35	
lk1	4.2	ET	2.77						0.09	0.95	-0.34	0.95	-0.46	0.00	1.11	0.86	1.05	
lk1	4.3	CSI	1.40						0.08	0.68	-2.87	0.64	-4.05	0.00	1.02	0.21	0.94	
lk1	4.4	CSI	0.27						0.09	1.08	0.70	1.10	0.94	0.00	1.45	3.22	1.45	
lk1	5.1	ET	0.49						0.00	1.06	0.64	1.10	0.98					
lk1	5.2	ET	0.42						0.00	2.56	11.11	2.61	11.52					
lk2	1.1	DC	1.66											0.00	1.64	4.86	1.63	
		I																
lk2	1.2	UC	0.88											0.09	1.12	1.21	1.16	
lk2	1.4	ET	1.90											0.00	2.39	8.52	2.46	
lk2	2	UC	3.58											0.00	7.24	23.59	7.52	
lk2	5	UC	2.23											0.10	1.44	3.87	1.41	
lk3	1.1	DC	1.69															
		I																
lk3	1.2	UC	1.62															
lk3	1.4	ET	3.43															
lk3	3.1	DC	2.01															
		I																
lk3	3.2	DC	2.43															
		I																
lk3	3.3	DC	2.26															
		I																
psta	3.2	DC	3.10															
		I																
psta	3.3	DC	2.57															
		I																

Note: SE of 0 indicates that the parameter has been anchored to its 1995 pretest value

Multidimensional modeling

Table 3a: Unidimensional parameter estimators and item information for 1994-95 field test data

w-t	se	Time 3 Fit			
		u-mnsq	unw-t	w-mnsq	w-t
0.00	1.18	1.16	1.13	1.03	
0.00	0.96	-0.09	0.89	-0.38	
0.00	2.77	7.50	2.88	9.98	
0.00	1.18	1.16	1.14	1.17	
0.00	1.89	4.88	1.82	5.87	
0.00	2.07	5.53	2.05	6.74	
0.00	0.93	-0.54	0.96	-0.33	
0.00	1.15	1.27	1.15	1.39	
0.00	1.73	5.25	1.70	5.19	
4.59					
3.08					
3.94					
3.14					
0.49					
-0.54					
3.53					
0.00	1.09	0.70	1.11	0.89	
0.00	1.48	3.38	1.46	3.28	
5.31	0.00	2.09	6.74	1.87	6.16
1.69	0.00	1.45	3.51	1.45	3.81
10.26	0.00	2.63	8.79	2.77	10.56
25.80	0.00	9.24	25.25	10.20	26.28
3.79	0.00	1.78	5.05	1.75	4.99
0.08	2.11	6.69	2.06	6.81	
0.12	1.39	2.81	1.40	2.96	
0.16	1.76	4.49	1.85	5.00	
0.00	1.75	4.87	1.69	5.19	
0.00	4.64	15.76	3.94	15.33	
0.00	1.76	4.94	1.82	5.98	
0.13	1.66	4.83	1.74	4.80	
0.11	1.36	2.84	1.38	3.15	

Note: SE of 0 indicates that the parameter has been anchored to its 1995 pretest value

Multidimensional modeling

Table 3b: Unidimensional step, rater, and population parameters for the 1994-95 field test data

Parameter	location	Time 0 Fit					Time 1 Fit					Time 2 Fit			
		se	u- msg	u-t	w- msg	w-t	se	u- msg	u-t	w- msg	w-t	se	u- msg	u-t	
DCI step1	DCI	-2.02	0.0	1.63	4.13	1.26	1.45	0.0	1.80	5.48	1.68	5.28	0.0	1.89	5.94
DCI step2	DCI	-1.62	0.0	1.10	0.76	1.08	0.83	0.0	1.84	5.70	1.74	5.88	0.0	1.54	3.94
DCI step3	DCI	0.61	0.0	0.97	-0.15	1.02	0.22	0.0	4.30	####	2.21	6.70	0.0	2.51	9.04
ET step1	ET	-2.91	0.0	1.06	0.58	0.98	-0.14	0.0	2.41	8.84	2.28	8.63	0.0	2.40	8.34
ET step2	ET	-1.27	0.0	1.65	4.71	1.64	5.73	0.0	1.40	3.13	1.49	4.31	0.0	1.68	4.64
ET step3	ET	0.75	0.0	4.81	####	3.19	####	0.0	1.76	5.41	1.42	3.33	0.0	1.48	3.47
UC step1	UC	-2.59	0.0	1.48	3.12	1.60	4.19	0.0	1.26	1.86	1.18	1.43	0.0	1.46	3.75
UC step2	UC	-0.73	0.0	1.09	0.67	1.14	1.32	0.0	1.62	4.08	1.52	4.16	0.0	2.59	####
UC step3	UC	0.40	0.0	1.25	1.72	1.12	1.04	0.0	2.43	7.95	2.08	5.51	0.0	2.59	####
CSI step1	CSI	-2.21	0.0	1.76	5.27	1.87	5.28	0.0	0.81	-1.64	1.08	0.58	0.0	1.52	3.63
CSI step2	CSI	-1.56	0.0	2.58	9.35	2.58	####	0.0	0.92	-0.62	0.97	-0.28	0.0	1.76	5.03
CSI step3	CSI	0.02	0.0	1.20	1.64	1.30	2.98	0.0	1.04	0.33	1.08	0.86	0.0	0.87	-1.07
rater 0		-0.63	0.0	1.00	0.06	0.98	-0.16								
rater 1		-0.06						0.0	1.07	0.71	1.05	0.57			
rater 2		0.09						3					0.0	1.35	3.19
rater 3		0.24											3		
	Mea	SD													
	n														
Sep-94	###	0.82													
Jan-95	0	1.12													
Apr-95	0	1.06													
Jun-95	0	0.94													
	7														
Deviance	474	#				48									
	78.	parameters													
	60														

Note: SE of 0 indicates that the parameter has been anchored to its 1995 pretest value

Multidimensional modeling

Table 3b: Unidimensional step, rater , and population parameters for the 1994-95 field test data

Time 3					
Fit					
w- msq	w-t	se	u- msq	u-t	w- msq
1.60	4.49	0.0	1.52	3.55	1.52
		0			
1.58	4.70	0.0	1.63	4.20	1.35
		0			
2.07	6.31	0.0	5.45	####	3.69
		0			
2.03	6.57	0.0	2.43	7.91	1.97
		0			
1.82	6.39	0.0	1.02	0.23	0.93
		0			
1.03	0.24	0.0	1.18	1.32	1.32
		0			
1.56	4.96	0.0	1.64	3.95	1.62
		0			
2.46	####	0.0	2.36	7.28	2.26
		0			
2.58	9.93	0.0	2.17	6.47	2.09
		0			
1.50	2.90	0.0	2.38	6.84	2.50
		0			
1.73	5.51	0.0	3.25	9.78	2.84
		0			
0.91	-0.96	0.0	1.36	2.25	1.53
		0			
1.38	3.44				
		0.0	1.16	1.27	1.14
		2			

Note: SE of 0 indicates that the parameter has been anchored to its 1995 pretest value

Multidimensional modeling

Table 4: reliabilities of person estimates for the unidimensional link test analysis

Time	Total			DCI			ET		
	pop var	var EAP	reliabilit	pop var	var EAP	reliabilit	pop var	var EAP	reliabilit
			y			y			y
0	0.89	0.70	.79	na	na	na	1.16	0.64	.55
1	1.04	0.92	.88	1.29	0.96	.74	1.63	1.30	.80
2	1.31	1.18	.90	1.44	1.15	.80	1.63	1.30	.80
3	1.14	1.04	.91	1.15	0.91	.79	1.15	0.84	.73

	UC			CSI		
	pop var	var EAP	reliabilit	pop var	var EAP	reliabilit
			y			y
	0.84	0.45	.54	1.09	0.64	.59
	2.00	1.36	.68	1.62	1.02	.63
	2.00	1.67	.84	1.31	0.82	.63
	1.44	1.09	.76	1.31	1.05	.80

Multidimensional modeling

Table 5a: Multidimensional item difficulty information for the 1994-95 field test data

Paramete r	location	Time 0 Fit					Time 1 Fit					Time 2 Fit				
		se	u-msq	u-t	w- msq	w-t	se	u-msq	u-t	w- msq	w-t	se	u-msq	u-t	w- msq	w-t
pret	1 UC	1.92	0.00	1.37	2.61	1.37	3.39									
pret	2 DC	2.59	0.00	1.30	1.21	1.30	1.22									
	I															
pret	3 UC	0.65	0.00	2.49	8.25	2.68	11.31									
pret	4 ET	2.28	0.00	1.92	5.60	1.78	5.98									
pre	5.1 CSI	0.87	0.00	0.70	-2.46	0.70	-3.71									
pre	5.2 CSI	0.08	0.00	1.11	0.84	1.07	0.65									
pra	1.1 ET	2.63	0.11	0.97	-0.24	0.96	-0.39									
pra	1.2 ET	3.96	0.10	1.24	2.27	1.22	2.20									
pra	2 DC	0.61	0.12	1.21	1.98	1.16	1.48									
	I															
pra	3.1 ET	2.86	0.00	1.70	4.72	1.59	4.78	0.00	1.71	4.80	1.46	3.94				
pra	3.2 UC	2.55	0.00	1.36	1.84	1.28	1.57	0.00	1.10	0.60	1.09	0.54				
pra	4.1 UC	2.76	0.00	1.24	1.86	1.27	2.46	0.00	0.89	-0.87	0.88	-1.12				
pra	4.2 UC	2.84	0.00	0.85	-0.82	0.89	-0.60	0.00	0.76	-2.12	0.75	-2.61				
pra	5.1 CSI	1.71	0.00	0.79	-2.18	0.76	-3.04									
pra	5.2 CSI	0.33	0.00	1.13	1.29	1.16	1.55									
lk1	1.1 DC	2.13						0.12	2.61	11.17	2.33	9.06				
	I															
lk1	3.1 DC	3.50						0.00	1.72	4.85	1.37	3.20	0.00	1.94	6.00	1.69 5.61
	I															
lk1	3.2 DC	3.55						0.00	1.51	3.58	1.28	2.46	0.00	1.49	3.47	1.34 3.02
	I															
lk1	3.3 DC	3.78						0.00	1.35	2.59	1.26	2.32	0.00	1.75	4.99	1.57 4.74
	I															
lk1	4.1 ET	2.05						0.09	1.08	0.68	1.09	0.85	0.00	1.12	0.99	1.07 0.67
lk1	4.2 ET	3.10						0.09	0.75	-2.21	0.79	-2.16	0.00	0.93	-0.54	0.92 -0.76
lk1	4.3 CSI	0.64						0.08	0.58	-3.94	0.58	-4.81	0.00	0.60	-3.77	0.57 -5.13
lk1	4.4 CSI	-0.37						0.09	0.88	-0.98	0.95	-0.42	0.00	1.03	0.29	1.02 0.17
lk1	5.1 ET	1.07						0.00	1.36	3.27	1.43	3.78				
lk1	5.2 ET	0.98						0.00	2.12	8.56	2.18	8.79				
lk2	1.1 DC	3.02											0.00	2.18	8.01	2.12 8.25
	I															
lk2	1.2 UC	3.59											0.11	1.49	4.29	1.58 5.01
lk2	1.4 ET	2.53											0.00	1.91	6.08	1.90 7.08
lk2	2 UC	4.33											0.00	1.26	2.03	1.30 2.74
lk2	5 UC	5.32											0.11	1.51	4.41	1.53 4.83
lk3	1.1 DC	3.14														
	I															
lk3	1.2 UC	3.96														
lk3	1.4 ET	4.40														
lk3	3.1 DC	3.40														
	I															
lk3	3.2 DC	2.56														
	I															
lk3	3.3 DC	3.66														
	I															
psta	3.2 DC	4.31														
	I															
psta	3.3 DC	3.89														
	I															

Note: SE of 0 indicates that the parameter has been anchored to its 1995 pretest value

Multidimensional modeling

Table 5a: Multidimensional item difficulty information for the 1994-95 field test data

Time 3 Fit				
se	u-msq	u-t	w- msq	w-t
0.00	2.48	6.96	2.37	7.89
0.00	0.86	-0.50	0.74	-1.37
0.00	4.63	12.34	4.72	14.73
0.00	1.09	0.64	1.09	0.72
0.00	0.98	-0.07	0.99	-0.05
0.00	1.02	0.17	1.04	0.37
0.00	1.79	5.62	1.81	6.19
0.00	2.30	8.35	2.26	9.23
0.09	0.77	-2.12	0.79	-2.38
0.00	1.34	2.49	1.36	2.43
0.00	2.29	7.58	2.25	6.95
0.00	1.12	1.00	1.13	1.16
0.00	1.55	4.18	1.55	4.36
0.00	2.98	10.12	2.76	10.80
0.00	1.87	5.37	2.05	6.94
0.00	3.70	12.90	3.05	10.68
0.07	1.50	3.46	1.43	3.28
0.14	1.68	4.50	1.70	4.85
0.15	1.51	3.18	1.64	4.03
0.00	1.33	2.39	1.28	2.40
0.10	1.53	3.63	1.66	5.09
0.00	1.41	2.91	1.44	3.37
0.12	1.28	2.26	1.36	2.32
0.10	0.97	-0.24	0.98	-0.15

Multidimensional modeling

Table 5b: Multidimensional step, rater, and population parameters for the 1994-95 field test data.

Parameter	location	se	Time 0 Fit				Time 1 Fit				Time 2 Fit						
			u- msq	u-t	w- msq	w-t	se	u- msq	u-t	w- msq	w-t	se	u- msq	u-t	w- msq	w-t	
DCI step 1	DCI	-3.14	0.00	2.65	8.87	1.14	0.92	0.00	2.03	6.73	1.10	0.83	0.00	2.81	10.36	1.28	2.13
DCI step 2	DCI	-2.83	0.00	1.20	1.48	1.16	1.37	0.00	1.25	2.00	0.99	-0.03	0.00	1.96	6.35	1.21	1.74
DCI step 3	DCI	1.25	0.00	0.96	-0.29	1.10	0.88	0.00	5.99	20.51	2.63	6.91	0.00	19.86	42.87	4.25	11.10
ET step 1	ET	-2.96	0.00	1.96	6.53	1.49	3.29	0.00	1.86	5.98	1.40	3.18	0.00	1.67	4.60	1.13	1.10
ET step 2	ET	-1.95	0.00	1.36	2.84	1.31	2.79	0.00	1.18	1.52	1.21	1.93	0.00	1.24	1.89	1.25	2.20
ET step 3	ET	0.80	0.00	1.28	2.25	1.08	0.77	0.00	2.83	10.74	1.61	4.18	0.00	2.32	7.97	1.29	1.66
UC step 1	UC	-3.22	0.00	1.37	2.49	1.17	1.36	0.00	1.57	3.77	1.14	1.06	0.00	1.99	7.13	1.49	3.35
UC step 2	UC	-1.57	0.00	0.98	-0.10	1.06	0.61	0.00	1.46	3.17	1.47	3.48	0.00	1.03	0.34	1.03	0.37
UC step 3	UC	0.74	0.00	1.96	5.58	1.51	3.49	0.00	0.79	-1.67	1.06	0.45	0.00	1.69	5.31	1.04	0.36
CSI step 1	CSI	-1.55	0.00	1.18	1.45	1.16	1.49	0.00	0.67	-3.02	0.97	-0.21	0.00	0.90	-0.77	1.04	0.39
CSI step 2	CSI	-1.74	0.00	1.31	2.43	1.27	2.60	0.00	0.94	-0.47	1.04	0.44	0.00	1.22	1.70	1.19	1.80
CSI step 3	CSI	-0.18	0.00	1.01	0.13	1.02	0.27	0.00	0.89	-0.86	0.94	-0.57	0.00	0.82	-1.49	0.86	-1.43
rater 0	DCI		na	na	na	na	na										
rater 0	ET	-2.55	0.05	1.40	3.58	1.40	3.59										
rater 0	UC	-0.37	0.06	1.20	1.89	1.19	1.86										
rater 0	CSI	-0.24	0.04	1.00	-0.01	0.95	-0.46										
rater 1	DCI	-0.17						0.06	0.94	-0.60	0.87	-1.26					
rater 1	ET	-0.26						0.05	0.98	-0.18	1.00	0.08					
rater 1	UC	-0.32						0.08	1.18	1.78	1.15	1.50					
rater 1	CSI	0.16						0.06	0.82	-1.94	0.82	-1.90					
rater 2	DCI	-0.68											0.05	1.55	4.70	1.51	4.30
rater 2	ET	0.27											0.06	1.08	0.78	1.08	0.80
rater 2	UC	0.49											0.06	1.03	0.31	1.00	0.02
rater 2	CSI	0.62											0.06	1.12	1.21	1.15	1.44
rater 3	DCI	-0.23															
rater 3	ET	-1.17															
rater 3	UC	0.35															
rater 3	CSI	-0.82															
Means and standard deviations																	
			DCI	ET	UC		CSI										
Sep-94	-	-	2.14	1.22	0.31	1.08	-0.15	0.71	769								
Jan-95	1.37	1.67	0.9	1.12	1.67	1.71	-0.43	1.12	452								
Apr-95	1.57	1.74	0.84	1.37	3.66	1.17	-0.22	1.14	442								
Jun-95	0.65	1.51	1.3	1.02	1.58	1.72	-0.03	1.02	467								
Correlations																	
	Sep-94					Jan-95				Apr-95				Jun-95			
	DCI	ET	UC	CSI	DCI	ET	UC	CSI	DCI	ET	UC	CSI	DCI	ET			
DCI	1.00	.37	.56	.53	1.00	.61	.59	.65	1.00	.65	.72	.86	1.00	.72			
ET	.37	1.00	.55	.83	.61	1.00	.79	.93	.65	1.00	.71	.82	.72	1.00			
UC	.56	.55	1.00	.72	.59	.79	1.00	.79	.72	.71	1.00	.82	.77	.63			
CSI	.53	.83	0.72	1.00	.65	.93	0.79	1.00	.86	.82	0.82	1.00	.63	.84			
deviance	3958	#				62											
	4.10	parameters															

Note: SE of 0 indicates that the parameter has been anchored to its 1995 pretest value

Multidimensional modeling

Table 5b: Multidimensional step, rater, and population parameters for the 1994-95 field test data.

Time 3 Fit				
se	u- msg	u-t	w- msg	w-t
0.00	1.32	2.30	1.17	1.44
0.00	1.06	0.48	1.02	0.20
0.08	1.05	0.42	1.08	0.64
0.00	1.35	2.45	1.13	1.10
0.00	1.20	1.49	1.01	0.12
0.00	1.65	4.19	1.61	3.94
0.00	1.16	1.14	1.19	1.54
0.00	1.08	0.64	1.02	0.20
0.00	1.06	0.47	1.13	1.04
0.00	1.28	1.79	1.45	3.06
0.00	1.62	3.57	1.56	4.31
0.00	1.14	0.93	1.36	2.80

0.03	1.10	0.81	1.06	0.57
0.05	1.05	0.41	1.09	0.87
0.06	1.09	0.73	1.01	0.15
0.05	1.18	1.38	1.07	0.64

UC	CSI
.77	.63
.63	.84
1.00	.62
0.62	1.00

Table 6: Reliabilities of person estimates for the multidimensional link test analysis

Multidimensional modeling

			ET			UC	
var	reliabi	pop	var	reliabi	pop	var	reliabi
EAP	lity	var	EAP	lity	var	EAP	lity
na	na	2.05	1.49	.73	1.69	1.17	.69
2.77	.83	1.51	1.26	.83	3.57	2.92	.82
3.01	.90	2.26	1.87	.83	1.60	1.36	.85
2.29	.85	1.28	1.03	.80	3.46	2.95	.85

CSI		
pop	var	reliabi
var	EAP	lity
0.77	0.50	.65
1.51	1.26	.83
1.65	1.29	.78
1.28	1.03	.80

Multidimensional modeling

Table 7a: Multidimensional item difficulty information for the 1994-95 field test activity data

Time	Activity	Variabl e	Estimate	se	Fit information			
					u-msq	u-t	w-msq	w-t
0	5	ET	2.63		1.74	7.07	1.72	6.67
0	5	CSI	0.39		0.86	-1.37	0.83	-1.73
0	5	CSI	-0.02		1.13	1.13	1.06	0.49
0	7	DCI	-0.23		0.37	-4.92	0.36	-9.11
0	7	DCI	-0.37		0.27	-7.17	0.27	-11.75
0	7	DCI	-0.25		0.42	-5.09	0.41	-10.37
0	9	UC	1.39		2.19	6.23	2.2	6.46
0	12	DCI	-0.22		0.36	-8.74	0.35	-18.33
0	12	DCI	-0.14		0.39	-8.25	0.37	-16.41
0	12	ET	3.90		1.64	4.67	1.66	4.88
0	12	ET	5.57		1.24	1.73	1.29	2.21
0	12	CSI	0.49		1.33	2.89	1.23	2.18
0	12	CSI	0.15		1.26	2.42	1.15	1.42
1	16	UC	0.83					
1	18	UC	1.60					
1	18	DCI	0.09					
1	18	DCI	0.89					
1	18	DCI	0.34					
1	18	DCI	0.65					
1	19	ET	1.56					
1	19	ET	2.47					
1	20	DCI	0.37					
1	20	DCI	0.48					
1	20	DCI	0.63					
1	25	DCI	-1.11					
1	25	DCI	-1.31					
1	25	DCI	0.74					
1	26	UC	2.14					
1	28	ET	0.93					
1	28	ET	1.93					
1	28	CSI	1.77					
1	28	CSI	1.86					
1	29	DCI	1.07					
1	31	UC	0.98					
1	31	ET	2.29					
1	32	DCI	0.43					
1	32	DCI	0.27					
1	33	UC	0.23					
1	35	UC	1.73					
1	36	UC	0.85					
1	36	UC	0.80					
2	37	UC	3.61					
2	38	ET	-0.07					
2	38	ET	1.58					
2	38	CSI	-0.55					
2	38	CSI	-0.60					
2	40	DCI	1.73					
2	40	UC	3.86					
2	42	ET	1.65					
2	44	DCI	0.33					
2	44	DCI	1.14					
2	44	DCI	1.31					
2	44	DCI	3.02					
2	52	UC	3.68					
2	52	UC	3.99					
2	52	DCI	3.05					
2	52	DCI	2.27					
2	50	DCI	-0.48					
2	50	DCI	1.65					
2	53	UC	3.61					
2	51	ET	0.49					
2	51	ET	2.73					

Multidimensional modeling

2 56 DCI 1.23

Multidimensional modeling

Table 7b: Step fit information person distribution information for the multidimensional model of the 1994-95 field test link test and activity data

Parameter	Time 0 fit				Time 1 fit				Time 2 fit			
	u- msq	u-t	w- msq	w-t	u- msq	u-t	w- msq	w-t	u- msq	u-t	w- msq	w-t
DCI step 1	24.3	46.5	28.0	41.8								
	1	6	6	3								
DCI step 2	36.3	56.7	43.7	35.9								
	4	4	9	3								
DCI step 3	44.0	62.1	56.7	28.7								
	8	3	1	9								
ET step 1	1.59	6.92	1.43	4.64								
ET step 2	1.07	1.01	1.07	1.08								
ET step 3	1.2	2.57	1.01	0.23								
UC step 1	1.21	3.14	1.05	0.75								
UC step 2	1.09	1.43	1.01	0.19								
UC step 3	2.43	16.6	1.24	3.57								
		4										
CSI step 1	1.14	1.92	1.32	4.44								
CSI step 2	1.43	5.21	1.47	7.03								
CSI step 3	1.13	1.74	1.14	2.3								

Deviance:
24277.07

Means and standard
deviations

	DCI	ET	UC	CSI
Sep-94	- 0.22	2.78	1.26	0.36
	1.2			
	8	0.04	9	
Jan-95	0.79	1.61	0.90	1.27
	1.28	1.5	2.23	1.2
	9		7	
Apr-95	1.57	1.59	0.89	1.35
	3.09	1.4		
	1	0.23	9	

Correlations Sep- Jan Apr-

Multidimensional modeling

	94			-	95			95		
	DCI	ET	UC	DCI	ET	UC	DCI	ET	UC	
ET	.36			0.6			0.67			
UC	.19	.71		0.5	0.63		0.81	0.6		
CSI	.59	.79	.61	0.6	1.00	0.6	0.66	0.8	0.6	
				0		3		0	7	

Multidimensional modeling

Table 8: Reliabilities of person estimates for the multidimensional analysis of link tests and activity set

	DCI			ET			UC		
Time	pop var	var EAP	reliability	pop var	var EAP	reliability	pop var	var EAP	reliability
Time 0	0.05	0.01	.21	1.59	1.15	.72	1.63	1.14	.70
Time 1	2.58	2.20	.85	1.61	1.33	.83	2.54	1.93	.76
Time 2	2.54	2.19	.86	1.81	1.47	.81	1.98	1.70	.86

CSI		
pop var	var EAP	reliability
0.80	0.56	.70
1.61	1.33	.83
1.40	1.05	.75

Multidimensional modeling