Complex Composites: Issues That Arise in Combining Different Modes of Assessment

Mark Wilson, University of California, Berkeley

Wen-chung Wang, National Taiwan University

Data from the California Learning Assessment System are used to examine certain characteristics of tests designed as the composites of items of different modes. The characteristics include rater severity, test information, and definition of the latent variable. Three different assessment modes—multiple-choice, open-ended, and investigation items (the latter two are referred to as performance-based modes)—were combined in a test across three different test forms. Rater severity was investigated by incorporating a rater parameter for each rater in an item response model that then was used to analyze the data. Some rater severities were found to be quite extreme, and the impact of this variation in rater severities on both total scores and trait level estimates was examined. Within-rater variation in rater severity also was examined and was found to have significant variation. The information contribution of the three modes was compared. Performance-based items provided more information than multiple-choice items and also provided greatest precision for higher levels of the latent variable. A projection-like method was applied to investigate the effects of assessment mode on the definition of the latent variable. The multiple-choice items added information to the performance-based variable. The results of the analysis also showed that the projection-like method did not practically differ from the results when the latent trait was defined jointly by both the multiple-choice and the performance-based items. *Index terms: equating, linking, multiple assessment modes, polytomous item response models, rater effects.*

Multiple-choice (MC) items have been used widely in psychological and educational testing for many years. Administrative convenience and computerized scoring make them very convenient. However, MC items have been criticized as being inadequate to fully assess examinees' abilities. Moreover, test-wiseness may seriously contaminate the measurement. Recently, there has been an increased interest in performance-based (PB) items (or constructed-response items) as an alternative to MC items. A PB item refers to any item format that requires the examinee to generate a response in any way other than selecting from a short list of alternative answers as in MC items (Pollack, Rock, & Jenkins, 1992). The different types of response formats, such as MC items and the many types of PB items, are referred to here as different assessment modes.

The main advantages of PB items are that: (1) they provide a more direct representation of content specifications (face validity and content validity), (2) they provide more diagnostic information about examinees' learning difficulties from their responses, (3) examinees prefer them to MC items, and (4) the test formats may stimulate the teaching of important skills, such as problem solving and essay writing (Grima & Liang, 1992). However, MC items are more economical to score and have well-established patterns of reliability. PB items are more difficult to score objectively and reliably, but they may have more systemic validity (Wainer & Thissen, 1993).

MC and PB items often are presented as antagonistic item format alternatives. However, Wilson (1994; Wilson & Adams, in press) suggested combining the scores from the different assessment modes in order to take advantage of the positive aspects of each and to attempt to avoid the negative aspects. Before such an

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 19, No. 1, March 1995, pp. 51–71

0146-6216/95/010051-21$2.30

option can be considered, it needs to be established that what each assessment mode measures is sufficiently similar to justify combining scores. In order to combine scores from the different assessment modes, an assumption of a common latent variable is necessary.

The 1992 California Learning Assessment System (CLAS) mathematics assessment combines scores from different modes. CLAS data were used to examine three issues that arise when scores from different assessment modes are combined. The first issue relates to the PB mode; the effect of using human raters in the assessment system was investigated. The second issue arises from the added flexibility inherent in using different assessment modes and addresses how many of each type of item should be used to attain a specified degree of precision. The third issue arises when a justification is needed for using the psychometric concept of dimensionality in establishing test validity; that is, should the vector sum of the different modes be used, or should it be controlled in some way? This is not an exhaustive list.

THREE ISSUES

Rater Effects in Performance-Based Items

There are many potential effects of raters on PB items. Two are focused on here: the variation of rater severity between raters and the variation of rater severity within raters across different items and modes (other rater effects are discussed below in the discussion section). Rater severity is the tendency of a rater to consistently assign scores that are higher (or lower) than the average rater.

Interrater variation in rater severity. The scores from a MC item should always be identical even when different raters are used. However, for a PB item, this is usually not the case. Thus, the scores of a PB item are rater dependent. Lunz, Wright, & Linacre (1990) pointed out that despite thorough training, raters still vary in severity. If an item is judged by a severe rater, the score will be lower than that given by a lenient rater. It is obviously unfair for examinees if rater severity is not identical and no compensation is made for the differences in scores due to rater severity. For a large-scale test in which a great many raters are involved, it may be reasonable to try to ensure that the raters follow the same rating criteria so that they have similar rating severities, but it will be very difficult to ensure that all have the same severity. Hence, it may be necessary to estimate the values of those severities and compensate for any differences accordingly.

Traditionally, the assessment of interrater consistency has focused on the reliability of the rating instrument and procedure. The intraclass correlation coefficient has been used as a measure of average interrater reliability by several authors (Shrout & Fleiss, 1979; Winer, 1962). Coefficient alpha (Cronbach, 1951) is another estimate of average rater reliability; however, this can conceal differences in severity. Dillon & Mulani (1984) applied latent class analysis to estimate the probability of each response pattern across raters. Van den Bergh & Eiting (1989) assumed multiple quantitative ratings to be congeneric, tau-equivalent, or parallel and then used LISREL (Jöreskog & Sörbom, 1988) to fit these models. Overall & Magee (1992) proposed several simple models, such as the disattenuation model, the common factor model, the external criterion model, the treatment effects model, and the regression model, to estimate individual reliabilities of raters from simple bivariate correlations among their ratings.

Item response modeling focuses on rater severity as an important aspect of rater consistency that needs to be examined. For example, an extension of the Rasch model, the FACETS model (Linacre, 1988, 1989), has been applied to examine the effects of items, examinees, raters, tasks, and rating scales on test scores; to clarify and control the multiple facets of an oral examination involving different protocols, raters, and candidates (Lunz, Stahl, Wright, & Linacre, 1989); to examine rater severity in essay, clinical, and oral forms of tests and grading sessions (Lunz & Stahl, 1990); and to measure writing ability, writing-task difficulty, domain difficulty, and rater severity (Engelhard, 1992). These studies all concluded that rater

severity plays a significant role in PB items and should not be neglected.

Within-rater variation in rater severity. Large-scale testing programs use several forms of a test. These forms may not be exactly parallel but can be expected to be very similar. Traditionally, horizontal equating is applied to calibrate item and person parameters among these forms by either common items or common examinees across forms. Assuming that the parameters of these common items or examinees remain stable in different forms, other item parameters can be calibrated accordingly.

It is also possible to propose test implementation designs that have no common examinees or items, but do have common raters; in fact, there may be situations in which such designs are, in a practical sense, the most convenient. This is, at least in part, the situation that pertains to the CLAS data. Just as it is plausible to link forms through common items, it is also plausible to link different forms through common raters—by assuming that the raters' parameters are unchanged across test forms. Both common items and common raters were used here to link forms. In the present context, it is also possible to consider whether assessment modes rather than forms might be the correct locus of the uniformity required for linking.

Information and Standard Errors

The issue of how many items from each mode should be used in a particular situation depends principally on the required standard error (SE) for the resultant measure and the SEs attributable to each mode. This is usually expressed using Fisher information, which is effectively the reciprocal of the square of the SE evaluated at each point of the measure. The consideration of information available from different assessment modes has a relatively long history in the study of polytomous item response models. Samejima (1969, 1977) and Thissen (1976) demonstrated typical patterns of relative information from dichotomous and polytomous items. They investigated whether more information was obtained by scoring polytomously rather than dichotomously. They found that more information was obtained using polytomous scoring, assuming that the scoring scheme was appropriate. The same procedures were followed here, but they were applied to a somewhat different situation. However, here the question was not how information changes depending on how the items are scored, as it was in the earlier research, because the items were considered to have a standard unchangeable scoring scheme. Instead, the effect on the total SE for different arrangements of items from different assessment modes was investigated.

Dimensional Definition of the Measured Variable

Because MC items and PB items were combined and a composite score was reported, the latent variable indicated by the composite score should be a combination of the two. In the method used above in investigating issues 1 and 2, this combination is a function of the relative scores assigned to the two different modes and of the correlation between the modes. Because this correlation varies somewhat from context to context, the exact dimensional definition of the resultant measure will tend to differ also. A higher degree of control could be exerted over the final measure. For example, perhaps one of these modes would be considered to be predominant as the basis for definition of the dimension. For example, the PB items might embody the most genuine definition of the latent variable, but there may still be interest in augmenting the PB information with the MC information that is consistent with it.

To investigate this idea, a projection-like approach is proposed. It is similar to a technique described by Luecht & Miller (1992), but it is considerably easier to use because it is based on parameter anchoring rather than factorial analysis. The procedure is as follows: (1) decide on one particular mode that defines a "best" direction (on the basis of substantive preference), for example, the PB items; (2) estimate the parameters for the PB items alone; (3) anchor the values for the PB items, and estimate the parameters for the MC items and for the examinees based on the anchored values. Effectively, this projects the information from the MC items onto the latent variable defined by the PB items alone.

THE ITEM RESPONSE MODEL

The random coefficients multinomial logit model (RCMLM; Adams & Wilson, 1992) was used in this study. To describe the items, there is a vector of p item parameters $\xi' = (\xi_1, \xi_2, ..., \xi_p)$. Linear combinations of these parameters are used in the response probability model and are identified as item difficulty parameters, step parameters, and so forth, by the structure of the linear combination. The linear combinations are defined by p-element design vectors $\mathbf{a}'_{ik}$, ($i = 1, ..., I$ and $k = 1, ..., K_i$), which for notational convenience can be denoted collectively by the design matrix

$$\mathbf{A} = \left(\mathbf{a}'_{11}, \mathbf{a}'_{12}, \ldots, \mathbf{a}'_{1k_1}, \mathbf{a}'_{21}, \ldots, \mathbf{a}'_{2k_2}, \ldots, \mathbf{a}'_{Ik_I}\right) . \tag{1}$$

$\mathbf{A}$ has

$$\sum_{i=1}^{I} K_i \tag{2}$$

rows and p columns. The dependent variable X_i is the response of an individual to item i and it can take K_i values ($k = 1, 2, ..., K_i$). In the measurement context, each of the K_i possible values for X_i is assigned a score or "level" that is given by a mapping function $B_i(k)$, which gives the performance level of the observed response k to item i. The response probability model then gives the probability of a response in category x of item i as

$$P_i(x; \mathbf{A}, \xi|\theta) = \frac{\exp\{\theta B_i(x) + \mathbf{a}'_{ik}\xi\}}{\sum_{k=1}^{K_i} \exp\{\theta B_i(k) + \mathbf{a}'_{ik}\xi\}} . \tag{3}$$

Equation 3 gives the probability of observing each type of response conditional on a particular value of the attribute or trait level, θ. The function B_i and the vectors in $\mathbf{A}$ define the RCMLM for item i. A detailed description of this model and the marginal maximum likelihood (MML) algorithm used to estimate its parameters is given in Adams & Wilson (1992, in press).

For example, consider the simplest unidimensional item response model, the Rasch (1960/80) model. In the usual parameterization of the Rasch model for a set of I dichotomous items there are I item difficulty parameters, δ, and there is also a constraint that locates the zero of the scale. In this model the general form of the response probability function is:

$$P_i(x = 1|\theta) = \frac{1}{1 + \exp(\theta - \delta_i)} \tag{4}$$

and

$$P_i(x = 2|\theta) = \frac{\exp(\theta - \delta_i)}{1 + \exp(\theta - \delta_i)} , \tag{5}$$

where $x = 1$ for a correct response, and $x = 2$ for an incorrect response, and it is assumed that the mean of the θ distribution is fixed at 0.

Each of the items has two response categories, $x = 1$ or $x = 2$, and in this model the scoring function is $B_i(x) = x - 1$ for all i. This is consistent with a score of 0 for the first category of each item and a score of 1 for the second category. The model for three items with parameter vector (δ_1, δ_2, δ_3) is specified by a design matrix $\mathbf{A}$ as follows:

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}. \tag{6}$$

The design matrix has I columns, one for each item, and $2I$ rows, one for each of the available responses (i.e., two for each item). The rows are grouped in pairs with the first of each pair corresponding to the first category of response to each item and the second corresponding to the second category of response.

The first pair of rows in **A** corresponds to the categories for Item 1, the first row for the first category and the second row for the second. The -1 in the (2,1) position may be interpreted as indicating that the first parameter will summarize how much more difficult it is to give the correct response to Item 1 than it is to give the incorrect response. Similar interpretations can be made for the other two parameters. Mathematically, note that setting the first row to all 0s and coupling this with a scoring function that gives a score of 0 for the first category when substituted into Equation 3 gives the numerator of Equation 4. Similarly, setting a -1 in the first column of the second row and a scoring function that assigns a score of 1 to the second category results in the numerator for Equation 5.

The constraint can be changed to the requirement that the sum of all the items be 0 by changing **A** to

$$\mathbf{A} = \begin{bmatrix} 0 & 0 \\ -1 & 0 \\ 0 & 0 \\ 0 & -1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}. \tag{7}$$

Note that this constraint is achieved by setting the parameter for the third item to be the negative sum of the remainder. That is why there are only two columns for **A**—there are only two free item parameters.

More complicated item response models may be expressed using equally straightforward matrices. For example, the partial credit model (Masters, 1982) is designed for assessment situations with multiple levels of achievement within each item. Here each item can be described by a set of parameters, one parameter for each successive step from one level to another (δ_{ij} is the step from category j to $j+1$ for item i). A general expression for this model is:

$$P(x_i = j|\theta) = \frac{\exp\left[(j-1)\theta - \sum_{h=1}^{j} \delta_{ih-1}\right]}{\sum_{k=1}^{K_i} \exp\left[(k-1)\theta - \sum_{h=1}^{k} \delta_{ih-1}\right]}, \tag{8}$$

with the convention that $\delta_{i0} \equiv 0$.

For an instrument with, say, three items and three categories in each, each of the items has three response categories, $x = 1, x = 2$, and $x = 3$, and in this model the scoring function is $B_i(x) = x - 1$ for all i. This is consistent with a score of 0 for the first category of each item, a score of 1 for the second category, and a score of 2 for the third category. Assuming that the usual constraint is applied in the form of a zero mean for the person distribution, the model for parameter vector (δ_{11}, δ_{12}, δ_{21}, δ_{22}, δ_{31}, δ_{32}) then is specified by a design matrix **A** as follows:

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 \end{bmatrix}. \quad (9)$$

In this matrix, the −1 in the (2,1) position indicates that the first parameter will summarize how much more difficult it is to give the score 1 response to Item 1 than it is to give the score 0 response. The −1 in the (3, 2) position indicates that the second parameter will summarize how much more difficult it is to give the score 2 response to Item 1 than it is to give the score 1 response [the −1 in the (3, 1) ensures that this comparison is conditional on attainment of score 1]. To impose a zero sum constraint on these step estimates, the design matrix can be altered by making the last row (1, 1, 1, 1, 1) and leaving out the last column (because the constraint expresses the last parameter as the negative sum of the remaining columns).

The RCMLM is a generalized Rasch model that integrates many existing Rasch models, such as the simple logistic model (Wright & Panchapakesan, 1969), the rating scale model (Andrich, 1978), the partial credit model (Masters, 1982), the FACETS model (Linacre, 1989), and the ordered partition model (Wilson, 1992). In addition, this model also provides a great deal of flexibility to design customized models according to particular test situations. It also allows different numbers of categories in different items, complex patterns of raters, and multilevel structures. Examples of such applications are given in Draney, Pirolli, & Wilson (in press), Moore (in press), Wang & Wilson (in press), and Wilson & Adams (in press).

MML (Bock & Aitkin, 1981) estimation is implemented in the RCML computer program (Adams & Wilson, 1992). Although joint maximum likelihood estimation is widely applied by several other computer programs, such as BICAL (Wright, Mead, & Bell, 1980) and LOGIST (Wingersky, Barton, & Lord, 1982), it does not provide consistent estimates as sample size increases. MML estimation alleviates this problem by integrating over the θ distribution and estimating the item parameters using the marginal distribution of θ. In the item response model, the latent variable θ is associated with each case (or person) in a certain population. The population distribution is given by the density function $g(\theta; \boldsymbol{\alpha})$ and the corresponding cumulative distribution $G(\theta; \boldsymbol{\alpha})$, both of which contain a vector of parameters $\boldsymbol{\alpha}$ to characterize the distribution.

EXAMPLE APPLICATION

Method

Data

Data from the CLAS 1992 spring field test for grade 4 mathematics Forms M0423, M0424, and M0425 (hereafter referred to as Forms 3, 4, and 5, respectively) were used. A primary purpose of the field test was to generate information on which to base a decision regarding combining MC and open-ended (OE) items that would satisfy some minimum requirements in terms of SEs of examinee estimates. The decision to combine information from MC and OE formats was made by the test constructors. The items in Forms 3, 4, and 5 were designed to measure an underlying mathematics ability. Although the three forms were not expected to be exactly parallel, all three forms contained 20 MC items, two OE items, and one investigation (IN) item.

The OE items required examinees to explain their reasoning processes used in solving problems and were scored from 0 to 5 by two raters independently. The IN items required examinees to solve three activities with a partner in approximately an hour and one-half and to answer the questions given. The

responses to the IN item were scored from 0 to 5 by one rater only.

435 examinees completed Form 3, 496 completed Form 4, and 488 completed Form 5. 49 raters scored the OE and the IN items in the three forms. Among them, 11 raters scored only the IN items, 5 raters scored both the OE and the IN items, and the other 33 scored only the OE items. The scores from each item type (assessment mode) were summed and this total sum was reported to the examinees. This assumed that the three forms were of equal difficulty and that all raters were equally severe.

Linking. Given that the three forms were taken by three separate samples of examinees, the three forms had to be linked. There was one common MC item across all three forms. Forms 4 and 5 had five additional MC items in common. Thus, there were 53 distinct MC items in the three forms (one item was common in all three forms). Forms 3 and 4 had 8 raters in common. Nothing else was common across the three forms. Under this circumstance, apart from the one common MC item for all three forms, the linking information between Forms 4 and 5 was common items they shared and the linking information between Forms 3 and 4 was common raters.

An alternative way to link the three forms would be to assume that the three samples of examinees (one for each form) were equivalent. Then the three distributions could be centered on the same value and, subject to that assumption, the linking would be accomplished. The item and rater linkage procedure was used in this case for illustrative purposes.

Data preparation. Prior to analyzing the data, the three samples from Forms 3, 4, and 5 were combined into a complete dataset of 1,419 examinees. Each examinee completed only 20 MC items, two OE items, and one IN item. For each examinee, the other 33 (53 – 20 = 33) MC items, 4 OE items, and 2 IN items were treated as missing at random. This is consistent with the testing arrangement because the forms were distributed randomly within each class.

In addition, the OE and IN items were changed from an item-oriented organization into a rater-oriented format. Because there were nine PB items (two OE items and one IN item per form) to be scored by some subset of the 49 raters, the data were organized as 441 (9 × 49) rater-items. Each examinee could have obtained only five scores from raters (two scores from each of the two OE items and one score from the IN item); the other 436 rater-items were treated as missing at random. This too was consistent with the testing arrangement, because raters were not assigned to ratees in a systematic way. With this organization, and given sufficient data, the parameters for the 441 rater-items could be estimated, in theory. However, because not every rater scored every item on every form, only 85 rater-items were required. Therefore, 94 parameters for these 85 rater-items were actually estimated, assuming a somewhat simpler model than is theoretically possible.

The Scoring Vector and the Design Matrix

Scoring vector. The scoring vector (0 = incorrect, 1 = correct) for these 53 MC items was the same as that for the Rasch dichotomous model, and the sum of the item difficulties was constrained to be 0. Therefore, the scoring vector of a MC item was defined as $\mathbf{b}_M = (1, 1)'$, where M denotes a MC item. Accordingly, the first 106 elements in the scoring vector were:

$$\mathbf{B}_M = (\mathbf{b}'_{1M}, \mathbf{b}'_{2M}, \ldots, \mathbf{b}'_{53M})' = (0\,1\,0\,1 \ldots 0\,1)'. \tag{10}$$

Each of the 441 rater-items had 6 response categories scored from 0 to 5, respectively, resulting in identical scoring vectors, $\mathbf{b}_R = (0\ 1\ 2\ 3\ 4\ 5)'$, where R denotes a rater-item. These vectors were collected into a scoring vector with 2,646 (441 × 6) elements:

$$\mathbf{B}_R = (\mathbf{b}'_{1R}, \mathbf{b}'_{2R}, \ldots, \mathbf{b}'_{441R})' = (0\,1\,2\,3\,4\,5\,0\,1\,2\,3\,4\,5 \ldots 0\,1\,2\,3\,4\,5)'. \tag{11}$$

The first six elements (0 1 2 3 4 5) are the scoring vector of the first rater-item, the second six elements are

the scoring vector of the second rater-item, and so on. The vectors in Equations 10 and 11 were collected into the scoring vector **B**,

$$\mathbf{B} = (\mathbf{B}'_{\mathrm{M}}, \mathbf{B}'_{\mathrm{R}})', \tag{12}$$

with 2,752 (106 + 2,646) elements.

Design matrix. The design matrix **A** was constructed using the following order of item parameters: 52 item difficulty parameters, one for each of the first 52 MC items; nine overall item difficulty parameters, one for each of the OE items and the IN items; nine sets of rating-scale style step difficulties, one set of four for each OE item and the IN items; and 49 rating severity parameters, one for each rater. There was a total of 146 parameters.

Because the sum of the 53 MC item difficulty parameters was constrained to be 0, the item difficulty parameter of the 53rd MC item was minus the sum of the first 52 MC item difficulty parameters. The other 94 (146 − 52) parameters pertained to 441 rater-items. In the dataset, these 441 rater-items were arranged in sequence of rater number; therefore, the first nine rater-items belonged to the first rater, the second nine rater-items belonged to the second rater, and so on. Among those nine rater-items, the first two belonged to the first and second OE items of Form 3, respectively, and the third to the IN item of the same form. Similarly, the subsequent three rater-items belonged to the first and second OE items and IN items of Form 4, respectively, and the last three rater-items to those of Form 5, respectively.

The design matrix can be partitioned into four parts:

$$\mathbf{A}_{(2{,}752\times 146)} = \begin{bmatrix} \mathbf{A}_{1_{(106\times 52)}} & \mathbf{A}_{2_{(106\times 94)}} \\ \mathbf{A}_{3_{(2{,}646\times 52)}} & \mathbf{A}_{4_{(2{,}646\times 94)}} \end{bmatrix}, \tag{13}$$

where $\mathbf{A}_1$ corresponds to the 53 MC items and their 52 free item parameters, and $\mathbf{A}_4$ corresponds to the 441 rater-items and their 94 free parameters. $\mathbf{A}_2$ and $\mathbf{A}_3$ are both zero matrices because the interaction between the MC items and the rater-items was not modeled. Matrix $\mathbf{A}_1$ was

$$\mathbf{A}_{1_{(106\times 52)}} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 & 0 \\ & & \vdots & \ddots & \vdots & & \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ -1 & -1 & -1 & \dots & -1 & -1 & -1 \end{bmatrix}. \tag{14}$$

The first row vector in $\mathbf{A}_1$ represents the linear combinations of the item parameters as they relate to the first response category (incorrect response) to the first MC item. They are all 0s because that category was treated as a reference. The second row vector corresponds to the linear combinations of the item parameters as they relate to the second response category (correct response) to the first MC item. The third and the fourth row vectors pertain to the first and second categories of the second MC item, respectively, and so on. The last row vector accomplished the constraint of making the parameters of the last MC item equal minus the sum of the item difficulties of the other 52 MC items.

$\mathbf{A}_4$ was a 2,646 × 94 matrix. For simplicity, let $\mathbf{A}_4$ be partitioned as follows:

$$\mathbf{A}_{4_{(2,646\times 94)}} = \begin{bmatrix} \mathbf{D} & \mathbf{S} & \mathbf{R} & \mathbf{O} & & \mathbf{O} & \mathbf{O} \\ \mathbf{D} & \mathbf{S} & \mathbf{O} & \mathbf{R} & \cdots & \mathbf{O} & \mathbf{O} \\ \mathbf{D} & \mathbf{S} & \mathbf{O} & \mathbf{O} & & \mathbf{O} & \mathbf{O} \\ & & \vdots & \ddots & \vdots & & \\ \mathbf{D} & \mathbf{S} & \mathbf{O} & \mathbf{O} & & \mathbf{O} & \mathbf{O} \\ \mathbf{D} & \mathbf{S} & \mathbf{O} & \mathbf{O} & \cdots & \mathbf{R} & \mathbf{O} \\ \mathbf{D} & \mathbf{S} & \mathbf{O} & \mathbf{O} & & \mathbf{O} & \mathbf{R} \end{bmatrix}, \tag{15}$$

where **D** is a 54 × 9 matrix, **S** a 54 × 36 matrix, **R** a 54 × 1 matrix, and **O** a 54 × 1 zero matrix. Every **R** corresponds to a specific rater. Because there were 49 raters, matrix $\mathbf{A}_4$ contained 49 **D**, **S**, and **R** submatrices. **D** can be viewed as a subdesign matrix pertaining to overall item difficulty, **S** contained step difficulties, and **R** contained rater severity. **D** can be partitioned into

$$\mathbf{D} = \begin{bmatrix} \mathbf{C} & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* \\ \mathbf{O}^* & \mathbf{C} & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* \\ \mathbf{O}^* & \mathbf{O}^* & \mathbf{C} & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* \\ \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{C} & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* \\ \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{C} & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* \\ \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{C} & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* \\ \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{C} & \mathbf{O}^* & \mathbf{O}^* \\ \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{C} & \mathbf{O}^* \\ \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{O}^* & \mathbf{C} \end{bmatrix}, \tag{16}$$

where $\mathbf{C} = (0\ 1\ 2\ 3\ 4\ 5)'$, $\mathbf{O}^*$ is a 6 × 1 zero matrix, and

$$\mathbf{S} = \begin{bmatrix} \mathbf{T} & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' \\ \mathbf{O}' & \mathbf{T} & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' \\ \mathbf{O}' & \mathbf{O}' & \mathbf{T} & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' \\ \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{T} & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' \\ \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{T} & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' \\ \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{T} & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' \\ \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{T} & \mathbf{O}' & \mathbf{O}' \\ \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{T} & \mathbf{O}' \\ \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{O}' & \mathbf{T} \end{bmatrix}, \tag{17}$$

where $\mathbf{O}'$ is a 6 × 4 zero matrix, and

$$\mathbf{T} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{18}$$

C can be viewed as a subdesign matrix pertaining to the overall item difficulty of its corresponding rater-item, and **S** to its step difficulties. Therefore, submatrices **C** and **T** in the first row of **D** and **S**, respectively, correspond to the first rater-item, which is the first OE item of Form 3; similarly, those in the second row and third row of **D** and **S** correspond to the second OE item and the IN item of Form 3, respectively. The following three rows correspond to Form 4, and the last three to Form 5, respectively.

R is a subdesign matrix pertaining to rater severity. Rater severities were assumed to remain stable across the OE items and the IN items and different forms (Forms 3, 4, and 5). In addition, item difficulties

and step difficulties were assumed to remain relatively constant for each rater. Therefore, **R** was defined as

$$\mathbf{R} = (0\,1\,2\,3\,4\,5\,0\,1\,2\,3\,4\,5 \ldots 0\,1\,2\,3\,4\,5)', \tag{19}$$

where the first six elements correspond to the first rater-item, the second six elements to the second rater-item, and so on.

The availability of data can impose limits on the estimability of complex models such as these. Thus, to estimate the full model defined by **A**, each of the raters would have had to rate several examinees on each of the 9 OE and IN items. This was not the case for the CLAS data, so that the actual design matrix used was somewhat smaller than **A**, leaving out entire rows where raters did not rate certain items. For this reason, and because of the need to impose a constraint on the rater facet, the actual design matrix used was a modification of **A**; however, it is important to see **A** as the generic model that was adapted to specific data-driven circumstances.

Calibration of Parameters

Parameter estimates. The θ distribution, the item difficulties of the MC items, the OE items, the IN items, and the rater severities are shown graphically in Figure 1, plotted on the logit (θ) scale (note that item difficulty was constrained to have a mean of 0). The θ estimates ranged from −1.37 to 2.71, with a mean of .23 and a variance of .61.

The 20 MC items in Form 3 are numbered from 1 to 20, those in Form 4 are numbered 16, and 21 to 39, and those in Form 5 are numbered 16, 24, 36 to 39, and 40 to 53. The item difficulties of the 53 MC items ranged from −2.42 for Item 47 to 2.10 for Item 5. The mean and the variance of the item difficulties of the 20 MC items in Form 3 were .19 and 1.33, respectively, those of Form 4 were −.18 and .74, respectively, and those of Form 5 were −.16 and 1.08, respectively. Therefore, the MC items in Form 3 were more difficult and more variable than those in the other two forms.

The difficulties of the OE and the IN items in Figure 1 were transformed into level thresholds from the overall difficulties and step difficulties in the estimation routines. The level threshold is the point on the θ continuum at which the probabilities of reaching and not reaching that level are both .5 (Masters & Wilson, 1991). That is, the first threshold is the point on the continuum at which the probability of being in the first category equals the sum of the probabilities of being in all those above, the second threshold is the point at which the probability of being in the first or second category is equal to the probability of being in the third or above, and so forth. This particular expression is useful for visual interpretation of the results of analyses of polytomous item response models.

For items in which there is a many-to-one relationship between categories and levels [see, e.g., Wilson's (1992) ordered partition model], a formulation like this is necessary for graphical presentation of the item parameter estimates. For an item with six categories, the thresholds are defined as follows. Let $P_0(\theta)$ be the probability of a response in Level 0 [i.e., $P_0(\theta) = P(x = 0|\theta)$ in Equation 8], $P_1(\theta)$ be the probability of a response in Level 1, $P_2(\theta)$ be the probability of a response in Level 2, and so forth. Because there are six response categories (Level 0 to Level 5), five step difficulties can be estimated, and consequently five thresholds also can be calculated. These five thresholds are the solutions—θ_1, θ_2, θ_3, θ_4, and θ_5—to the following five equations:

$$P_0(\theta_1) = 1/2, \tag{20}$$

$$P_0(\theta_2) + P_1(\theta_2) = 1/2, \tag{21}$$

$$P_0(\theta_3) + P_1(\theta_3) + P_2(\theta_3) = 1/2, \tag{22}$$

Figure 1
θ Distribution, and Item and Rater Parameters for the Multiple-Choice, Open-Ended, and Investigation Items
(Rater Numbers for Raters Who Rated Only Investigation Items are in **Bold**; Those Who Rated Both Open-Ended and Investigation Items Are in *Italics*; the Remainder of the Raters Rated Open-Ended Items Only)

$$P_0(\theta_4) + P_1(\theta_4) + P_2(\theta_4) + P_3(\theta_4) = 1/2\,, \tag{23}$$

and

$$P_0(\theta_5) + P_1(\theta_5) + P_2(\theta_5) + P_3(\theta_5) + P_4(\theta_5) + P_5(\theta_5) = 1/2\,. \tag{24}$$

The solutions are calculated numerically.

The CLAS variable. Of the six OE items, the second item of Form 4 (column 4B in Figure 1) and the first item of Form 5 (column 5A in Figure 1) were the most difficult and the least variable. In general, the IN items were more difficult and more variable than the OE items. The IN item of Form 5 (column 5 in Figure 1) was the easiest of the three IN items.

Comparing the θ distribution with the item difficulties of the MC items, the OE items, and the IN items, all of the examinees had probabilities greater than .5 of answering MC Items 47, 4, 14, 25, 39, 50, and 31 correctly (i.e., $\theta > \delta$), and of obtaining one point on both the OE and the IN items (assuming a rater severity of 0 logits). MC Items 17, 3, and 5 and the highest score (5) of all the OE and IN items were very difficult for all examinees.

Figure 1 shows that the item difficulties of the three forms were not equivalent. Form 3 had somewhat more difficult MC items and an easier IN item. Most item difficulties of the MC items in the three forms clustered between −1.00 and 1.00 logit, which correspond to Levels 2, 3, and 4 in the OE items and the IN items. The IN items were somewhat more difficult than the OE items. The fit of the calibration is not reported here; Wang and Wilson (in press) report fit results.

Results

Rater Effects

Interrater Pearson correlations on the three forms for the OE items, based on two ratings per item, are shown in Table 1. The interrater correlations were between .63 and .81, which reflects some inconsistency between raters although they are typical for such rating tasks. For instance, Dunbar, Koretz, & Hoover (1991) reviewed nine studies of writing assessment and found that average rater reliabilities were between .33 and .91. However, although the interrater correlations were satisfactorily high, it is still possible that an examinee's performance was rated by two severe raters while another examinee was rated by two lenient raters. In this case, the scores derived from each pair of raters may be very consistent, resulting in a high interrater correlation, which is misleading.

Table 1
Interrater Correlations of the OE Items in the Three Forms

Form	1st Item	2nd Item
Form 3	.66	.81
Form 4	.63	.71
Form 5	.79	.67

Interrater variation in rater severity. Figure 1 shows that the range of rater severities was 2.67 logits, between −.61 for Rater 48 and 2.06 for Rater 46. The mean, median, and the standard deviation were .94, 1.09, and .56, respectively. There were 11 raters (bold format in Figure 1) who judged only the IN items. The severities of these 11 raters tended to be somewhat more extreme than the others; for example, Raters 48, 41, 39, 40, 44, and 47 were relatively lenient, and Rater 46 was the most severe.

Figure 2 shows 95% confidence intervals for the 49 rater severities. The intervals do not all overlap and the χ^2 statistic for testing equal severity was 771.14 with 48 degrees of freedom. Therefore, it was con-

cluded that the raters were operating with different severities. This is an important finding in the present context because in its field trials CLAS has simply summed rater judgments without making any adjustments for rater variation.

Figure 2
95% Confidence Intervals of the 49 Rater Severities

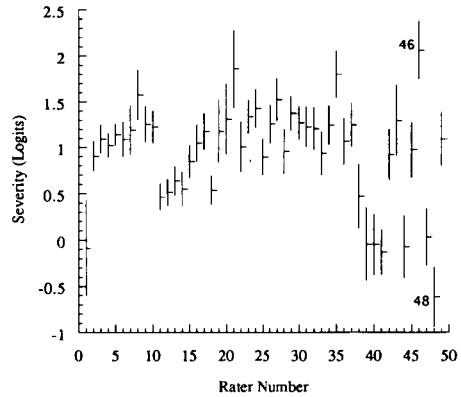


Figure 3 shows item response functions of the IN item of Form 3 rated by Raters 48 (Figure 3a; the least severe rater) and 46 (Figure 3b; the most severe rater). These data illustrate that it was much more difficult for examinees to obtain higher scores from Rater 46 than from Rater 48.

Figure 3
Probability Distributions of the Investigation Item of Form 3 Judged by Raters 48 and 46
a. Rater 48 b. Rater 46

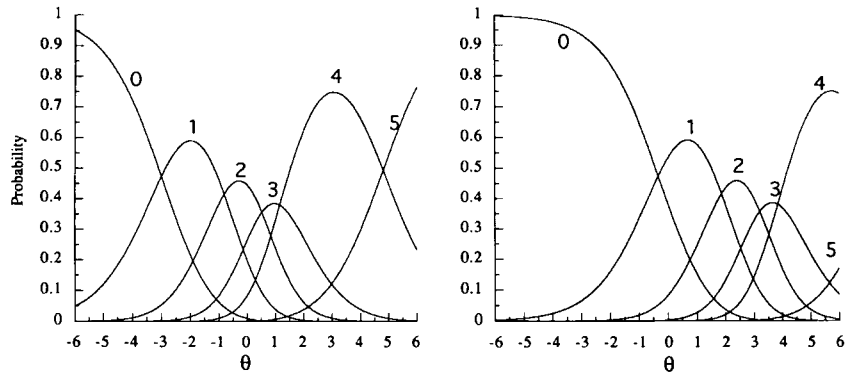
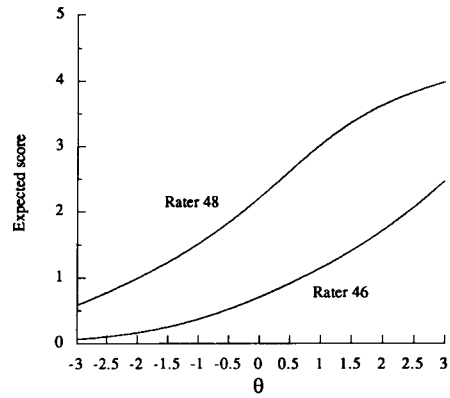


Figure 4 shows the expected scores for the IN item rated by Raters 48 and 46. An examinee with $\theta = 0.0$ would be predicted to have an expected score of 2.2 from Rater 48, and an expected score of .7 from Rater 46. An examinee with $\theta = 2.0$ would have an expected score of approximately 3.7 from Rater 48, and approximately 1.7 from Rater 46. The maximum difference of expected scores derived from these two raters is approximately 2 points (when examinees' trait levels are located between .5 and 2.5). Because all of the OE items and the IN items were judged on a six-point scale (0 to 5), a difference of 2 points represents an important bias.

Figure 4
Expected Scores on the Investigation Item of Form 3
When the Examinees Were Judged by Raters 48 and 46



This bias would not always be detected by a comparison of the original ratings. To put it another way, a rating of 2 derived from Rater 48 represents a θ estimate of −.3, but $\hat{\theta} = 2.4$ for the same rating from Rater 46. Therefore, in a case in which raters vary in severity, the same rating given by two raters does not necessarily reflect the same θ estimates. In other words, total scores are no longer sufficient statistics for θ estimates (as in the simple logistic model), hence analysis of the consistency of total scores is not a guarantee against significant problems in rater consistency.

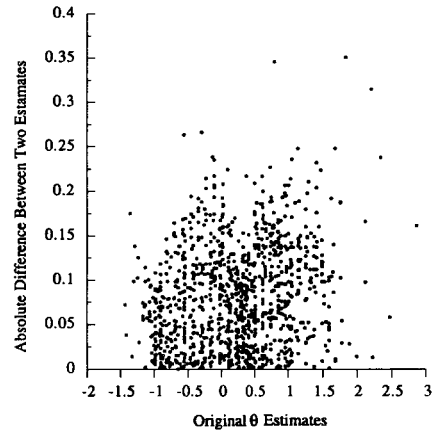
One way that the effect of variations in rater severity on the results can be examined is as follows. If severe raters and lenient raters were defined as those whose severities are located 1 standard deviation (.56 logits) above and below the mean, respectively, there were 4 severe raters (8, 21, 35, and 46) and 7 lenient raters (1, 39, 40, 41, 44, 47, and 48). Suppose the 49 raters were randomly assigned to examinee tests; then the probability of an examinee being judged by a severe rater on an IN item would be .082 (4/49) and by a lenient rater .143 (7/49). Similarly, the probability of an examinee being judged by two severe raters on an OE item would be .007, and by two lenient raters it would be .02. Fortunately, these are small probabilities. If the probabilities had been larger, it would call into question the fairness of the procedures CLAS raters use to grade PB items. Because there were different numbers of severe and lenient raters for OE and IN items, these probabilities differed for the two item types, but the method used to calculate them was the same.

The impact of rater severity on this particular dataset also was investigated by constraining all of the rater severities to be identical (assuming raters are equal in severity) and re-estimating the θs. These new estimates were compared to the original estimates in which different rater severities were taken into account. The mean of the absolute differences in θ estimates between these two models was .08, and the maximum difference was .35. The standard deviation of the absolute differences was .06. Figure 5 shows the absolute differences as a function of the original θ estimates. This mean absolute difference was not very large compared to the SEs of measurement (Table 4).

The variations in rater severities for this dataset were not very influential, because only a few raters differed in severity and because these extreme raters tended to be the raters who scored only the IN items. This consistency problem in the IN items may be due to the lack of a second rating (which was used as a quality control method in the OE items) for the IN items.

However, although the differences were not great on average, the differences in rater severities can have large effects on individual examinees. Assuming a normal distribution of the θ estimates, an approximate

Figure 5
Absolute Differences in θ Estimates With and Without the Equal Rater Severity Assumption



index of the changes in percentiles when the raters were assumed to have equal severities was derived and is shown in Table 2. Because the variance of the original θ estimates was .61, a maximum absolute difference of .35 logits corresponds to a *Z* score of .45, which in turn corresponds to a change in percentiles of approximately 17, assuming the person's original position was located at approximately the mean. (If it was further from the mean, the change in percentiles would be less.) Similarly, the changes in percentiles were below 4 for half of the examinees, below 6 for approximately 75% of the examinees, below 8 for approximately 90% of the examinees, and above 9 for approximately 95% of the examinees.

Table 2
Changes in Percentiles of the θ Estimates When the Raters Were Assumed to Have Equal Severities

Percentage of Examinees	Difference (in logits)	*Z* Score	Percentile
Maximum	.35	.45	17.36
50%	.08	.10	3.98
75%	.12	.15	5.96
90%	.15	.20	7.93
95%	.18	.23	9.10

These findings indicate that the rater severities in this example were not equal and should not be treated as such. If the variation of the rater severity was not taken into account, some of the examinees would be affected significantly on their relative position to other examinees.

Within-rater variation in rater severity. The calibration assumed, in part, that each rater was adequately modeled by a single severity parameter, no matter what he or she was rating. Given that there can be important variation between raters in severity, the question of whether there are important variations within raters naturally arises.

Two related ways to consider within-rater variation were examined here: across modes and across items. In the former case, the design matrix gains extra rows, so that whenever a rater rated a specific mode of item (either OE or IN), a column (and hence, a rater severity parameter) was added. In the latter case, each rater was

modeled to have a different severity whenever he or she rated a different item—so that the design matrix gained rows again, and now had one column (and hence, a rater severity parameter) for every item that a particular rater rated.

The two models described above, and the one described in the previous section, were fit and their respective likelihood ratio χ^2 statistics (G^2) and number of parameters are shown in Table 3. Because each model is a special case of the one below it in Table 3, standard likelihood ratio tests were used to investigate whether there were statistically significant differences in fit between the models. Both tests were significant at the 5% level [Test 1: $\Delta G^2 = 41.48$, degrees of freedom (df) = 5, $p \leq .001$; Test 2: $\Delta G^2 = 113.32$, $df = 31$, $p \leq .001$]. Thus, there were statistically significant amounts of within-rater variation when the differences between modes were considered, and there were additional amounts of statistically significant within-rater variation when individual items within modes were examined.

Table 3
Comparison of Models Invoking Different Degrees of Within-Rater Variation

Model	Number of Parameters	G^2
No within-rater variation in rater severity	146	45,975.32
Across-mode variation in rater severity	151	45,933.84
Across-item variation in rater severity	182	45,820.52

The statistical tests provide evidence of significant within-rater variation. The estimates of rater severity themselves must be examined to determine whether those variations are substantively meaningful. Consider first the across-mode variation. There were only five raters who rated across the three assessment modes, and none of the estimated rater severities were different at a 5% level of statistical significance across the modes. The largest difference was .5 for Rater 25, which was approximately one-third of the maximum between-rater variation in severity.

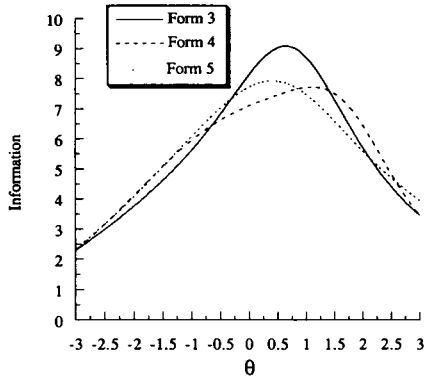
However, examination of the across-item variation showed considerably more and larger effects. Altogether 29 of the 49 raters scored more than one item, and of those 29 raters 26 scored a sufficient amount of examinee data to make comparison reasonable. Of these 26, 8 rater severities were statistically significantly different from 0.0 at the $\alpha = .05$ level. Seven of the discrepancies were between .6 logits and 1.4 logits, but one (for Rater 25) was 2.1 logits. Note that this result for Rater 25 came from items within an assessment mode, as did the other seven. These results cast serious doubt on the validity of using a common rater severity across different items. This is consistent with the way that the item scoring is organized. Note that, although these results cast doubt on the use of common rater severity to link the forms, the illustrative results below are reported for the common rater severity model in order to focus on other aspects of the results without unduly complicating the presentation.

Information and Standard Errors

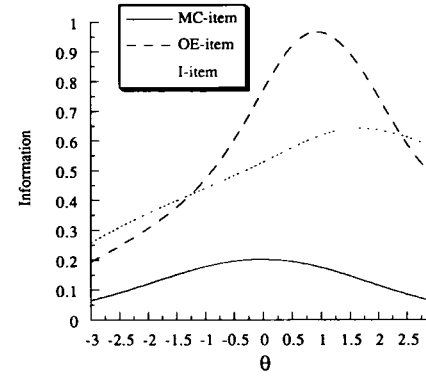
Figure 6a shows the information for the three forms. The three forms did not differ to a marked extent, but some differences can be noted. In the range between $\theta = -1.0$ and 2.0 (where most examinees were located), Form 3 was the most accurate because it provided the highest information in that range. Form 4 provided the greatest information for examinee θ levels around 1.5, and Form 5 provided greatest accuracy around $\theta = .5$.

Figure 6
Information Functions

a. Forms 3, 4, and 5



b. Average Information for Multiple-Choice, Open-Ended, and Investigation Items



Averaging the information across the three forms, Figure 6b shows the average information of a MC, an OE, and an IN item. These results are consistent with results of Samejima (1969, 1976) and Thissen (1976). The maximum information for a MC item is approximately .2 at $\theta = 0.0$, approximately .95 at $\theta = 1.0$ for an OE item, and approximately .65 at $\theta = 1.5$ for an IN item. The difference in location of the maximum points on the variable may be taken as evidence that the item constructors created OE items that were targeted at a higher level than were the traditional MC items and IN items targeted at a higher level than the OE items.

To calculate a typical SE for examinees in the sample, information was averaged across the range of the sample ($\theta = -1.25$ to 2.75), weighted by the distribution of the person estimates, and then converted into SEs. The results of this averaging process are shown in Table 4. Table 4 shows that an average OE item provided approximately 4.5 times as much information than a MC item, whereas an IN item provided approximately 3.2 times as much. Note that the relatively lower information for the IN mode (compared to the OE mode) also reflects the smaller number of examinees at the higher θ values at which the IN items peak. Based on these values, projections of information available from different arrangements of MC, OE, and IN items were calculated, assuming that the examinee distribution remained the same. Table 4 also provides the information and the SE for the actual arrangement of the different assessment modes in this field test

Table 4
Average Information and Associated Standard Error of Item Types and Item Combinations

Type of Item	Average Information	Standard Error (in logits)
Single Item		
MC	.18	2.38
OE	.80	1.12
IN	.56	1.33
Item combinations		
20MC + 2 × 2OE + 1IN (reference)	6.54	.39
2 × 4OE + 1IN	6.91	.38
40MC + 1IN	7.64	.36
5MC + 2 × 2OE + 1IN	4.44	.48
5MC + 4 × 2OE + 1IN	7.61	.36
30MC +1 × 2OE + 1IN	6.36	.40

(20 MC items, 2 OE items rated twice, and 1 IN item), as a reference.

The information and the SE for the reference item combination were 6.54 and .39, respectively. If all MC items were eliminated and the number of OE items was doubled, the information and the SE were approximately the same. Eliminating all OE items and doubling the number of MC items would result in somewhat larger information and a smaller SE. 30 MC items, one OE item, and one IN item would provide almost identical information to the reference combination.

Dimensional Definition of the Measured Variable

For comparison, following the application of the "projection-like" approach described above, three θ estimates were calculated for each examinee: one for PB items alone ($\hat{\theta}_1$), one for all items anchored by the PB items ($\hat{\theta}_2$), and one for all items without any anchoring ($\hat{\theta}_3$).

The correlation matrix of these three $\hat{\theta}$s is shown in Table 5. The elements in the diagonal of Table 5 are the reliability estimates. The reliability of $\hat{\theta}_1$ was approximately .16 lower than $\hat{\theta}_2$ and $\hat{\theta}_3$, which had almost identical reliabilities. This illustrates that the MC mode did indeed contribute extra information to the latent variable defined by the performance mode. The correlation of .81 between $\hat{\theta}_1$ and $\hat{\theta}_2$ indicates that adding the MC information altered the variable somewhat, but not substantially. The correlation of .99 between $\hat{\theta}_2$ and $\hat{\theta}_3$ indicates that the projection-like technique resulted in almost no information loss compared to the use of all items simultaneously and, indeed, produced almost identical results when items were used simultaneously. For these data, the correlation between the total scores on the MC items and the total scores on the PB items was .43, a typical value in CLAS analyses. These results are dataset-specific; that is, a different correlation between the MC items and the PB items would result in a different pattern of correlations if another dataset using different examinees or different tests were used.

Table 5
Correlations and Reliabilities (In the Diagonal) of Different $\hat{\theta}$s

$\hat{\theta}$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
$\hat{\theta}_1$	.63		
$\hat{\theta}_2$	.81	.79	
$\hat{\theta}_3$	.80	.99	.80

Discussion

There are advantages and disadvantages to using only multiple-choice or performance-based items. A test constructor usually considers reliability, validity, cost, testing time, and policy issues when selecting the proper test format. It is proposed here that other information from these different item types should be considered, and another index—the average information of each mode (which is related to reliability)—should be used to help determine the best combination of item types for a specific application.

In the dataset used here, a typical open-ended item and a typical investigation item provided approximately 4.5 and 3.2 times more information, respectively, than a typical multiple-choice item. These multiplicative factors must be weighed against the much greater discrepancies in cost between multiple-choice and performance-based items. An additional point of interest is that, according to the information distributions of the three kinds of items, the investigation items and the open-ended items provided more information for high ability examinees, but the multiple-choice items provided more information for average examinees.

Conventionally, test equating is based on common items or common examinees across different tests. In the example, the item parameters of both the open-ended and the investigation items were decomposed

into linear combinations of the item difficulties and rater severities. Common items and common raters then were used to link two forms. Therefore, all three forms were linked together but not in the same way. The item parameters of the three forms were estimated simultaneously rather than separately.

The information gains, however, may not suffice to justify the use of performance-based items because there are many other factors to be considered. The most critical issue for performance items—the variation in rater severity—was addressed here by introducing rater parameters. Although the interrater correlations were moderately high for the open-ended and investigation items, the range of rater severities was quite large. Hence, any procedure that did not take into account the variation of rater severities may be misleading.

The effects of single-mode definition of the latent variable were examined by a projection-like procedure. Assuming that the latent variable was defined by the performance-based mode, the multiple-choice mode was projected onto the latent variable and it was found that the measurement accuracy of the θ estimates improved significantly. However, the results of this approach did not differ from the approach in which the latent variable was defined jointly by both modes simultaneously. Similarly, if the latent variable was considered better defined by the multiple-choice items, then the performance-based items could be projected onto that latent variable, and the improvement of the measurement accuracy could be examined.

Conventional computer programs for item response modeling usually do not offer the flexibility to deal with different item types and rating tasks such as those investigated here. The newly developed RCMLM program provides users a great deal of flexibility to design customized models by manipulating the scoring vectors and the design matrix.

The results above indicate some possible directions for further research on the problem of combining scores from different assessment modes, and on the three specific issues investigated. Concerning rater effects, it would be useful to know whether more stringent and continuous evaluation would help keep between-rater variation low. CLAS conducted a series of field tests in the summer of 1994 to determine whether feedback of rater severity information was helpful to raters to, in part, address this issue (Wilson & Case, 1994). Knowledge about within-rater variation across items would make it easier to solve some technical problems in linking data from situations like that discussed here. There are other rater effects that are evident when the actual rating process is observed. For example, some raters tend to give more extreme scores and some give less extreme scores. Patterns such as these will not be detected by the models used here. Suitable models for the detection of such effects need to be developed.

Concerning information and SEs, a possibility that was not considered here is that nonlinear combinations of item responses may be used to combine the scores. CLAS has done this recently (California Learning Assessment System, 1994). A method of incorporating these nonlinearities into the item response model is needed.

Concerning the "projection-like" technique for the dimensional definition of the measured variable, the results above are tentative. A major effort in understanding the limitations and applications of the technique is needed, involving both mathematical work and extensive simulations. A multidimensional item response model which extends the RCMLM approach to multiple dimensions has been developed by Wang (1994).

References

Adams, R. A., & Wilson, M. (1992, April). *A random coefficients multinomial logit: Generalising Rasch models.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Adams, R. A., & Wilson, M. (in press). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. III). Norwood NJ: Ablex.

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

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

California Learning Assessment System. (1994). *1993/4 technical manual.* Sacramento CA: Author.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Dillon, W. R., & Mulani, N. (1984). A probabilistic latent class model for assessing inter-judge reliability. *Multivariate Behavioral Research, 19,* 438–458.

Draney, K., Pirolli, P., & Wilson, M. (in press). Using the RCML to investigate linear logistic test models in a complex domain. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. III). Norwood NJ: Ablex.

Dunbar, S. B., Koretz, D., & Hoover, H. D. (1991). Quality control in the development and use of performance assessment. *Applied Measurement in Education, 4,* 289–302.

Engelhard, G., Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5,* 171–191.

Grima, A., & Liang, J. (1992, April). *The effect of response rate to multiple-choice and open-ended items on differential item functioning.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Jöreskog, K. G., & Sörbom, D. (1988). *LISREL VII: A guide to program applications.* Chicago: SPSS, Inc.

Linacre, J. M. (1988). *FACETS: Computer program for many-faceted Rasch measurement.* Chicago: MESA Press.

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

Luecht, R., & Miller, T. (1992, April). *Multidimensional considerations for polychotomous item response models.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Lunz, M. E., & Stahl, J. A. (1990, April). *Severity of grading across time periods.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Lunz, M. E., Stahl, J. A., Wright, B. D., & Linacre, J. M. (1989, April). *Variation among examiners and protocols on oral examinations.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3,* 331–345.

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

Masters, G. N., & Wilson, M. (1991). *Partial credit models: Advanced session for statistical analysis and measurement staff.* Workshop at Educational Testing Service, Princeton NJ.

Moore, S. (in press). Estimating and testing differential item functioning with the RCML model. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. III). Norwood NJ: Ablex.

Overall, J. E., & Magee, K. N. (1992). Estimating individual rater reliabilities. *Applied Psychological Measurement, 16,* 77–85.

Pollack, J. M., Rock, D. A., & Jenkins, F. (1992, April). *Advantages and disadvantages of constructed-response item formats in large-scale surveys.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press. (Originally published 1960)

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement,* No. 17.

Samejima, F. (1977). The use of the information function in tailored testing. *Applied Psychological Measurement, 1,* 233–247.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Thissen, D. M. (1976). Information in wrong responses to Raven Progressive Matrices. *Journal of Educational Measurement, 13,* 201–214.

Van den Bergh, H., & Eiting, M. H. (1989). A method of estimating rater reliability. *Journal of Educational Measurement, 26,* 29–40.

Wainer, H., & Thissen, D. M. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6,* 103–118.

Wang, W. (1994). *Implementation and application of the multidimensional random coefficients model.* Unpublished doctoral dissertation, University of California, Berkeley.

Wang, W., & Wilson, M. (in press). Comparing multiple-choice and performance-based items. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. III). Norwood NJ: Ablex.

Wilson, M. (1992). The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement, 16,* 309–325.

Wilson, M. (1994). Community of judgment: A teacher-centered approach to educational accountability. In Office of Technology Assessment (Ed.), *Issues in educational accountability* (pp. 1–48). Washington D.C.: Office of Technology Assessment, United States Congress.

Wilson, M., & Adams, R. A. (in press). Evaluating progress with alternative assessments: A model for Chapter 1. In M. B. Kane (Ed.), *Implementing performance as-*

sessment: Promise, problems and challenge. Hillsdale NJ: Erlbaum.

Wilson, M., & Adams, R. A. (in press). Rasch models for item bundles. *Psychometrika.*

Wilson, M., & Case, M. (1994). *Dynamic rater calibration.* Paper presented at the meeting of the CLAS Technical Study Group, Sacramento CA.

Winer, B. J. (1962). *Statistical principles in experimental design.* New York: McGraw-Hill.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide*. Princeton NJ: Educational Testing Service.

Wright, B. D., Mead, R. J., & Bell, S. R. (1980). *BICAL: Calibrating items and scales with the Rasch model* (Research Memorandum No. 23C). Chicago: University of Chicago, Statistical Laboratory.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23–48.

Acknowledgments

This research was supported by grants from the National Science Foundation (RED-9255272), and from the California Learning Assessment System, California Department of Education. We thank the editor of this Special Issue, Fritz Drasgow, and two anonymous reviewers for numerous suggestions that have greatly improved the manuscript. Any errors or omissions are solely the responsibility of the authors.

Author's Address

Send requests for reprints or further information to Mark Wilson, Quantitative Methods in Education, Graduate School of Education, University of California, Berkeley CA 94720, U.S.A. Internet: Mark_Wilson@maillink.berkeley.edu.