A Comparison of Item- and Testlet-Level Scoring on Scale Stability in the Presence of Test Speededness

James A. Wollack
Craig S. Wells

and

Allan S. Cohen

University of Wisconsin
1025 W. Johnson St., #373
Madison, WI 53716

April 22, 2003

Portions of this paper were presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Running Head: ITEM VS. TESTLET SCORING

A Comparison of Item- and Testlet-Level Scoring on Scale Stability in the Presence of Test Speededness

Abstract

The present study was designed to examine the impact of test speededness on scale stability for a special class of items known as testlets. Previous work has demonstrated that speededness does have a negative effect on scale stability using dichotomously scored items (Wollack, Cohen & Wells, in press). This study extends the work of Wollack et al. to tests comprised of testlets. Results indicated that calibrating testlets using only nonspeeded examinees produced a more unidimensional scale with fewer drifting items and a more stable mean scaled score than was found using the total sample. The inclusion of speeded examinees was found to mask evidence of local item dependencies.

Index Terms: Partial credit model, item response theory, testlets, scale stability, item parameter drift

A Comparison of Item- and Testlet-Level Scoring on Scale Stability in the Presence of Test Speededness

Speededness effects arise when examinees modify their response strategies due to time limits for a test(Evans & Reilly, 1972). To the extent that examinees do not have sufficient time to complete a test, speededness effects will add an unwanted component to the construct being measured (Lord & Novick, 1968). This may result in poor estimation of ability for speeded examinees and poor estimation of item parameters, particularly for those items located at the end of the test (Douglas, Kim, Habing, & Gao, 1998; Oshima, 1994). Items at the end of speeded tests often appear harder than they would be on an unspeeded test. This occurs, because examinees often hurry through or even fail to respond to items at the end of the test (Bejar, 1985; Bolt, Cohen, & Wollack, 2002; Oshima, 1994). Bolt et al. (2002) used a mixture Rasch model (MRM; Rost, 1990) to classify examinees into latent speeded or nonspeeded groups, based upon the difference in performance on items at the beginning and end of speeded tests. Parameter estimates for end-of-test items using only responses from the nonspeeded group were found to be very similar to estimates for those same items when they were administered in nonspeeded locations on a different form of the test.

Wollack, Cohen, and Wells (in press) applied the Bolt et al. (2002) method to investigate the impact of speededness on scale stability. Eleven years worth of data from a college-level English Placement Test were analyzed. Items at the end of the test were found to be speeded. When these same items were moved to locations earlier in the test, their difficulties decreased. Furthermore, calibrating items using only nonspeeded examinees produced a much more stable and unidimensional score scale than was produced by including all examinees in the calibration.

This study extends the results of Wollack et al. (in press) to the case in which inter-item dependencies make testlet scoring appropriate. Wollack et al. scored all items with a dichotomous Rasch model, even though the reading comprehension section consisted of a number of reading passages each with 4 to 8 associated questions. It has been suggested that items of this type are better treated as testlets to control for the local dependency that typically exists among items for a common passage (Thissen, Steinberg, & Mooney, 1989; Wainer & Kiely, 1987; Wainer & Lewis, 1990; Yen, 1993).

In this paper, data from a college-level test of reading were analyzed under two separate models: the dichotomous Rasch model and the polytomous partial credit model (PCM; Masters, 1982). The PCM is a natural extension of the Rasch model that was used by Bolt et al. (2002) and Wollack et al. (in press). Use of polytomous models for testlets has been suggested as one means of handling tests composed of the testlets (Lee, Kolen, Frisbie, & Ankenmann, 2001; Wainer, Sireci, & Thissen, 1991). This paper compares the effects of using the PCM and the Rasch model on (a) the identification of speeded and nonspeeded examinees, (b) score scale stability, and (c) test unidimensionality for a reading comprehension test with sets of locally dependent items.

Modeling Test Speededness

Mixture Rasch Model Method

Bolt et al. (2002) and Wollack et al (in press) have modeled test speededness using a MRM (Rost, 1990) in which examinees are assumed to belong to one of two distinct classes, a speeded or nonspeeded class. Performance of individuals in the speeded class is assumed to be affected by time constraints, whereas that of examinees in the nonspeeded class is assumed to be

unaffected. The Rasch model is assumed to hold within each class; however, the item parameters are allowed to differ in the speeded and nonspeeded classes. In particular, performance on items early in the test is not thought to be hindered by any test speededness concerns that might arise later in the test. Therefore, in the Bolt et al. model of test speededness, item parameters are assumed to be equal in the speeded and nonspeeded groups for beginning-of-test items. In contrast, the end-of-test items are modeled as more difficult for the examinees in the speeded group. This pattern is modeled by constraining the item difficulty parameters for a set of items early in a test to be equal in the speeded and nonspeeded classes and constraining the difficulty parameters for a set of end-of-test items to be higher (i.e., more difficult) in the speeded class than in the nonspeeded class.

The MRM applied to test speededness is given by

$$P_i(U = 1|g, q_{jg}) = \frac{\exp(q_{jg} - b_{ig})}{1 + \exp(q_{jg} - b_{ig})},$$

where

- i indexes the item, i = 1, . . . , n,
- j indexes the examinees, j = 1, . . . , N,
- g indexes the latent class, g = 1, 2,
- 2_{jg} is the latent ability of examinee j in class g,
- b_{ig} is the Rasch item difficulty parameter for item i in class g, and
- $P_i(U = 1|g, 2_{jg})$ is the probability of a correct response to item i by examinee j in class g with ability 2_{jg}.

Mixture Partial Credit Model Method

When data are polytomous, such as the case on a test comprised of testlets, a mixture form of the partial credit model (Masters, 1982) may be more appropriate to use to identify classes of speeded and nonspeeded examinees in much the same way as the MRM was used. The mixture partial credit model (MPCM) is given by

$$P_i\left(U = k|g, q_{jg}\right) = \frac{\exp\left[\sum_{n=0}^{k}\left(q_{jg} - b_{ing}\right)\right]}{\sum_{c=0}^{m_i}\exp\left[\sum_{n=0}^{c}\left(q_{jg} - b_{ing}\right)\right]},$$

where

- k indexes the item category, k = 1, . . . , m_i,
- $b_{i<g}$ is the PCM item step parameter for category < of item i in class g,
- $P_i(U = k|g, 2_{jg})$ is the probability of examinee j in class g with ability 2_{jg} receiving a score of k on item i.

All other values are as previously defined. By definition, b_{i0g} equals 0 for both latent classes. Furthermore, item step parameters may be decomposed into an overall item difficulty, equal to the average of the k step parameters for item i, and a category-specific mean deviation:

$$b_{ikg} = \$_{ikg} + *_{ig},$$

where

$$d_{ig} = \frac{\sum_{n=1}^{k} b_{ing}}{k}$$

is the item's location on the ability scale and $\$_{ikg} = b_{ikg} - *_{ig}$ is the location of step k relative to the location of item i. It should also be noted that $G_{<} \$_{i<g} = 0$. In essence, this parameterization mirrors that of Andrich's rating scale model (1978a, 1978b, 1979), except that the $\$$s are allowed to differ for each item.

To identify latent speeded and nonspeeded classes of examinees, the model is constrained so that $b_{ik1} = b_{ik2}$ for items early in the test where speededness is not expected to be a factor. For items at the end of the test where speeded examinees are likely to fare less well than nonspeeded examinees, the model is constrained so that $*_{i1} > *_{i2}$. This constraint assures that the item is more difficult in class 1 (the speeded class) than in class 2 (the nonspeeded class). The $\$_{ikg}$, however, are freely estimated in the two latent classes for end-of-test items, meaning that the relative category locations may differ across classes.

Parameter Estimation

Parameters in the MRM and MPCM were estimated using a Markov chain Monte Carlo algorithm (MCMC; Gilks, Richardson, & Spiegelhalter, 1996; Patz & Junker, 1999a, 1999b) as implemented in the computer program WinBugs (Spiegelhalter, Thomas, & Best, 2000). Under MCMC, model parameters are estimated by repeatedly sampling each parameter from its posterior distribution, conditional on the data and the most recent estimates of all other parameters. After an initial burn-in period, it is possible to create a Markov chain such that the sampled values are drawn from the parameter's full conditional distribution. The value of each parameter is estimated as the mean of the Markov chain.

Sampling from posterior distributions requires the specification of prior distributions for all MCMC parameters. Bolt et al. (2002) and Wollack et al. (in press) provided prior distributions

for the MRM that allowed the Markov chains to converge to the parameters' full conditional distributions. These specifications are given in Table 1, along with the prior distributions used for the MPCM. In Table 1, $c_j = (1, 2)$ represents the class membership parameter for examinee j, and the B_g (g = 1, 2) are the mixing proportions, indicating the percentage of examinees in class g. All other parameters are as previously specified. Note that the only difference in priors between the MRM and MPCM involves the item difficulty parameters, b_{ig} for the MRM and $*_{ig}$ and $\$_{ikg}$ for the MPCM.

Insert Table 1 About Here

Research Design

Data

This study re-analyzed a four-year subset of the English Reading Comprehension Test (RCT) data used by Wollack et al. (in press). The RCT is a subtest on an 80-minute English Placement Test (EPT) which is administered annually to 15,000-18,000 students at a large midwestern university system. The test is used by advisors to help place entering undergraduate students into the initial composition sequence. The RCT is always the last section of the test. A different form of the test is published each year, though the test is designed to have some items overlap across years. For the four years used in this study, each form of the RCT had 10 reading passages with 52 or 53 operational (i.e., scored) items. The RCT subtest comprises approximately half the items on the EPT. The alpha coefficients for the operational portion of the RCT over that 4-year period ranged from .88 to .90.

In addition to the 10 operational reading passages, each RCT had three pilot forms, each containing one pilot reading passage and its associated items. The RCT pilot items are always located as the eleventh passage at the very end of the test, and so are the most susceptible to speededness effects. Poorly estimating item parameters on these pilot passages becomes a problem in subsequent years. Because the very end of the test is reserved for pilot passages and their associated items, any pilot passages that work well (or appear to work well) and are included on a future form of the RCT. These passages necessarily get moved out of the eleventh passage position reserved for pilot testing and into a less speeded portion of the test. For criterion-referenced tests such as a college placement test, this may be particularly problematic, since estimates of item difficulty which are systematically too high (i.e., too difficult) will result in a greater-than-expected number of students scoring above the criterion (Wollack et al., in press).

The particular four-year segment of data analyzed here, 1993-1996, is particularly well-suited for tracking the stability of items and score scales over time. The operational items in 1994, 1995, and 1996 consisted only of items that were administered, either operationally or as pilots, on the form from the previous year. Typically, the operational passages and their associated items remained in the same locations. The pattern of passage locations is given in Table 2. In 1994, seven passages (1, 2, 5, 7, 8, 9, and 10) were also administered in 1993 in those same locations. The third, fourth, and sixth passages in 1994 were all piloted in 1993 (and were all administered as the eleventh passage). In 1995, eight passages (1, 2, 3, 5, 6, 7, 8, and 10) were administered the previous year, 1994, in those same locations. Passages 4 and 9 on the 1995 test were both piloted in 1994. In 1996, eight passages (1, 2, 3, 4, 6, 7, 8, and 9) were

administered in 1995, though the locations changed for all but passage 2. Five of the seven passages that changed locations were in very nearly the same location, each appearing one passage later in 1996. Passages 1 and 3 in 1996 were located as passages 9 and 10 in 1995. The two 1996 passages which were not operational in 1995 (passages 5 and 10) were both 1995 pilots.

Insert Table 2 About Here

Identifying Nonspeeded Classes

For each form of the RCT between 1993 and 1996, both the MRM and the MPCM method described earlier were implemented to identify speeded and nonspeeded classes of examinees. We selected the PCM because it enabled us to implement a mixture model based on the PCM, which is a natural extension of the MRM used by Bolt et al. (2002).

To distinguish the speeded and nonspeeded classes, it is necessary to place certain constraints on the data. For the dichotomous MRM analysis, the Rasch difficulties, b_{ig}, associated with the first three reading passages (either 18 or 19 items) were constrained to be equal for both classes, while the difficulties for the items associated with the pilot passage (passage 11) were constrained to be larger (i.e., harder) for class 1 than for class 2. Therefore, class 1 defines the speeded class and class 2 defines the nonspeeded class. These were the same constraints imposed by Wollack et al. (in press).

The polytomous MPCM analysis was similar to that of the MRM. For the MPCM analysis, the item step parameters, b_{ikg}, were constrained to be equal in the two classes for the testlets associated with the first three reading passages. To distinguish the classes, difficulty parameters,

δ^{*}_{ig}, for the testlet associated with the pilot passage, were constrained to be larger for class 1 than for class 2.

Each form of the test was analyzed separately using both the MRM and MPCM. Although there were three pilot forms of the RCT each year, in 1994 and 1995 only two of the pilot passages became operational the following year. Speededness analyses were performed only for those forms that produced operational pilot items. Because 1997 data were not analyzed in this study, none of the1996 pilot passages were tracked. However, speeded and nonspeeded classes were estimated for one 1996 pilot form to allow for estimation of the scale stability, drift, and unidimensionality in the total group and in the nonspeeded class.

For each form of each year's test, a random sample of at least 1,500 examinees taking that form was analyzed for purposes of estimating model parameters. Class membership was estimated by fixing these parameter estimates and applying them to all examinees taking that particular form. This two-step process for each dataset has been shown to be an effective and efficient way to estimate class membership for all examinees (Cohen, Wollack, Bolt, & Mroch, 2002). All MCMC chains were run to a minimum of 6,500 iterations, with the first 500 comprising the burn-in and the mean of the remaining iterations taken as the parameter's estimate.

The prior distributions listed in Table 1 were used for the MCMC analyses. Random computer-generated starting values were used for all model parameters. An example of the WinBUGS syntax for performing the MPCM is shown in Appendix A.

Item Parameter Estimation

For each year, two sets of marginal maximum likelihood (Bock & Aitkin, 1981) estimates of Rasch difficulties and PCM item step parameters were obtained using MULTILOG (Thissen, 1991): One set of estimates was obtained using only those examinees who were classified as nonspeeded and one set was obtained using all examinees, regardless of class membership. Although the mixture analysis described above for identifying speeded and nonspeeded examinees involved only a subset of items and testlets from the RCT, item parameters for all items and testlets were estimated for each form using MULTILOG. The number of EM cycles was set to 1,000 for all runs. A total of 47 quadrature points ranging from -4.6 to 4.6 in increments of .2 were used to improve testlet parameter estimation in the PCM.

Detection of Item Parameter Drift and Equating

Partial Credit Model. Drift analyses were conducted first for the 93-94 years, then for the 93-94-95 years, and finally for the 93-94-95-96 years. This was done to represent the way that drift analyses are normally conducted within a testing program. That is, data are available initially only for the first two years, then for the first three years, and so on. In this study, PCM item step parameter estimates from each year were tested for parameter drift (Bock, Muraki, & Pfeiffenberger, 1988; Goldstein, 1983) from the base 1993 scale by concurrently calibrating all testlets and comparing parameter estimates from a compact model with estimates from an augmented model. The compact model imposed a series of equality constraints on common testlets. The augmented model relaxed those constraints to estimate the parameters for each testlet, one at a time.

The specifics of this analysis are as follows. The 93-94 drift analysis included data from 25 testlets: 10 operational testlets from 1993, 3 pilot testlets from 1993, 10 operational testlets from 1994, and 2 pilot testlets from 1994. As can be seen in Table 2, 10 testlets were common to 1993 and 1994. In the compact model, item step parameters for the 10 common testlets were constrained to be equal in the two years. In each of the 10 augmented models, the compact model was relaxed to let the PCM parameters be freely estimated for one of the common testlets.

The 93-94-95 drift analysis included data from 37 testlets: 10 operational testlets from 1995, 2 pilot testlets from 1995, and all 25 testlets from the 93-94 analysis. The 1995 compact model placed equality constraints on parameters for the 10 testlets common between 1994 and 1995. Also, equality constraints were imposed on all 93-94 testlets that were free of item parameter drift. Similarly, the 93-94-95-96 analysis included data from all 48 testlets: 10 operational testlets from 1996, 1 pilot testlet from 1996, and all 37 testlets from the 93-94-95 analysis. The 1996 compact model imposed equality constraints on all testlets common to 1995 and 1996, and between any pair of testlets found to be free of drift in either the 93-94 or 93-94-95 analysis. All other parameters were freely estimated. Each augmented model in the 1995 and 1996 analyses allowed one common testlet to be freely estimated.

The PCM drift analysis was performed twice, once using all examinees and once using only those examinees classified as nonspeeded. Constraints based on patterns of model parameter drift were imposed only within the group of examinees (nonspeeded or total) for which that pattern held.

Following the drift analysis, model step parameter estimates were equated to the 1993 scale. This was also done by concurrent calibration. For the 1994 estimation, parameters for all 1993

items were fixed at their scale values and equality constraints were imposed on all drift-free items. All drifting or 1994 pilot items were freely estimated. For the drift analyses including the 95 and 96 testing years, parameters for all previously estimated testlets were fixed at their scale values, equality constraints were imposed on all drift-free items, and all others were freely estimated.

Drift Criterion. Although our original intention was to test for item step parameter drift using the likelihood ratio test for differential item functioning (DIF; Kim, Cohen, DiStefano, and Kim, 1998; Thissen, Steinberg, and Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993) with $\alpha = .05$, because of the large number of examinees, this criterion proved to be neither practical nor feasible. Using an $\alpha = .05$ (or even an $\alpha = .10$ criterion) would have resulted in no drift-free testlets between 1993 and 1994 for the total group. Linking back to the 1993 scale requires having at least one common drift-free testlet. Therefore, a different criterion for inferring drift needed to be established. To address this concern, for each common testlet, we plotted testlet information curves (TIC) based on the sets of estimates from the augmented models, and calculated the average absolute difference (AAD) between the curves, as follows:

$$\mathbf{AAD} = \frac{\sum_{t=1}^{T} \left| \mathbf{I}_1\left(\theta_t\right) - \mathbf{I}_2\left(\theta_t\right) \right|}{\mathbf{T}},$$

where $I_1(\theta_t)$ and $I_2(\theta_t)$ are the amount of testlet information at θ_t (t = 1, . . . , T) for the two years. For this study, $| I_1(\theta_t) - I_2(\theta_t) |$ was observed at 91 evenly spaced θ-values ranging from -4.5 to 4.5. Based on visual inspection for dozens of testlets, a decision was made to regard AAD values less than or equal to .05 as negligible, and to treat all corresponding testlets as drift-free. To

illustrate this measure of similarity in item parameter estimates, plots of TICs are presented in Figure 1 for 10 sets of reference and focal groups and their corresponding AAD values. The plots in the left column all correspond to information curves that produced AAD values less than or equal to .05. The second column shows curves for which AAD was greater than .05.

Insert Figure 1 About Here

Rasch Model. Similar drift and linking procedures to those described above were also performed on the individual dichotomous Rasch items using the MRM nonspeeded and total groups. The drift analyses included a total of 138 items in 93-94, 203 items in 93-94-95, and 262 items in 93-94-95-96. In each of these three analyses, there were 53 items common to the last two years. As in the PCM analysis, equality constraints were also imposed between any pair of items which had previously been found to be drift free.

For the Rasch model analysis, it was not appropriate to use a .05 information criterion for identifying item parameter drift. Because Rasch items have only two categories and offer a maximum information of .25 (at $\theta = b$), average differences of .05 represent substantial deviations. Instead, items were tested for item parameter drift from the 1993 scale using the likelihood ratio test (Thissen, Steinberg, and Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993) with concurrent calibration. Items for which -2 times the log likelihood was greater than 3.84 (i.e., $\chi^2(1)$ at $\alpha = .05$) were identified as having drifted.

Creating Score Scales

As mentioned previously, the score scale was defined using operational 1993 data. For subsequent years, the estimation of θ was based on the most recent parameter estimates for each

operational testlet or item. Note that these values differed from the 1993 scale values for testlets/items whose parameter estimates were found to have drifted from the base scale values following its most recent administration.

Separate score scales were computed for items calibrated under the PCM and the Rasch model using parameter estimates either for all examinees or only for examinees in the nonspeeded class. Furthermore, within each of the four model (PCM vs. Rasch) × sample (total vs. nonspeeded) combinations, two separate score scales were created–one representing pre-equating and one representing post-equating. In pre-equating, the score scale is determined before the test is administered using data from previous administrations. Pre-equating is useful when it is important for examinees to know their scores shortly after testing, or when the testing window is sufficiently long that it is impractical for examinees to wait for the window to end before receiving scores. Examples of testing scenarios that often require pre-equating include computer adaptive testing, placement testing, and diagnostic testing (Wollack et al., in press). In contrast, for post-equating, the score scale is determined after all data are collected. When feasible, post-equating is preferable because it allows for re-estimation of any testlet or item parameters that may no longer be functioning the same as they once were.

A testing program with the ability to post-equate would appear to be less susceptible to speededness concerns than a program that relies on pre-equating (Wollack et al., in press). With post-equating, any items or testlets whose parameters have drifted would be re-estimated prior to publishing the score scale. In the case of speededness, any end-of-test items (such as the pilots) that become much easier as they get moved into earlier sections of the test will have their parameters re-estimated. In pre-equating, however, the scale is published using estimates of

item/testlet difficulty. When speededness is present, these estimates may be inappropriately high, given the new location of the item. This could result in scores that are systematically higher than expected.

Interestingly, Wollack et al. (in press) did not find the anticipated results with regard to pre- and post-equating. In particular, Wollack et al. found only a very slight tendency for root mean square errors to be smaller in post-equating situations. However, the post-equating biases were, in general, substantially larger than those for pre-equating.

RCT scores are reported to examinees on a metric from 150 to 850 with a mean of 500 and a standard deviation of 100. Therefore, for purposes of this study, linear transformation coefficients A and K were found such that

$$E(A\hat{q}_j + K) = 500 \text{ and } E[\,(A\hat{q}_j + K) - 500\,]^2 = 100^2$$

for the 1993 operational form of the test. Four separate sets of A and K coefficients were found, one each for the four model × sample combinations. The 1993 A and K coefficients from these four groups were used to transform RCT $\hat{q}$ scores for 1994 - 1996 onto the 1993 score-reporting metric.

Evaluative Measures

Assessing Local Item Dependence. To provide some insight into the extent to which testlet scoring was necessary for the RCT, and the degree to which we might expect to see differences between PCM and Rasch scoring, Yen's Q_3 (1984) statistic was used to measure the amount of local item dependence (LID) among all pairs of items. Q_3 is computed as the correlation of item residuals, after fitting the data with an IRT model. In theory, the Fisher r-to-z transformation of

Yen's Q_3 is distributed approximately normal (0, 1/(N - 3)). In practice, however, under the null hypothesis of local independence, Q_3 is known to have a negative bias. This is because the correlations between a pair of item residuals are contaminated by the inclusion of those items in estimating 2. This bias can be shown to be approximately -1/(n - 1) (Yen, 1993). It is clearly larger when n is smaller, e.g., such as when testlets are the unit of analysis as compared to individual dichotomous items

It was expected that Q_3 statistics based on Rasch model item and person parameter estimates would be slightly negative (on the order of -1/(n - 1)) for between-passage items, but would be positive (suggesting LID) for within-passage items. Furthermore, it was expected that Q_3 statistics between testlets based on PCM item step and person parameters would be slightly negative (again approximately equal to -1/(n - 1)).

Evaluating Scale Stability.

Three different methods were used to evaluate the stability of the score scales under the PCM and Rasch model for the nonspeeded examinees and all examinees. First, each form of the RCT was subjected to a principal components analysis (PCA), first using (a) only the nonspeeded examinees and then (b) all examinees. Two PCAs were conducted for each group of examinees, one based on all individual items and one based on testlet scores. For purposes of this analysis, only operational items were used. Therefore, the 1993 item-level PCA included 52 items, whereas the testlet-level PCA included 10 testlets. In 1994, 1995, and 1996, item- and testlet-level PCAs included 53 items and 10 testlets, respectively.

All PCAs were conducted using correlations that corrected item scores for continuity. That is, item-level PCAs were performed on tetrachoric correlations (as estimated by the computer

program TESTFACT (Wilson, Wood, & Gibbons, 1991)) and testlet-level PCAs were performed on polychoric correlations (as estimated by the computer program LISREL (Jöreskog, K. G. & Sörbom, D. (2002)). Matrices of these correlations were then input into SPSS (SPSS Inc., 2002) to complete the PCA.

The ratio of the first to second eigenvalue provides a measure of the extent to which the data are unidimensional (Wollack, et al., in press). This index was computed from the item- and testlet-level PCAs for both nonspeeded examinees and the total sample.

As a second measure of stability, we examined the mean and standard deviation of each form of the RCT for both groups of examinees, both types of IRT models, and both types of equating. In a stable scale, one would expect the mean and standard deviation to remain close to the 1993 values of 500 and 100, respectively. Two stability measures, the root mean square difference (RMSD) and bias, were calculated for each group on both the means and standard deviations, to quantify the amount of scale drift from the 1993 base values.

Finally, we compared the number of drifting items for each form of the RCT, for both groups of examinees and both models. In many regards, this provides the best measure of scale stability. Any differences in examinee ability level over time will result in changes in the mean scale value, but because drift analyses involve conditioning on $\hat{q}$, item parameter estimates can remain stable even if the overall ability level of examinees changes over time. Observing a small percentage of drifting items, therefore, provides evidence of scale stability. It was expected that the estimates based on responses for the total group would result in more drifting items than for those based on the nonspeeded group. Furthermore, it was expected that fewer items would drift when modeled with the PCM than with the Rasch model.

Results

Bolt et al. (2002) and Wollack et al. (in press), working with MRMs, and Wollack, Bolt, Cohen, and Lee (2002), working with the nominal response model (Bock, 1972), found that the MCMC chains converged very quickly to their stationary distributions (typically within the first 50 or 60 iterations). Figure 2 shows the sampling histories of 6,500 iterations of the Markov chains for several model parameters associated with one form of the 1993 RCT. Histories are provided for the latent class ability means, $:_1$ and $:_2$, latent class mixing proportions, B_1 and B_2, latent class item difficulties, $*_1$ and $*_2$, for the pilot testlet, and the latent class step deviation parameters, $\$_1$ and $\$_2$ for the pilot testlet. From Figure 2, one can see that, in general, model parameters here converged to their stationary distributions relatively quickly. However, there was a good deal of volatility in the estimates of the standardized step parameters for class 1 (the speeded class).

Insert Figure 2 About Here

Chains were run for a minimum of 6,500 iterations. The first 500 iterations were discarded as the burn-in. Average values across the remainder of the Markov chain were taken as the estimates for each parameter. Although estimates of the :s, *s and $s varied considerably by form and by item, the mixing proportions, B, were fairly stable, ranging between .2 and .3 for all datasets.

Wollack et al. (in press) emphasized the importance of checking the solution from a mixture model analysis to see that it makes sense. The speededness algorithm used here forces a two-class solution, regardless of the actual latent structure in the data. Because the mixture models

analyze only a subset of test questions (items associated with the first three passages and the last passage), Wollack et al. suggested comparing the characteristics of items used in the mixture analysis with those not used in the mixture analysis. In particular, they recommended examining the proportion correct scores for all items for examinees in both the speeded and nonspeeded classes, to see if the structure is consistent with the presence of test speededness.

Table 3 presents the proportion correct and average testlet score data for a form of the 1993 test for three categories of items/testlets: (a) those for which equality constraints were imposed in the mixture analysis, (b) those which were not included in the mixture analysis, and (c) those for which ordinal constraints were imposed in the mixture analysis. From Table 3, one can see that the items/testlets in category (a) have very similar statistics across the two classes. Also, the statistics associated with category (c) reveal that the speeded group performed dramatically less well at the end of the test. These results were anticipated, due to the nature of the mixture constraints. The statistics for category (b), however, are interesting. These items were excluded from the mixture analysis for two reasons. First, they were administered late enough in the test that we were not confident that equality constraints were appropriate. Second, they were not administered late enough in the test that we were comfortable imposing equality constraints. However, even though these items were not included in the analysis, it is still likely that speededness becomes an increasing concern throughout the test. Therefore, we expect to see the difference in difficulty between the nonspeeded and speeded classes gradually increase across this set of items. This is precisely the pattern that was observed in this study for both the MRM and MPCM.

Insert Table 3 About Here

Assessment of Local Item Dependence

Yen's (1984) Q_3 statistics were computed between all pairs of items administered in 1993. Because each pilot form contained a different reading passage and associated items, Q_3 statistics could not be computed between pilot items from different forms. To distinguish the separate forms, each form is numbered with a three digit number. The first two digits represent the testing year (and a common set of operational items), and the third digit is used to identify the particular pilot passage. Q_3 statistics computed between pilot items and operational items in this study were based on 2,525 examinees for the form 934 pilots, 3,453 examinees for the form 935 pilots, and 2,541 examinees for the form 936 pilots. Q_3 statistics between operational items were based on all 8,519 examinees completing a reading pilot form.

This analysis resulted in Q_3 statistics for 2,250 pairs of items. To facilitate interpretation, all Q_3 statistics between items associated with the same reading passage were averaged together to provide a measure of the average within-passage LID. As an example, for reading passage 1, which contained four items, Q_3 statistics between item pairs 1-2, 1-3, 1-4, 2-3, 2-4, and 3-4 were averaged to provide a within-passage 1 Q_3. Also, all Q_3 statistics between items from the same two passages were averaged together to provide a measure of the average between-passage LID. The four passage 1 items and the five passage 10 items, for example, produced $4 \times 5 = 20$ item pairs, each of which produced a Q_3 statistic. The average across these 20 statistics provided an average between-passage Q_3 for passages 1 and 10.

The average within- and between-passage Q_3 statistics based on Rasch model parameter estimates are provided in Table 4. To help interpret the structure in these data, the lower half-matrix of average Q_3 statistics is divided into five sections, labeled A, B, C, D, and E, as shown in the legend. The triangle labeled A contains the average between-passage Q_3 statistics between passages 1 through 9. The triangle labeled B contains the between-passage Q_3 statistics between passage 10 and the three pilot passages, numbered 11, 12, and 13 (although all were administered as the 11th passage on their respective forms). The rectangle labeled C contains the between-passage Q_3 statistics between one passage from among the first nine, and one passage from among the last four. The main diagonal section labeled D contains the within-passage Q_3 statistics for passages 1 through 9. The main diagonal section labeled E contains the within-passage Q_3 statistics for passages 10 through 13. The average Q_3 statistics in each of the five sections are given at the bottom.

Insert Table 4 About Here

Inspection of Table 4 reveals an interesting pattern. Q_3 statistics in section A of the table revealed a very slight negative bias, consistent with what one would expect between passages. Q_3 statistics in section C were also negative, but noticeably larger (in absolute value). Q_3 values in section B were all slightly-to-moderately positive, suggesting some degree of LID among between-passage items for end-of-test passages. Among the within-passage Q_3 statistics, those in section D were positive, but the magnitude was sufficiently small to suggest that LID may not be a serious concern, even for items within passages. In section E, however, Q_3 statistics were very large. Taken collectively, this pattern is consistent with what one would expect from a speeded

test, because speededness can result in LID among affected items. Here, between-passage items early in the test showed no LID, whereas between-passage items late in the test did show some LID. Also, within-passage items late in the test showed a great deal of LID. The between- and within-passage Q_3 statistics began to become substantial around passage 10, suggesting that the last two passages (10 and the pilot 11th passage) may be speeded. Interestingly, within-passage items early in the test did not show much LID.

To better examine the impact of within-passage items on LID, the Q_3 analysis was repeated using only nonspeeded examinees from the 1993 MRM analysis. The results of this analysis are presented in Table 5 and present a very different picture than the results from Table 4. As expected, removal of speeded examinees caused the Q_3 statistics in sections A, B, and C to become very similar. Q_3 values ranged from !.03 to .00, and were all very close to -1/(n - 1) = !.018, the expected Q_3 value for locally independent items. The results of the Q_3 analysis differed for the nonspeeded and total samples with respect to their within-passage values (i.e., sections D and E). When nonspeeded examinees only were analyzed, all the values along the main diagonal (except that for passage 1) indicated considerable LID. When all examinees were used, only the end-of-test items showed LID. The pattern in Table 5 is more consistent with our expectations of item behavior when administering multiple testlets of related items. Further, it suggests that the test conforms to a pattern that is better fitted with a PCM than a Rasch model, i.e., because of the higher LID within each testlet. This pattern in Table 4 was obscured by the presence of examinees for whom the test was overly speeded.

Insert Table 5 About Here

Table 6 shows the pattern of Q_3 statistics obtained by using all examinees, but estimating expected item scores using Rasch item parameter estimates based on the nonspeeded group only. As one can see from Table 6, average Q_3 statistics in sections A, B, and C are quite similar to their values in Table 4. The statistics gradually increase in section C and are positive in section B, reflecting the presence of test speededness. As a result of using item parameter estimates based on the nonspeeded class, however, the average Q_3 statistics in sections D and E were larger and positive, indicating that the within-passage items were not locally independent.

Insert Table 6 About Here

The above analysis was also performed on testlets using the PCM to compute expected testlet scores (i.e., testlet true scores). Within-passage comparisons were not possible for testlets. Tests with 11 testlets, have an expected Q_3 equal to !.10. Looking at the off-diagonal elements, the patterns were identical to those observed in Tables 4 to 6. The magnitudes of Q_3 statistics were higher, but were again similar to their expected values. As an example, in the PCM parallel of Table 5, where only nonspeeded examinees were analyzed, average Q_3 statistics in sections A, B, and C were !.089, !.050, and !.096, respectively.

Principal Components Analysis

The results of the PCA are given in Table 7. The top half of Table 7 shows the eigenvalues from a PCA on operational testlets using either all examinees or only the nonspeeded examinees from the MPCM analyses. The bottom half of Table 7 shows the eigenvalues from a PCA on individual operational items using either all examinees or only the nonspeeded examinees from the MRM analyses.

There was little difference between the magnitude of θ_1 between the total and nonspeeded groups. There was, however, a large difference in θ_2 and in the ratios of θ_1 to θ_2. In both the PCM and Rasch model, the nonspeeded group produced smaller values of θ_2 and larger eigenvalue ratios. This suggests that the nonspeeded group is more homogeneous than the entire group.

Interpreting differences between item-based and testlet-based PCAs is difficult because the number and nature of variables is so different in the two analyses. As expected, there was a large difference between the testlet-based and item-based PCAs with respect to the percentage of variance explained by the first two principal components. In the PCM analysis, where testlets were analyzed, the first component explained, on average, nearly half the variance in scores. In the individual item Rasch model analyses, the first component explained only one quarter of the variance in scores.

Score Scale Stability

Means and standard deviations across the four year period are presented in Table 8, along with RMSD and bias statistics. Note that in Table 8, the statistics listed for the nonspeeded group were found by estimating item/testlet parameters using only the nonspeeded group, and then applying these estimates to all examinees. The most discernable pattern from Table 8 is that the means were more stable when based on nonspeeded item/testlet parameter estimates. In addition, pre-equating using the total sample consistently provided the least stable RCT scores. Otherwise, clear patterns do not appear to be present. There was no meaningful difference between choice of model, nor between type of equating with respect to the stability of RCT scores.

Insert Table 8 About Here

Even though model parameter estimates were transformed back to the original 1993 scale, the differences between methods, or lack thereof, still may be obscured somewhat by differences in ability across the four years. The expectation that means and standard deviations remain constant across time rests in the reasonableness of the assumption that the samples are drawn from the same population. In fact, there was evidence that the samples here were not of equal ability. As an ad hoc analysis, we simultaneously calibrated all 40 operational testlets using the PCM, placing equality constraints on all common, drift-free items for purposes of linking. Average 2 estimates for 1993, 1994, 1995, and 1996 samples were 0.0, 0.02, 0.11, and -0.03, respectively. Of course, one limitation to this method is that estimates of 2 differ depending on the model, sample, set of linking items, and set of calibrating items used. Unfortunately, there is no empirical way to judge which set of 2 estimates is best. As an example, repeating the analysis above using only the nonspeeded examinees (and linking with the corresponding set of drift-free testlets), average values of $\hat{q}$ were 0.0, 0.01, 0.05, and -0.17 for years 1993 - 1996, respectively. Although the two sets of ability estimates differ somewhat, both show that the four years are not uniform with respect to ability.

Item/Testlet Parameter Drift

Results of the drift analysis are shown in Table 9, where items are classified into one of four categories, based on the magnitude of drift. In the upper half of the table, for which the PCM was applied, estimates of testlet parameters for two adjoining years are classified as either

similar, moderately similar, or dissimilar based on the difference between TIFs for the two years in an augmented model. To be classified as similar, TIFs had to differ by an average of less than or equal to .05 across all 2 values. Note that this was the criterion for identifying drift-free items for the PCM. Differences between .05 and .10 were classified as moderately similar, and differences greater than .10 were classified as dissimilar. Also, the number of items which were drift-free at $"$ = .05 according to the likelihood ratio test are recorded in the LR column.

Insert Table 9 About Here

Results of the Rasch drift analysis are presented in the bottom half of Table 9. Item parameter estimates were categorized as similar if they differed by no more than .15, moderately similar if they differed by .15 to .30, and dissimilar if they differed by more than .30. Again, the number of items yielding nonsignificant likelihood ratio tests are recorded in the LR column. For the Rasch analysis, the likelihood ratio test was used to identify drifting and nondrifting items.

Item parameter estimates were far more stable over time within the nonspeeded group than within the total group. As an example, within the nonspeeded group, an average of 6 testlets and 27 items per year were drift-free for the PCM and Rasch model, respectively, whereas within the total group, an average of 2.7 testlets and 10.3 items were drift free. Furthermore, for the Rasch analysis, calibrating on the basis of the nonspeeded group resulted in half the number of dissimilar items (an average of 8) when compared to the total group (an average of 15.7). A similar pattern for the PCM was not observed.

Comparisons across model type are difficult because the types of items and the criteria for classification were different. As mentioned previously, using the likelihood ratio criterion

proved unworkable for the PCM, because for the total group in 1993, zero drift-free testlets were identified. However, using the criteria developed for this study, there was little difference between the PCM and Rasch model with respect to item stability. An average of 60% of the testlets (and an average of 31.3 score points per year) in the PCM remained stable for the nonspeeded group compared to 51% of the items (and an average of 27 score points) in the Rasch model.

Comparison of Classifications

As a final comparison, we examined the similarity of proportions of examinees classified into speeded or nonspeeded groups by each of the two models for each of the four years. These data are provided in Table 10. Across the four years, the percentage of students classified in the same groups varied from 93.8% (1996) - 96.2% (1995). Overall, 94.9% of the students received the same classification. The models did not differ meaningfully with respect to the proportion of examinees classified as speeded. A total of 2.9% of the examinees were classified as speeded under the MPCM, but nonspeeded under the MRM, compared to 2.6% of the examinees who were classified as speeded under the MRM, but nonspeeded under the MPCM. Overall, roughly 22% of the examinees were classified as speeded (with the remaining 78% being classified as nonspeeded), regardless of model.

Insert Table 10 About Here

The correlations between the latent group membership variables ranged from .87 (1993) - .95 (1994), again indicating that the classifications under the two models were very similar.

Discussion

Wollack et al. (in press) demonstrated that calibrating a test using only nonspeeded examinees from a MRM results in a more unidimensional and stable scale across time. The purpose of this paper was to compare the effectiveness of a MPCM to identify test speededness with that of a MRM speededness algorithm for stabilizing a score scale on a reading comprehension test comprised of testlets. This study looked at the RCT over a four year period.

The results of this study largely replicated those of Wollack et al. (in press), in that calibrating using responses only from examinees in the nonspeeded class greatly improved the scale integrity: The scale was more unidimensional, showed better stability of average standard scores, and resulted in more drift-free items on which to equate forms. In fact, the criteria we used to define a drifting PCM item had to be altered because the likelihood ratio test failed to identify a single nondrifting testlet between 1993 and 1994, when the total sample was used. Overall, using a likelihood ratio test for drift criterion, nonspeeded samples identified five-and-a-half times more drift-free testlets using the PCM than did total samples. Using the $| I_d | \# .05$ criterion, nonspeeded samples identified two-and-a-quarter times the number of drift-free testlets for the total group. Nonspeeded samples with the Rasch model identified over two-and-a-half times more drift-free items than did total samples. There were no noticeable differences between the PCM and Rasch model with regards to scale stability measures.

Interestingly, few meaningful differences were observed between fitting the PCM and the Rasch model to the data. Patterns of unidimensionality, stability in average RCT scores, and parameter drift were very similar for both models. Also, classifications of examinees into speeded or nonspeeded classifications were identical in the Rasch model and PCM 95% of the

time. Nevertheless, the PCM was still the more appropriate model for the data. The RCT is comprised of testlets, and as such, the within testlet items violated the local independence assumption necessary to use the Rasch model.

The results of the LID analyses were interesting and emphasize the importance of removing test speededness effects from the data before subjecting it to analysis. When the Rasch model was fitted to all the data and Q_3 statistics were computed, the pattern of statistics (see Table 4) appeared to be counter-intuitive. Q_3 statistics for between-passage items were generally small, suggesting that the items were essentially locally independent. The magnitude of the Q_3 statistics increased with passage number, and the passages at the very end of the test had moderate, positive Q_3 statistics. This pattern is consistent with what one would expect from a speeded test where examinees responses to end-of-test questions were hurried and may not have accurately reflected their ability levels. The most noteworthy aspect of Table 4 is that the within-passage Q_3 statistics were all very small. It would have been reasonable to have stopped at this point and conclude that all the LID was caused by test speededness and that, in spite of the testlet nature of the RCT, the items appear sufficiently locally independent to allow use of the Rasch model if the speededness effects were first removed. However, this conclusion would have been made in error. In fact, when one removes the speeded examinees from the data, the structure of the remaining dataset is markedly different. The pattern among only the nonspeeded examinees fitted with the Rasch model (Table 5) again showed no LID among between-passage items, but this time showed substantial within-passage LID. This pattern was still evident when parameter estimates from nonspeeded examinees were used to compute $\hat{q}$ and $P_i(\hat{q})$ for all examinees.

This demonstrates an important reason for purifying the dataset of speeded examinees prior to

performing calibration work: The inclusion of speeded examinees presents a source of LID that obscures the real pattern among the items, thereby making it difficult to select an appropriate model.

The results of the Q_3 analysis were also interesting in that they appear to shed light on the point in the test where test speededness began to become a problem. By examining Table 4, one can see that all passages prior to passage 10 exhibit the same pattern: low negative between-passage statistics and low positive within-passage statistics. Beginning with passage 10, this pattern changed. Q_3 statistics in section C between one of the first nine passages and one of the last two passages became bigger, and Q_3 statistics in section B between passages 10 and 11 through 13 became positive and substantial. Also, Q_3 statistics within passages 10 - 13 were all large and positive. Although in this study ordinal constraints were placed upon only the pilot testlet, in fact it appears as though performance on the final two testlets was affected by test speededness.

It was also interesting that the type of equating did not seem to affect score scale stability. In particular, it had been expected that pre-equating on the basis of the entire sample would result in a systematically increasing RCT mean score because all items were systematically harder in their pilot locations than they were when administered in operational locations. However, no discernable pattern was observed with respect to type of equating. This is consistent with Wollack et al. (in press) who also found no meaningful differences between pre- and post-equating.

Wollack et al. (in press) studied drift over 11 years, whereas this study looked at only 4 years. Although studying drift over a longer time is desirable, it comes with certain drawbacks.

The amount of similarity in test items across years was much higher in this study. The four year period used here was selected because 100 percent of the operational items were common to the test from the previous year (either as operational or pilot items), and very often in the same locations. In Wollack et al., the amount of similarity between RCT forms in adjacent years was as little as one testlet with five items. Because the number of common items between adjacent forms was potentially so small, Wollack et al. defined "common item" as any item that had been previously administered, even if it wasn't administered on the immediately preceding form. This increased the number of "common items," and was the only reasonable way to equate the 11 forms used. It may not be the most appropriate way to view drift, however.

Another difference between Wollack et al. (in press) and this study is the way items were equated and tested for scale drift. Wollack et al. used iterative linking (Candell & Drasgow, 1988) through the test characteristic curve method (Stocking & Lord, 1983), and Lord's chi-square (Lord, 1980) to identify drifting items. The Wollack et al. drift analyses were performed as a chain of 11 independent analyses. As they mentioned in their discussion, ideally, they would have included all data into a single analysis in which all model parameters and all drift analyses were simultaneously estimated. This simultaneous solution would have the advantage of minimizing estimation errors by including information from the entire variance-covariance item matrix. Unfortunately, with 11 years of data, this solution was too large. However, this ideal design was precisely the one used in this study.

The results of this study are important because they demonstrate that it is difficult to hold together a scale for tests which are, to some degree, speeded. The solution is to purify the calibration dataset by systematically removing all examinees classified into a latent speeded

class. The rationale is that item behavior among the examinees in the nonspeeded class was more stable over time than item behavior for all examinees. Furthermore, item behavior for end-of-test items in the nonspeeded class was a better predictor of how the same items would perform for the entire group when moved into earlier positions in the test.

Purifying datasets of speededness is particularly important for tests that use end-of-test locations for piloting items or for tests on which an item's location changes across forms. Locations at the end of the test tend to make items appear more difficult than locations at the beginning or middle of the test. Item location is normally a nuisance variable, contributing an unintended source of variability. The speededness model presented here provides one way to address and reduce location effects (for another speededness model, see Yamamoto & Everson (1997)). As shown in this study, as well as those by Bolt et al. (2002) and Wollack et al. (in press), an item's behavior is much more stable within the nonspeeded group, regardless of that item's location, than it is within the total group. Scale purification will result in more drift-free items on which to equate, a more stable score scale, and ultimately, a more interpretable score.

This study implemented a MPCM because it is the natural extension of the MRM used by Bolt et al. (2002) and Wollack et al. (in press). There exist several other item response models which could have been used to analyze the data in this study. Examples of item response models available for scoring polytomous item responses include the PCM (Masters, 1982), the graded response model (Samejima, 1969), and the nominal response model (Bock, 1972). The testlet model (Bradlow, Wainer, & Wang, 1999; Li & Cohen, 2003; Wainer, Bradlow, & Du, 2000) is also available for accounting for local dependence among sets of dichotomously scored items. The effectiveness of these models for use in identifying test speededness is an area that should be studied further.

It is important to mention that test speededness effects, as modeled in this study, are confounded with fatigue effects (Wollack et al., in press). That is, it is impossible to know whether examinees classified as speeded actually had insufficient time, or if their diminished performance was attributable to fatigue. This distinction is important because, while increasing testing time can reduce speededness effects, it will also serve to increase fatigue effects.

In addition, speededness effects in this study were confounded with content effects. Because ordinal constraints were applied only to the last testlet, it is possible that something relating to the content of the final passage, rather than insufficient time, contributed to the poor performance of examinees in the speeded class on the final testlet. Had multiple testlets been used to distinguish the classes, it would have been possible to tease apart those suffering from content effects associated with the pilot passage and those suffering from speededness or fatigue.

Finally, although the testlets used in this study were composed of sets of individually-scored items associated with a common reading passage, the results of this study should generalize to other testing situations for which the PCM is appropriate, such as for scoring constructed response items or performance measures.

References

Andrich, D. (1978a). Scaling attitude items constructed and scored in the Lickert tradition. Educational and Psychological Measurement, 38, 665-680.

Andrich, D. (1978b). Applications of a psychometric rating model to ordered categories which are scored with successive integers. Applied Psychological Measurement, 2, 581-594.

Andrich, D. (1979). A model for contingency tables having ordered response classification. Biometrics, 35, 403-415.

Bejar, I. I. (1985). Test speededness under number-right scoring: An analysis of the Test of English as a Foreign Language (Report No. ETS-RR-85-11). Princeton, NJ: Educational Testing Service.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 46, 443-459.

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

Bock, R., Muraki, E. & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. Journal of Educational Measurement, 25, 275-285.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Applications of a mixture Rasch model with ordinal constraints. Journal of Educational Measurement, 39, 331-348.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. Psychometrika, 64, 513-168.

Candell, G. L, & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. Applied Psychological Measurement, 12, 253-260.

Cohen, A. S., Wollack, J. A., Bolt, D. M., & Mroch, A. A. (2002). A mixture Rasch model analysis of test speededness. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Douglas, J., Kim, H. R., Habing, B., & Gao, F. (1998). Investigating local dependence with conditional covariance functions. Journal of Educational and Behavioral Statistics, 23, 129-151.

Evans, F. R., & Reilly, R. R. (1972). A study of speededness as a source of test bias. Journal of Educational Measurement, 37, 123-131.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J., Eds. (1996). Markov chain Monte Carlo in practice. London: Chapman & Hall.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. Journal of Educational Measurement, 20, 369-377.

Jöreskog, K. G., & Sörbom, D. (2002). LISREL 8.5.3. [Computer program]. Chicago, IL: Scientific Software, Inc.

Kim, S-H., Cohen, A. S., DiStefano, C. A., & Kim, S. (April, 1998). An investigation of the likelihood ratio test for detection of differential item functioning under the partial credit model. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Li, Y., & Cohen, A. S. (2003, April). Equating tests composed of testlets: A comparison of a testlet response model and four polytomous response models. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

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

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Addison-Wesley.

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

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. Journal of Educational Measurement, 31, 200-219.

Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. Journal of Educational and Behavioral Statistics, 24, 146-178.

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. Journal of Educational and Behavioral Statistics, 24, 342-366.

Rost, J. (1990) Rasch models in latent classes: An integration of two approaches to item analysis. Applied Psychological Measurement, 14, 271-282.

Samejima, F. (1969). Estimation of a latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 17.

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (2000). WinBUGS version 1.3 [Computer program]. Robinson Way, Cambridge CB2 2SR, UK: Institute of Public Health, Medical Research Council Biostatistics Unit.

SPSS Inc. (2002). SPSS 11.5.0 [Computer program]. Chicago, IL.

Stocking, M., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 207-210.

Thissen, D. (1991). MULTILOG: Multiple, categorical item analysis and test scoring using item response theory [Computer program]. Chicago, IL: Scientific Software, Inc.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. Psychological Bulletin, 99, 118-128.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. Journal of Educational Measurement, 26, 247-260.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), Test Validity (pp. 147-169). Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), Differential Item Functioning (pp. 67-113). Hillsdale, NJ: Erlbaum.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An anolog for the 3-PL useful in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), Computerized Adaptive Testing: Theory and Practice (pp. 245-270). Boston, MA: Kluwer-Nijhoff.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement, 24, 185-201

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. Journal of Educational Measurement, 27, 1-14.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. Journal of Educational Measurement, 28, 197-219.

Wilson, D. T., Wood, R., & Gibbons, R. (1991). TESTFACT: Test scoring, item statistics, and item factor analysis [computer program]. Chicago, IL: Scientific Software, Inc.

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y. -S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. Applied Psychological Measurement, 26, 337-350.

Wollack, J. A., Cohen, A. S., & Wells, C. S. (in press). The effects of test speededness on score scale stability. Journal of Educational Measurement.

Yamamoto, K. & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.) Applications of latent Trait and Latent Class Models in the Social Sciences. New York: Waxmann.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8, 125-145.

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

Table 1

Prior Distributions for MRM and MPCM

MRM Priors	MPCM Priors
$b_{ig} \sim \text{Normal}(0, 1)$	$\theta_{ig} \sim \text{Normal}(0, 1)$
	$\gamma_{ikg} \sim \text{Normal}(\theta_{ig}, 1)$
$\zeta_{jg} \sim \text{Normal}(\xi_g, 1)$	$\zeta_{jg} \sim \text{Normal}(\xi_g, 1)$
$\xi_g \sim \text{Normal}(0, 1)$	$\xi_g \sim \text{Normal}(0, 1)$
$c_j \sim \text{Bernoulli}(\beta_1, \beta_2)$	$c_j \sim \text{Bernoulli}(\beta_1, \beta_2)$
$(\beta_1, \beta_2) \sim \text{Dirichlet}(0.5, 0.5)$	$(\beta_1, \beta_2) \sim \text{Dirichlet}(0.5, 0.5)$

Table 2

Tracking of Passage Locations for RCT

1993	1994	1995	1996
1	1	1	
2	2	2	2
3			
4			
5	5	5	6
6			
7	7	7	8
8	8	8	9
9	9		
10	10	10	3
11	4		
11	3	3	4
11	6	6	7
	11	9	1
	11	4	
		11	10
		11	5
			11

Table 3

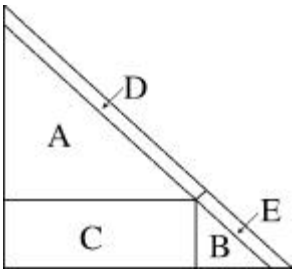
Proportion Correct for Speeded and Nonspeeded Classes

Item	Average Item Score: Nonspeeded Class	Average Item Score: Speeded Class		Item	Proportion Correct: Nonspeeded Class	Proportion Correct: Speeded Class
1	2.76	2.78	Items/Testlets with Equality Constraints on Item/Testlet Difficulty	1	.85	.84
2	4.25	4.30		2	.54	.54
3	3.31	3.31		3	.65	.62
				4	.72	.71
				5	.88	.90
				6	.72	.66
				7	.71	.69
				8	.69	.65
				9	.69	.68
				10	.60	.60
				11	.77	.78
				12	.45	.37
				13	.92	.91
				14	.64	.65
				15	.55	.55
4	3.92	3.70	Items/Testlets not Included in MRM and MPCM analyses	17	.69	.62
5	3.80	3.57		19	.51	.43
6	2.89	2.66		21	.73	.69
7	3.79	3.41		23	.80	.69
8	3.37	2.75		25	.80	.76
9	3.09	2.11		27	.58	.54
10	3.16	1.69		29	.54	.48
				31	.47	.41
				33	.44	.38
				35	.72	.64
				37	.71	.59
				39	.79	.66
				41	.63	.48
				43	.61	.40
				45	.60	.37
				47	.72	.48
				49	.55	.33
				51	.47	.23
11	4.31	1.06	Items/Testlets With Ordinal Constraints on Item/Testlet Difficulty	53	.82	.20
				54	.74	.21
				55	.88	.10
				56	.60	.07
				57	.70	.03
				58	.66	.02

Table 4

Average Between- and Within- Passage Q_3 Statistics–Rasch Total Group

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	.03												
2	.01	.04											
3	.00	-.01	.01										
4	.01	.00	.00	.02									
5	.01	.01	.00	.00	.03								
6	-.01	-.01	.00	-.01	-.01	.01							
7	-.01	-.01	-.02	-.01	-.01	-.01	.02						
8	-.03	-.03	-.03	-.03	-.02	-.03	-.01	.04					
9	-.03	-.04	-.03	-.03	-.03	-.03	-.02	.00	.03				
10	-.06	-.05	-.05	-.06	-.05	-.05	-.04	.00	.02	.11			
11	-.07	-.06	-.05	-.06	-.04	-.06	-.06	-.03	.00	.07	.21		
12	-.04	-.04	-.03	-.04	-.03	-.05	-.04	-.02	.00	.04	.	.10	
13	-.06	-.06	-.05	-.06	-.05	-.06	-.05	-.01	.02	.08	.	.	.16

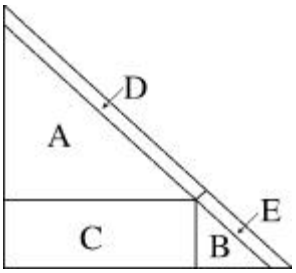


Average Between-Passage Q3		Average Within-Passage Q3	
A	-.014	D	.023
B	.065	E	.143
C	-.039		
Total	-.023	Total	.060

Table 5

Average Between- and Within- Passage Q_3 Statistics–Rasch Nonspeeded Group Only

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	.01												
2	.00	.18											
3	-.02	-.02	.20										
4	-.01	-.01	-.01	.17									
5	.00	.00	-.01	-.01	.21								
6	-.02	-.02	-.01	-.02	-.02	.20							
7	-.01	-.02	-.02	-.02	-.02	-.02	.17						
8	-.03	-.03	-.03	-.02	-.01	-.02	-.02	.21					
9	-.02	-.02	-.02	-.02	-.02	-.02	-.01	-.02	.20				
10	-.04	-.03	-.03	-.04	-.03	-.03	-.03	-.01	-.01	.24			
11	-.01	.00	-.01	-.01	-.02	-.02	-.02	-.02	-.02	-.01	.15		
12	.00	.00	.00	-.02	-.02	-.03	-.02	-.02	-.01	-.02	.	.17	
13	-.01	-.02	-.02	-.02	-.02	-.02	-.01	-.02	-.01	.00	.	.	.17

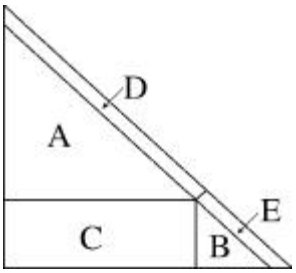


Average Between-Passage Q3		Average Within-Passage Q3	
A	-.017	D	.172
B	-.010	E	.183
C	-.019		
Total	-.017	Total	.175

Table 6

Average Between- and Within- Passage Q_3 Statistics–Rasch Total Group, Nonspeeded Class Statistics

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	.02												
2	.01	.19											
3	.00	-.01	.21										
4	.00	.00	.00	.18									
5	.01	.01	.00	.00	.22								
6	-.01	-.01	.00	-.01	-.02	.21							
7	-.01	-.01	-.02	-.01	-.01	-.02	.18						
8	-.03	-.03	-.03	-.03	-.02	-.02	-.02	.23					
9	-.03	-.04	-.03	-.03	-.03	-.03	-.02	.00	.22				
10	-.06	-.05	-.05	-.06	-.04	-.05	-.04	.00	.01	.28			
11	-.06	-.05	-.04	-.05	-.04	-.06	-.06	-.03	.00	.06	.34		
12	.04	-.04	-.03	-.04	-.03	-.05	-.04	-.02	.00	.03	.	.29	
13	-.06	-.05	-.05	-.06	-.05	-.06	-.04	-.01	.02	.06	.	.	.30



Average Between-Passage Q3		Average Within-Passage Q3	
A	-.014	D	.184
B	.050	E	.303
C	-.035		
Total	-.022	Total	.221

Table 7
Comparison of Eigenvalues for Total and Nonspeeded Groups in PCM and Rasch Model

Partial Credit Model

	Total Group			Nonspeeded Group		
Year	θ_1	θ_2	θ_1 / θ_2	θ_1	θ_2	θ_1 / θ_2
1993	4.67 (46.7%)	.926 (9.3%)	5.04	4.78 (47.8%)	.768 (7.7%)	6.22
1994	4.89 (48.9%)	.948 (9.5%)	5.16	4.93 (49.3%)	.732 (7.3%)	6.73
1995	4.81 (48.1%)	.861 (8.6%)	5.59	4.86 (48.6%)	.712 (7.1%)	6.83
1996	4.94 (49.4%)	.861 (8.6%)	5.74	5.06 (50.6%)	.713 (7.1%)	7.10

Rasch Model

	Total Group			Nonspeeded Group		
Year	θ_1	θ_2	θ_1 / θ_2	θ_1	θ_2	θ_1 / θ_2
1993	12.31 (23.7%)	2.31 (4.4%)	5.33	12.15 (23.4%)	1.58 (3.0%)	7.69
1994	13.75 (25.9%)	2.45 (4.6%)	5.61	13.26 (25.0%)	1.67 (3.1%)	7.94
1995	12.82 (24.2%)	2.11 (3.2%)	6.08	12.64 (23.9%)	1.95 (3.7%)	6.48
1996	13.89 (26.2%)	2.33 (4.4%)	5.96	13.85 (26.1%)	1.75 (3.3%)	7.91

Table 8

RCT Means and Standard Deviations, 1993 - 1996

			Mean						Standard Deviation					
			1993	1994	1995	1996	RMSD	Bias	1993	1994	1995	1996	RMSD	Bias
PCM	Pre-Equating	Total Group	500	527.4	525.1	508.1	22.0	20.2	100	109.6	110.3	105.6	8.7	8.5
		NS Group		508.4	504.9	484.1	10.8	-0.9		119.6	119.9	116.6	18.8	18.7
	Post-Equating	Total Group		506.3	517.0	494.4	11.0	5.9		119.5	116.6	98.6	14.8	11.6
		NS Group		498.3	503.6	476.9	13.5	-7.1		120.0	120.2	113.9	18.3	18.0
Rasch Model	Pre-Equating	Total Group	500	536.4	531.3	519.8	30.0	29.2	100	121.5	118.7	123.0	21.1	21.1
		NS Group		507.5	511.3	495.7	8.2	4.8		121.8	119.1	122.8	21.3	21.2
	Post-Equating	Total Group		509.4	517.0	507.0	11.9	11.1		122.0	119.6	122.6	21.4	21.4
		NS Group		500.5	504.7	488.9	7.0	-2.0		121.5	119.6	122.3	21.2	21.1

Table 9

Comparison of the number of nondrifting items

PCM

	Total Group				Nonspeeded Group			
Year	LR	$\lvert I_d \rvert \# .05$	$.05 < \lvert I_d \rvert \# .10$	$\lvert I_d \rvert > .10$	LR	$\lvert I_d \rvert \# .05$	$.05 < \lvert I_d \rvert \# .10$	$\lvert I_d \rvert > .10$
93-94	0	1	6	3	4	6	1	3
94-95	1	5	3	2	5	7	2	1
95-96	1	2	4	4	2	5	2	3

Rasch Model

	Total Group				Nonspeeded Group			
Year	LR	$\lvert d \rvert \# .15$	$.15 < \lvert d \rvert \# .30$	$\lvert d \rvert > .30$	LR	$\lvert d \rvert \# .15$	$.15 < \lvert d \rvert \# .30$	$\lvert d \rvert > .30$
93-94	6	10	21	21	29	31	13	8
94-95	16	20	23	10	37	41	8	3
95-96	9	11	26	16	15	25	15	13

Table 10

Partial Credit Model Versus Rasch Model Classifications and Group Membership Correlations

1993

$r_{gmem(R),\ gmem(PCM)} = .87$

PCM Classifications	Rasch Classifications: SP	NS	Total
SP	1638	199	1837
NS	303	6379	6682
Total	1941	6578	8519

1994

$r_{gmem(R),\ gmem(PCM)} = .95$

PCM Classifications	Rasch Classifications: SP	NS	Total
SP	1199	304	1503
NS	34	3963	3997
Total	1233	4267	5500

1995

$r_{gmem(R),\ gmem(PCM)} = .94$

PCM Classifications	Rasch Classificatons: SP	NS	Total
SP	1380	101	1481
NS	178	5615	5793
Total	1558	5716	7274

1996

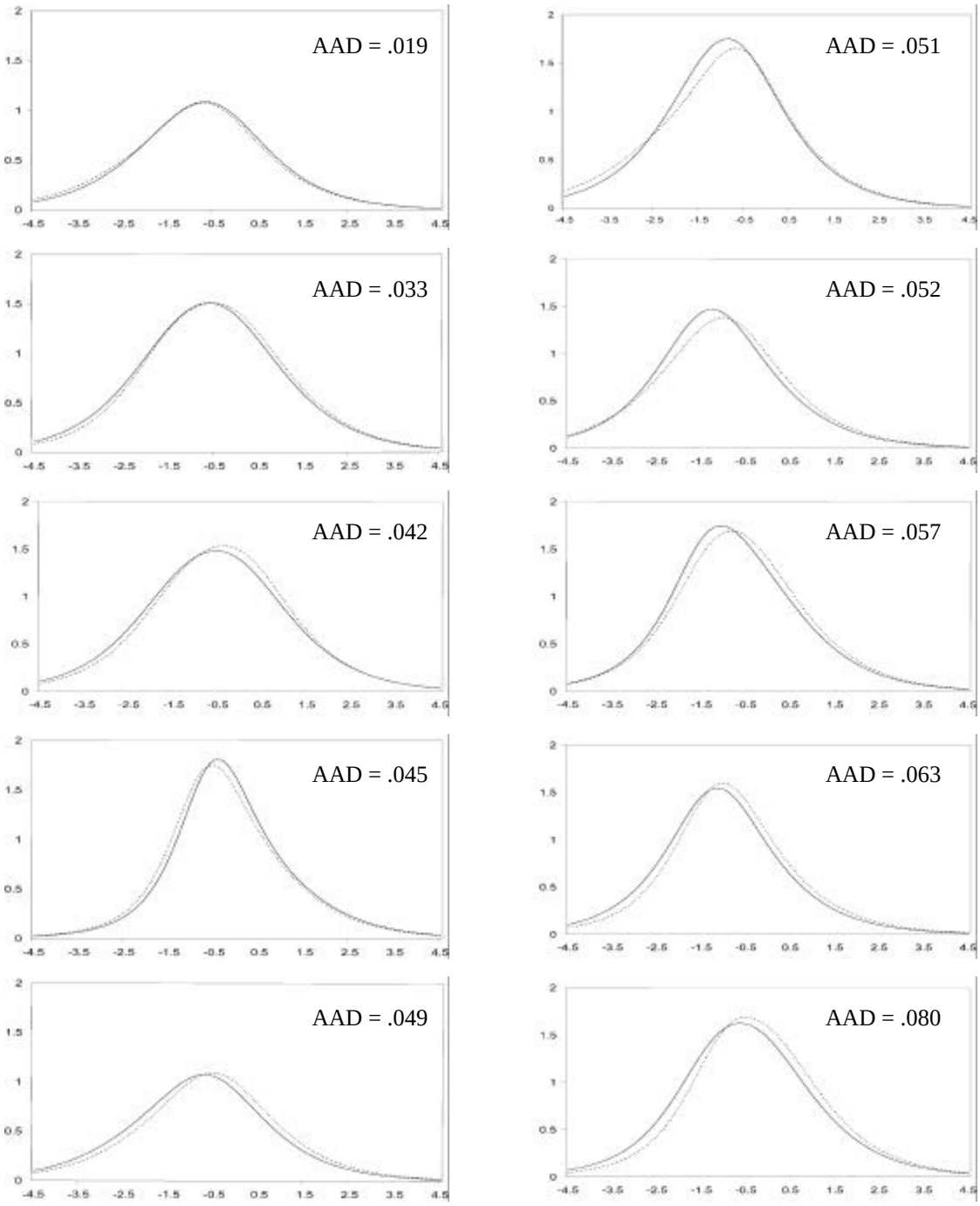
$r_{gmem(R),\ gmem(PCM)} = .91$

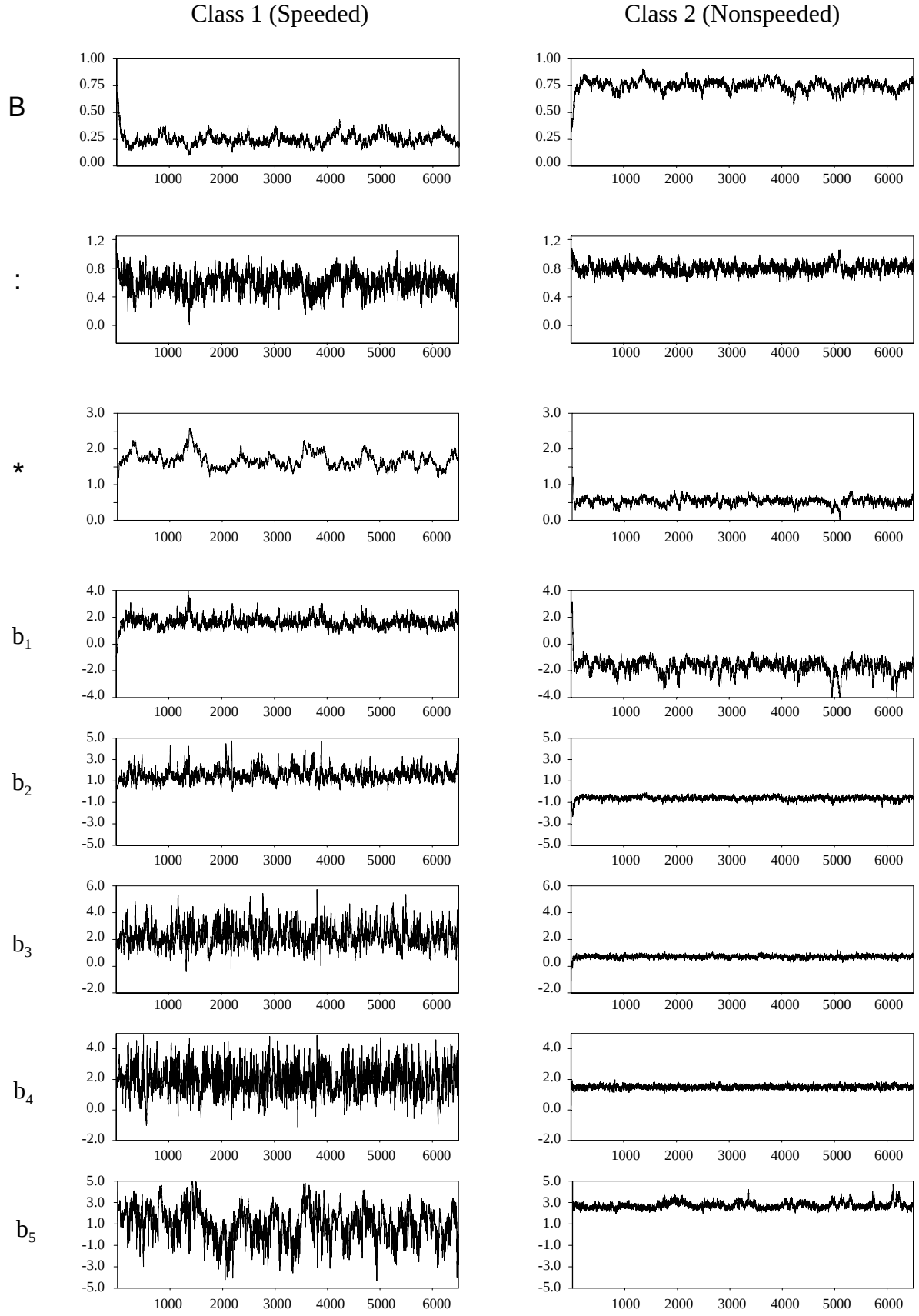
PCM Classifications	Rasch Classifications: SP	NS	Total
SP	702	133	835
NS	146	3517	3663
Total	848	3650	4498

Figure 1

Average Absolute Differences for Various Testlet Information Functions

AAD # .05 AAD > .05




Class 1 (Speeded)
Class 2 (Nonspeeded)
B
:
*
b1
b2
b3
b4
b5
1000
2000
3000
4000
5000
6000

Appendix A

```
 model
{
# Form 934--equality constraints on operational items

   for (j in 1:N)   {
      for (i in 1:T)      {
         r[j,i]<-resp[j,i]
   }}

   for (g in 1:G){
      alpht[g]<-alph[g]
   }

   for (j in 1:N)   {
      for (i in 1:T)   {
         denom[j,i,1] <- 1
         numer[j,i,1] <- 0
         enumer[j,i,1] <- 1
   }}

# Partial Credit Model
   for (j in 1:N)   {
      for (i in 1:T)   {
         for (k in 2:mI[i])  {
            numer[j,i,k] <- (theta[j]-step[gmem[j],i,k] + numer[j,i,k-1])
            enumer[j,i,k] <- exp(numer[j,i,k])
            denom[j,i,k] <- enumer[j,i,k] + denom[j,i,k-1]
   }
            denom2[j,i,1] <- denom[j,i,mI[i]]
   }}

   for (j in 1:N)   {
      for (i in 1:T)   {
         for (k in 1:mI[i])  {
            p[j,i,k] <- enumer[j,i,k]/denom2[j,i,1]
   }
         r[j,i]~dcat(p[j,i,1:mI[i]])
   }
      theta[j] ~ dnorm(mut[gmem[j]],1)
      gmem[j] ~ dcat(pi[1:G])
   }

 # Priors
 # Equality constraints
    for (i in 1:(T-1)){
      beta.pre[1,i]~dnorm(0.,1.)
   }

    for (i in 1:(T-1)){
         beta.pre[2,i]<-beta.pre[1,i]
   }

    beta.pre[1,T]~dnorm(0,1.)

    beta.pre[2,T]~dnorm(0,1.) I(,beta.pre[1,T])

      for (i in 1:T){
             for (g in 1:G){
                      beta[g,i]<-beta.pre[g,i]-mean(beta.pre[g,1:T])
   }}

            for (i in 1:T) {
               for (g in 1:G) {
                  b[g,i,1] <- 0
   }}
```

Appendix A, Cont'd.

```
   for (i in 1:T){
          for (k in 2:(mI[i]-1)) {
             b[1,i,k]~dnorm(beta[1,i],1.)
   }}

   for (i in 1:T){
             b[1,i,mI[i]]  <- (mI[i]-1)*(beta[1,i])-sum(b[1,i,2:(mI[i]-1)])
   }

   for (k in 2:(mI[T]-1)) {
      b[2,T,k]~dnorm(beta[2,T],1.)
   }

   b[2,T,mI[T]]  <- (mI[T]-1)*(beta[1,T])-sum(b[2,T,2:(mI[T]-1)])


   for (i in 1:(T-1)){
      for (k in 2:(mI[i])) {
         b[2,i,k] <- b[1,i,k]
   }}


   for (i in 1:T){
      for (k in 2:mI[i]){
         for (g in 1:G){
            b.std[g,i,k]<-b[g,i,k]-mean(b[g,i,2:mI[i]])
   }}}

   for (i in 1:T) {
      for (g in 1:G) {
         b.std[g,i,1] <- 0
   }}

   for (i in 1:T){
      for (k in 2:mI[i]){
         for (g in 1:G){
            step[g,i,k] <- b.std[g,i,k] + beta[g,i]
   }}}

   for (i in 1:T) {
      for (g in 1:G) {
         step[g,i,1] <- 0
   }}
      pi[1:2]~ ddirch(alpht[1:2])
      mut[1]~ dnorm(0.,1.)
      mut[2]~ dnorm(0.,1.)
   }

list(N=1500, T=4, G=2,alph=c(.5,.5),mI=c(6,7,6,7),
resp=structure(.Data=c(
3,5,5,7,
3,6,5,7,
.
.
.
3,5,2,2,
3,4,5,4), .Dim=c(1500,4)))
```