A Mixture Rasch Model Analysis of Test Speededness

Allan S. Cohen, James A. Wollack, Daniel M. Bolt and Andrew A. Mroch
University of Wisconsin–Madison

March 15, 2002

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, April, 2002

Running Head: Test Speededness

Abstract

Speededness effects arise when examinees do not have enough time to complete a test and may affect the performance of some examinees. One consequence of speededness effects is that item response theory model parameters may be inaccurately estimated. A mixture Rasch model was used in this study and appears to have resulted in removing at least some, if not all, speededness effects from item parameter estimates. Characteristics of examinees classed in the speeded and non-speeded groups were examined and found to be somewhat consistent with previous work on this model. Membership in a speeded or non-speeded class was associated with gender for the higher difficulty test but not for the lower difficulty one. Some differences were noted in ability between the latent groups: Examinees in the speeded class appeared to perform better on the test as a whole, albeit not on the speeded items, than examinees in the non-speeded group. In addition, these same examinees appeared to have slightly higher academic achievement and make faster progress toward the degree.

A Mixture Rasch Model Analysis of Test Speededness

Speededness effects result when examinees have insufficient time to complete a test (Evans & Reilly, 1972). Speededness effects may alter the test performance of some examinees and are generally a problem as speed is not usually an intended component of the construct being measured (Lord & Novick, 1968). In the context of item response theory (IRT), speededness effects can result in model parameters being inaccurately estimated, particularly for items located at or near the end of a test (Oshima, 1994). Bolt, Cohen and Wollack (in press b) described an estimation strategy using a mixture Rasch model (MRM: Rost, 1990) for reducing contamination due to test speededness on item parameter estimates on a college-level mathematics placement test. That strategy consisted of applying ordinal constraints to a MRM so as to distinguish two latent classes: (1) a "speeded" class of examinees for whom the time limits were not sufficient to adequately answer end-of-test items, and (2) a "non-speeded" class of examinees for whom the time limits were adequate. The item parameter estimates obtained for end-of-test items from the responses of examinees in the non-speeded class were shown to be more similar to the difficulties of those same items, when they were administered at earlier, non-speeded locations on a different form of the test.

Although it appears possible to be able to remove some of the effects of speededness from item parameter estimates, it is also important to understand how examinees classified into speeded or non-speeded groups differ. This should help explain why some examinees respond differently at the end of a test, when time limits are too short, and why other examinees do not. Bolt, et al. (in press b) found that the members of speeded and non-speeded classes differed in terms of ethnic group but not gender. Although gender differences did not seem to be present between the speeded and non-speeded classes, it is possible that other important differences might exist. For example, it is possible that the speeded and non-speeded classes differ in their academic performance. This might be reflected in higher admissions or placement test scores for members of the non-speeded class. It might also be the case that non-speeded examinees perform better in the classroom, earning higher grades. The use of mixture models in this context enables us to identify examinees who are members of different latent classes, thereby, letting us study their class-specific characteristics. In this study, we seek to understand more about how members of the speeded and non-speeded classes differ, in particular with respect to academic achievement related variables.

Mixture Rasch Model. Previous research with mixture IRT models has shown that they can be used to identify latent classes who use different problem-solving strategies (Mislevy & Verhelst, 1990) or who demonstrate different skills needed to solve test items (Rost, 1990). Results from Bolt, et al. (in press b) suggest that a MRM can also be used to identify a class of examinees for whom the test is speeded and one for whom it is not. Mixture IRT models, such as the MRM, have recently been suggested as a useful means of investigating how qualitative examinee differences, such as use of different problem-solving strategies, may lead to differences in responses to test items (Embretson & Reise, 2000). In addition, recent work with a mixture nominal response model has demonstrated that it is possible to detect different latent classes in multiple choice data and to obtain diagnostic information about examinees based on their class membership (Bolt, in press a).

The MRM described by Rost (1990) considers an examinee population assumed to be composed of a fixed number of discrete latent classes of examinees. The Rasch model is assumed to hold within each class, but each with different item difficulty parameters. Members of a class may differ in ability. The MRM describes each examinee with a class membership parameter, g, which determines the relative difficulty ordering of the items for that examinee, and a continuous latent ability parameter in class g, θ_g which affects the number of items the examinee is expected to answer correctly. The probability of a correct response is written in the MRM as:

$$P(U=1|g,\theta_g) = \frac{\exp(\theta_g - b_{ig})}{1+\exp(\theta_g - b_{ig})} \quad , \tag{1}$$

where b_{ig} is the Rasch difficulty parameter of item i for class g.

The within-class item difficulty estimates are subject to the norming constraint $\sigma_i b_{ig} = 0$. This is necessary for identification, and also ensures that differences between the $g = 1, \ldots, G$ classes are attributable to some items being differentially difficult across the classes and not just to differences in the number of items answered correctly (Rost, 1990). This constraint also enables comparisons of θ_g across classes so that differences in distributions of θ_g can be related to differences in the number of items that members of each class are expected to answer correctly. In this study, we model the assumption that speeded examinees are expected to perform less well than non-speeded examinees on items located at the end of tests by constraining the item difficulties to be higher in Class 1 than in Class 2. In the MRM, this constraint indicates an expectation that end-of-test items will be more difficult in the speeded class (Class 1) compared to the non-speeded class (Class 2).

Class parameters μ_g and σ_g denote the mean and standard deviation for ability, respectively, for class g. Differences among classes in their θ_g distributions can account for the number of items that members in each class can be expected to answer correctly. We assume here that these θ_g are normally distributed. Finally, a set of parameters called mixing proportions, π_g, are specified in the model to indicate the proportion of examinees in each latent class. These mixing proportions are constrained so that $\Sigma_{g=1}^{G} = 1$.

Once the item parameters are estimated for the two classes, they can be used to estimate class membership for other previously unclassified examinees. Classification of a new sample of examinees not included in the estimation of model parameters is done by holding the item parameters fixed at their estimated values and then re-running the estimation algorithm using the new response vectors. We use this two-stage approach to first estimate item and class parameters for speeded and non-speeded classes and then examine academic and additional demographic characteristics of the members of both.

Methods

Data. The 100 operational multiple-choice items on a college-level mathematics placement test were analyzed for this study. The placement test is used by advisors and faculty to place students into courses in the pre-calculus sequence or the first calculus course. These items are divided into three sections: Section A (35 items) measures achievement in arithmetic, intermediate algebra, and intuitive geometry; Section B (46 items) measures college-level algebra and plane geometry; Section C (35 items) measures analytic geometry, trigonometry, and functions and graphs. Item locations 10, 20, and 30 on Sections A and C and items 10, 20, 30, and 40 on Section B are for item tryouts. Although these items are not used to compute examinees scores and were not included in the analyses, they do contribute to test length and, therefore, to possible speededness. All three sections are contained in the same test booklet, but examinees take only two sections, either the 74 operational items on sections A and B (AB Test) or the 74 operational items on sections B and C (BC Test). Examinees who have less than 2.5 years of high school mathematics and who have not studied trigonometry are advised to take the AB Test, and examinees who have 2.5 years or more of high school mathematics and who have studied trigonometry are advised to take the BC Test. The 42 operational items on Section B, therefore, are common to all examinees. The trigonometry items were omitted from the end-of-test items analyzed in this study.

Items from the first two-thirds of either the AB Test or the BC Test were assumed to be non-speeded. Items in the remaining third were assumed to contain some amount of speededness. The last eight items on either the AB or BC Tests were modeled in the MRM as potentially speeded for purposes of this study. It is useful to note that the items assumed to be speeded at the end of the AB Test were actually located between items 36 to 46 on Section B and, consequently, were near the middle of the BC Test. In their locations on the BC Test, these same items were assumed to be unspeeded.

Estimation of Rasch Item Parameters. Two samples were randomly drawn without replacement from the 20,349 examinees who had responded to either the AB or the BC Tests during the 1997/98 school year: One sample of 3,000 examinees was drawn from the 13,102 examinees who responded to the AB Test (i.e., Sections A and B) and a second sample of 3,000 was drawn from the 7,073 examinees who responded to the BC Test (i.e., Sections B and C). These two samples were used to obtain estimates of the Rasch item difficulty parameters for the non-speeded portions of the AB Test and BC Test. These difficulty parameters were estimated for the non-speeded portions of the AB and BC Tests using the computer program MULTILOG (Thissen, 1991) and are identified in Table 1 as difficulty constraints. The values in Table 1 were then used to fix the difficulties for these items in the subsequent mixture model analysis to determine latent speeded and non-speeded groups.

Insert Table 1 About Here

Estimation of Mixture Rasch Model. Two samples of data were used for the estimation of MRM parameters: 3,000 examinees were randomly selected from the sample of 13,102 examinees who had responded to the AB Test, and a similar 3,000 examinees were selected from the sample of 7,073 examinees who responded to the BC Test. A Markov chain Monte Carlo (MCMC) estimation algorithm employing adaptive rejection sampling was then used to estimate the remaining model parameters for the AB Test and for the BC Test, respectively. This algorithm is implemented in the WinBUGS software (Spiegelhalter, Thomas, & Best, 2000). MCMC estimation algorithms have been receiving increasing attention in IRT and offer great promise for use in estimating parameters of more complex types of IRT models (Patz & Junker, 1999a, 1999b, Baker, 1998, Kim, in press, Wollack, Bolt, Cohen & Lee, in press). The appeal of MCMC in this paper stems from its ability to handle ordinal constraints on model parameters. Imposing ordinal constraints results in restrictions on the domain of permissible values that can be sampled. In this study, we used ordinal constraints to model a speeded and a non-speeded group by constraining the Rasch item difficulty estimates to be larger in Class 1, the speeded class, than in Class 2, the non-speeded class.

The MCMC algorithm used here samples a class membership for each examinee at each stage of the chain and then samples item and class parameter values conditional on those class memberships. This is done by first sampling a class membership for each examinee, $j, c_j = (1, 2)$, along with an ability, θ_{jg}, at each stage of the Markov chain, proportional to the probability of the examinee's membership in that class, conditional upon all class parameters. In this study, only the speeded items were sampled. All non-speeded items were fixed at values obtained from the MULTILOG analysis (described above). This meant that the speeded items and the class parameters were sampled from their full conditional posterior distributions, given the already sampled class memberships and examinees' abilities.

The MCMC algorithm requires initial values for each parameter that is to be sampled. These values were generated within the WinBUGS program. Some information from the initial iterations is discarded because sampled values tend to be dependent on the starting values. These discarded iterations are referred to as the *burn-in* iterations. The remaining iterations are based on a chain that is assumed to have converged to its stationary distribution. Estimates of sampled values are obtained from these final iterations. Over the course of the Markov chain, the class parameters come to be defined according to the frequency with which each examinee is sampled into each class. The frequency with which each examinee is sampled into each class over the course of the Markov chain defines the posterior probability of each examinee's membership in that class.

By imposing priors for each parameter in the model, MCMC methods essentially estimate the full conditional posterior of each parameter given the data and the other parameters in the model. The estimated posterior is obtained by simulating a Markov chain in which the stages represent a sample from the posterior distribution of the parameter. The sample mean of the chain gives an estimate of the mean of the posterior and is generally taken as the estimate of the parameter.

To derive the posterior distributions for each parameter, it is first necessary to specify the prior distribution for each. The following priors were used in the two-class MRM in this study:

$$b_{ig} \sim Normal(0,1), \qquad i = 1, \ldots, I, g = 1, 2$$

$$\theta_{jg} \sim Normal(\mu_g, 1), \qquad j = 1, \ldots, N$$

$$c_j \sim \text{Bernoulli}(\pi_1, \pi_2), \qquad j = 1, \ldots, N$$

$$\mu_g \sim Normal(0, 1), \quad g = 1, 2$$

$$(\pi_1, \pi_2) \sim Dirichlet(100, 300),$$

where I is the number of items and N is the total number of examinees. Results from Bolt, et al. (in press b) suggested that mixing proportions of .25 and .75 could be expected for the speeded and non-speeded groups, respectively, for these data. Consequently, we used priors on the mixing proportions, π_g, which were strong and caused the MRM to yield mixing proportions that were close to these values. In addition, in this paper, σ_g was fixed at 1 in both classes.

The MRM presented here includes two sets of constraints on the item difficulty parameters, one set to reflect the non-speeded items and a second set to reflect the end-of-test items assumed to be speeded. The difficulty estimates for the non-speeded items were fixed at the values estimated using the computer program MULTILOG (Thissen, 1991). The items at the end of the test were not fixed at any values but rather were constrained to be easier for the members of the non-speeded class (Class 2). These constraints are illustrated in the WinBUGS code in Appendix A.

Results

Determination of a suitable burn-in was based on 12,000 iterations of the Markov chain. The chains for these iterations are illustrated in Figure 1 for $b_{1,25}$, $b_{2,25}$, $b_{1,26}$,$b_{2,26}$,μ_1,μ_2,π_1, and π_2 for the AB Test. The chains for all b_{ig}s were similar to those in Figure 1 and were observed for all items, speeded and non-speeded on both the AB and BC Tests. WinBUGS provides several indices which can be used to determine an appropriate length for the burn-in. These indices suggested that burn-in lengths of less than 100 iterations were reasonable for all the parameters sampled. This is evident in Figures 1 and 2, as each of the chains converged relatively quickly to its stationary distribution within about the first 50 iterations. A conservative estimate of 1,000 iterations for the burn-in was used in this study. For each chain, therefore, the initial 1,000 iterations were discarded and the algorithm was run to sample an additional 11,000 iterations. Estimates of model parameters were based on the means of the sampled values from the iterations following burn-in.

Insert Figure 1 About Here

Although the AB and BC Tests both contained 74 items, to simplify these analyses, only 26 items were used – 18 items assumed to be non-speeded and 8 end-of-test items assumed to be speeded. The two-class MRM was constrained for both the AB and BC Tests so that the Rasch item difficulties for the non-speeded items were the same in Class 1 and Class 2 (i.e., $b_{i1} = b_{i2}$ for $i = 1, \ldots, 18$). In addition, ordinal constraints were imposed on the difficulties of the items assumed to be speeded (i.e., $b_{i1} > b_{i2}$ for $i = 19, \ldots, 26$). These constraints were actually imposed on the non-normalized beta's in the WinBUGS code shown in Appendix A. The constraints are shown in Table 1 along with the subsequent normalized item difficulties for the non-speeded and speeded items. As can be seen in Table 1, the normalizing resulted in difficulties (noted in the WinBUGS code as b_{ig}s) which differed by a constant of approximately .47 for the AB results and .43 for the BC results.

Two sets of results are presented below. The first set compares the item difficulties estimated for the latent speeded and non-speeded classes. The second examines characteristics of examinees in the speeded and non-speeded classes with an eye toward understanding why the test is speeded for some and not for other examinees. The difficulty estimates for the speeded items are given at the bottom of Table 1.

Item Parameter Estimation

The normalized item parameter estimates, b_{ig}s, obtained for the speeded and non-speeded classes are given in Table 1. As noted above, the non-speeded items (items 1 - 18) for both tests differed by a constant across the two classes. This is due to the equality constraints imposed and also to the normalizing. Differences between the speeded items (items 19 - 26) on both the AB Test and the BC Test, however, are not constant. All the speeded items were constrained to be harder for members of Class 1, but some items were relatively more difficult than others for the members of Class 1 than for the members of Class 2. AB items 23, 25, and 26, for example, were much higher in difficulty for Class 1 than for Class 2 indicating these items contributed more to distinguishing between the classes than did items such as 19 and 21. Similarly, BC items 25 and 26 were harder for Class 1 than Class 2, and contribute more to differentiating the two classes, whereas BC item 23 showed little difference between classes.

The priors on the mixing proportions were strong and the resulting values were very close to those obtained by Bolt, et al. (in press b). For the AB Test, the proportions were $\pi_1 = .26$ and $\pi_2 = .74$; for the BC Test, they were $\pi_1 = .25$ and $\pi_2 = .75$. The mean abilities for the AB Test were $\mu_1 = -.26$ and $\mu_2 = -.24$, reflecting no difference between Class 1 and Class 2. For the BC Test, the mean ability for the speeded class was lower than the mean in the non-speeded class ($\mu_1 = .35$ and $\mu_2 = .61$). The results for the BC Test agree with those observed by Bolt, et al. (In press b), but the lack of differences between the classes for the AB Test did not.

Three TCCs are shown in Figure 2 for the eight end-of-test items on the AB Test (items B36 to B46) to illustrate the improvement in estimation of the item difficulty parameters: (1) AB Total, based on parameter estimates obtained from MULTILOG for the total AB Test sample; (2) BC Total, based on MULTILOG parameter estimates as obtained for the total BC Test sample; and, (3) mixture Rasch difficulty estimates for only the non-speeded examinees administered the AB Test. If the MRM used in this study is effective at identifying examinees for whom the test is speeded, the difficulty estimates from this last group, the non-speeded examinees taking the AB Test, should be more similar to the estimates from the total group taking the BC Test (where speededness should not exist for these items), than should the estimates for the total AB Test sample, which includes examinees in both Classes 1 and 2. Item difficulty estimates for all comparisons were equated to the AB Test scale using the characteristic curve method (Stocking & Lord, 1983) as implemented in the computer program EQUATE (Baker, 1993). Six items at the beginning of the BC Test provided the common item link to the AB Test scale. These six items were assumed to be sufficiently close to the middle of the AB Test so as to not be speeded. DIF items from this link were removed using iterative linking (Candell & Drasgow, 1988). The equating constant obtained from these common items was 2.17 and was used to place the estimates from the BC Total onto the AB Total scale. Equated Rasch parameter estimates for the eight common items from the end of the BC Test are reported in Table 2.

Insert Table 2 and Figure 2 About Here

The TCC in Figure 2 for the non-speeded class taking the AB Test is very close to the TCC for the total sample taking the BC Test, indicating that the item parameter estimates are similar in these two groups. The TCC estimated from the AB Total sample, however, is lower than the other two and indicates that the difficulty estimates are harder in the AB Total sample. The similarity of the TCCs for the total sample taking the BC Test and the non-speeded sample for the AB Test indicates that much of the bias due to speededness appears to have been removed from the difficulty estimates from the AB Test sample. It does appear, in other words, that we have improved the estimation of item parameters in the AB Test sample by using only the responses from members of the non-speeded class.

It is of interest to note that this comparison between eight end-of-test items on the AB Test and the same items on the BC Test is actually a vertical equating situation in which the items on the AB Test require less knowledge of mathematics than the items on the BC Test. The 36 common items on the B section of the test, only 8 of which were assumed to be non-speeded for the AB Test, provide a convenient means of equating the two tests. In addition, the vertical equating situation is one in which examinees taking the AB Test had lower ability than examinees taking the BC Test. The examinees taking the AB Test had lower raw scores on the 36 common B items (M = 16.34 (SD = 7.20)) than those taking the BC Test (M = 28.95 (SD = 7.90)). Using responses only from the non-speeded examinees taking the AB Test improved the estimates by reducing the effects of speededness from the parameter estimates of the eight items at the end of the AB Test.

Characteristics of Latent Groups

The full sample of 13,102 examinees taking the AB Test and 7,073 examinees taking the BC Test were classified into speeded and non-speeded groups using MCMC with the model parameters fixed at the values obtained above so that only group memberships were estimated. These two samples were then evaluated to determine whether gender or age might be associated with speededness (all significance tests in this paper were evaluated at the $\alpha = .01$ level). Table 3 reports the proportions of examinees by gender classified into either the speeded or non-speeded groups for both samples. An association was found between gender and class membership in the BC sample but not in the AB sample.

Insert Table 3 About Here

Pearson chi-squares revealed no associations between age and class membership in the AB sample but did show an association in the BC sample ($p < .01$). In the AB sample, the majority of examinees (91%) were between ages 17 and 21 (Mode = 18). In the BC sample, the majority of examinees (97%) were between 18 and 19 (Mode = 18).

Gender differences were observed on both sections for this sample ($p < .01$), however, the effect sizes were small (approximately .10 for both). The raw score means for the males were slightly higher for both Section A ($M_{males} = 19.80$ vs $M_{females} = 19.15$) and Section B ($M_{males} = 16.71$ vs $M_{females} = 16.11$). Both class and gender differences in raw scores were also observed for the BC sample ($p < .01$). Effect sizes for the Section B raw score were .86 between latent groups and .30 for gender. Interestingly, it was the speeded class, and not, as might have been expected, the non-speeded class, that produced higher raw scores. The raw score means for the speeded class were higher for both Section B ($M_1 = 34.66$ vs $M_2 = 28.12$) and Section C ($M_1 = 18.46$ vs $M_2 = 16.19$). For gender, the mean for males was higher than for females for both Section B ($M_{males} = 32.04$ vs $M_{females} = 30.74$) and Section C ($M_{males} = 18.27$ vs $M_{females} = 16.38$).

Background and Achievement Data. High school background information, entrance examination scores, and first three years of college achievement information were available for a sample of examinees from one university. A number of variables were examined in both the AB Test and BC Test samples. These included information from the high school transcripts (e.g., numbers of high school academic and nonacademic units, units in biology, in chemistry, in English, in first and second foreign languages, in mathematics, and in physics, rank in graduating class, and graduation year), scores from the admissions data on each student (e.g., ACT scores, SAT scores, placement test scores in English, foreign language, and mathematics, and number of transfer credits), and college achievement data (e.g., first and second semester grade point averages (GPA), cumulative GPA, number of degree credits, number of failure credits, number of courses dropped, dates of course drops, and mathematics GPA). Univariate gender by latent group ANOVAs were used to analyze these data. Results are summarized in Table 4. Only results which were significant at $\alpha = .01$ are included in Table 4.

Insert Table 4 About Here

Most of the differences reported in Tables 4A and 4B are gender-related. No gender × class interactions were observed. Females in the AB Test sample had a higher number of foreign language units, higher ACT-English scores, higher English Placement Test scores, higher first semester and cumulative GPAs, higher mathematics GPAs, more degree credits and transfer credits, and lower ACT-Mathematics scores and lower numbers of failure credits. In the AB Test sample (Table 4A), only three variables had differences between the speeded and non-speeded groups: Cumulative GPA, mathematics GPA and number of degree credits were all higher in the speeded group, although all effect sizes were small. Gender-related differences in the AB group also had effect sizes which were small.

More differences were noted in the BC Test sample between gender groups and between latent groups (Table 4B) than were observed in the AB Test sample. Even so, the majority of these differences also had small effect sizes. The college algebra placement test score did show a moderate effect size, but this is most likely because most of the items used to classify examinees into speeded and non-speeded groups were the same as used to calculate this score. A number of gender differences were noted for the BC sample, but the effect sizes were almost all small.

First and second academic majors at the end of the third year of courses were also available for students in the University sample (see Table 5). These majors could change before graduation, but do provide an indication of the types of majors that students in this sample selected. Majors were categorized into one of five main discipline areas: humanities, biological sciences, physical sciences, social sciences, and undeclared. The majority of students had a single major whether they were in the AB or BC samples (74.2% and 69.4%, respectively). In the AB sample, the majority of students had either a Social Science (43.3%) or a Humanities major (19.7%). In the BC sample, the majority had a Social Science major (34.3%) followed by majors in Biological Sciences (19.6%) and Physical Sciences (18.3%).

Insert Table 5 About Here

Most students were classified into the non-speeded class in both the AB and BC samples (80.3% and 84.7%, respectively). Differences in selection of majors between the latent classes appeared to be small. In the AB Test sample, although there were relatively few majors in the Physical Sciences, the majority of these (28 of 31) were in the non-speeded class. In the BC Test sample, the largest group of majors in the non-speeded class was the Biological Sciences (89.0%) and the smallest was the Physical Sciences (79.2%).

Discussion

Test speededness effects introduce a higher level of difficulty on end-of-test items (Douglas, Kim, Habing, & Gao, 1998; Oshima, 1994). Previous work has shown that a MRM can be useful in removing some of the effects of speededness on item parameter estimates (Bolt, et al., in press b). The focus of this paper was on studying characteristics of individuals classified by a MRM into latent speeded or non-speeded groups, with an eye toward trying to understand how speededness effects might influence test performance. We first applied the same MRM approach used by Bolt, et al. to responses on a college-level mathematics placement test in order to identify those examinees for whom the test was speeded and those for whom it was not. This was done on two samples of examinees who took one of two different forms of the test. The second half of the AB Test was the same as the first half of the BC Test. This meant that items at the end of the AB Test were actually in the middle of the BC Test. Item parameter estimates were compared with and without speededness effects removed. Eight end-of-test items on each test were assumed to be in speeded locations. Parameter estimates for these eight end-of-test items from the AB Test obtained on the sample with the speeded class of examinees removed were found to be nearly the same as estimates for these same items on the BC Test. This result agrees with previous research by Bolt, et al. (in press, b).

The data for this study were further interesting in that comparison of the parameter estimates for items from the two forms of the test was done in the context of a vertical equating situation. The two forms of the tests differed in difficulty and the two samples of examinees differed in their levels of mathematics ability. Wollack, Cohen, and Wells (2002) have shown that poor item parameter estimation due to test speededness has a negative effect on the quality of horizontal equating, thereby affecting the interpretation of the equated scaled scores. The use of the MRM in this study appears to have removed at least some of the effects of speededness in a vertical equating context such that TCCs from the parameter estimates for the common items did not differ.

Some differences were found between examinees in the two latent classes on gender and academic achievement variables. Although gender was not found to be related to latent class membership for the AB Test sample, an association was found for examinees taking the BC Test. Results for the BC Test agree with previous research (Bolt, et al., in press b). In addition, gender differences were found in raw score performance on the B and C sections of the BC Test.

Examinees classified in the speeded groups actually seem to have performed better than non-speeded examinees on some measures of academic achievement. This is somewhat counter-intuitive given the way we modeled speededness. Speeded examinees were found to have had somewhat higher admissions and placement test scores, and in their subsequent college careers, higher grade point averages, fewer credits of grade F (i.e., fewer failure credits), more transfer credits, and more credits toward degree. Further, the magnitudes of raw score differences on BC tests suggest that speeded examinees were able to perform very well on non-speeded portions of the test. The impact of speededness on these examinees appears to have been to limit their performance on the end-of-test items. This result occurred because members of the speeded group could only be identified provided they performed noticeably worse on the end-of-test items than on the earlier items. As a consequence, examinees who performed consistently poor on all items were classed into the non-speeded group. Such examinees may or may not also be speeded examinees, but we could not tell from their response patterns.

It is important to note that the results of this use of MCMC for estimation of latent groups are in no small part a function of the ways in which the problems were structured and the data were analyzed. It is possible, in other words, that some of the differences observed between latent groups may have been due to the particular ways in which test speededness was modeled. Had we modeled speededness differently, perhaps by constraining fewer non-speeded items, we might have observed other differences between the latent groups or other compositions of the groups. Use of the MRM may have resulted in different latent classes than might have been observed had we used a more highly parameterized model such as a mixture nominal response model. It is also possible that different forms of constraints, such as focusing on specific components of test items or on specific response strategies might have resulted in still different formations of latent speeded and non-speeded classes. It seems clear that attention needs to be paid to the kinds of constraints and priors used in order to make certain their impact on the resulting solutions is understood.

The application of the MRM proposed in this paper is important as it provides a means of identifying those examinees for whom the time limits are too short to be able to use the same response strategies they used in non-speeded portions of the test. The qualitative differences noted between examinees in the speeded and non-speeded classes do help to provide insight into why these individuals are classed into one of these groups. Additional research using mixture models can help identify other important examinee differences which may lead to a better understanding of how time constraints affect certain examinees.

References

Baker, F.B. (1993). EQUATE 2.0: A computer program for the characteristic curve method of IRT equating. *Applied Psychological Measurement, 17*, 20.

Baker, F.B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement, 22*, 153-169.

Bolt, D.M., Cohen, A.S., & Wollack, J.A. (in press a). A mixture item response for multiple-choice data. *Journal of Educational and Behavioral Statistics.*

Bolt, D.M., Cohen, A.S., & Wollack, J.A. (in press b). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement.*

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.

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.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Evans, F.R.,& Reilly, R.R. (1972). A study of speededness as a source of test bias. *Journal of Educational Measurement, 9*, 123-131.

Kim, S.-H. (2001). An evaluation of a Markov chain Monte Carlo method for the Rasch model. *Applied Psychological Measurement, 25*(2), 163-176.

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

Mislevy, R.J. & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195-215.

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. (199b). 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.

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Spiegelhalter, D., Thomas, A. & Best, N. (2000). *WinBUGS version 1.3* [computer program]. Robinson Way, Cambridge CB2 2SR, UK: MRC Biostatistics Unit, Institute of Public Health. http://www.mrc-bsu.cam.ac.uk/bugs.

Thissen, D. (1991). *MULTILOG, version 6.0* [computer program]: Multiple, categorical item analysis and test scoring using item response theory. Chicago, IL: Scientific Software, Inc.

Wollack, J.A., Bolt, D.M., Cohen, A.S., & Lee, Y.-S. (in press). 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.*

Wollack, J.A., Cohen, A.S., Bolt, D.M., & Wells, C.S. (April, 2002). *The Effects of Test Speededness on Score Scale Stability.* Paper presented at the annual convention of the American Educational Research Association, New Orleans, LA.

Table 1. Rasch Difficulty Estimates for Speeded and Non-Speeded Items on AB and BC Tests

AB Test Results					BC Test Results				
			MRM Difficulty Estimates					MRM Difficulty Estimates	
		Difficulty	Class 1	Class 2			Difficulty	Class 1	Class 2
Item	Location	Constraints	(Speeded)	(Non-Speeded)	Item	Location	Constraints	(Speeded)	(Non-Speeded)
1	A23	.07	-.54	-.07	1	B27	-.48	-.31	.13
2	A24	-.29	-.90	-.43	2	B28	-.69	-.51	-.08
3	A25	-.01	-.62	-.15	3	B29	-1.18	-1.00	-.57
4	A26	.12	-.49	-.02	4	B33	-1.22	-1.04	-.61
5	A27	-.47	-1.08	-.61	5	B36	-.96	-.78	-.35
6	A28	.43	-.18	.29	6	B37	-1.76	-1.58	-1.15
7	A29	-1.72	-2.33	-1.86	7	B39	-1.08	-.90	-.47
8	A31	.59	-.02	.45	8	B41	-1.22	-1.04	-.61
9	A32	.20	-.41	.06	9	B42	-.31	-.13	.30
10	A33	-.45	-1.06	-.59	10	B43	-1.75	-1.57	-1.14
11	A34	.35	-.26	.21	11	B45	-1.68	-1.50	-1.07
12	A35	-1.48	-2.09	-1.62	12	B46	-.16	.02	.45
13	B1	.26	-.35	.12	13	C3	-2.32	-2.14	-1.71
14	B2	-.18	-.79	-.32	14	C7	-.94	-.76	-.33
15	B4	.98	-.37	.84	15	C8	.76	.94	1.38
16	B5	-.51	-1.12	-.65	16	C9	.23	.41	.84
17	B6	.22	-.39	.08	17	C11	-.42	-.24	.19
18	B7	.63	.02	.49	18	C12	-.17	.01	.44
19	B36		1.18	.79	19	C27		.82	.04
20	B37		.89	.18	20	C28		-.03	-1.14
21	B39		.82	.40	21	C29		.97	.19
22	B41		1.61	.87	22	C31		2.07	1.33
23	B42		2.56	1.12	23	C33		1.54	1.43
24	B43		1.30	-.74	24	C34		1.56	.38
25	B45		1.43	-.05	25	C35		2.56	.77
26	B46		2.51	1.17	26	C36		2.62	1.28

Table 2. Rasch Difficulty Estimates

Of Speeded Items on AB and BC Tests

Results for AB Test			
	Location	Class 1	Class 2
	AB Test	(Speeded)	(Non-Speeded)
19	B36	1.18	.79
20	B37	.89	.18
21	B39	.82	.40
22	B41	1.61	.87
23	B42	2.56	1.12
24	B43	1.30	-.74
25	B45	1.43	-.05
26	B46	2.51	1.17
Results for BC Test			
19	C27	.82	.04
20	C28	-.03	-1.14
21	C29	.97	.19
22	C31	2.07	1.33
23	C33	1.54	1.43
24	C34	1.56	.38
25	C35	2.56	.77
26	C36	2.62	1.28

Table 3. Gender Characteristics of Speeded and Non-Speeded Samples

Sample	Gender	N	Class 1	Class 2	χ^2	df	p-value
AB Test	Male	5,228	1,002 (19.2%)	4,226 (80.1%)			
	Female	7,713	1,572 (20.4%)	6,141 (79.6%)	2.89	1	NS*
BC Test	Male	3,735	555 (14.9%)	3,019 (85.1%)			
	Female	3,313	294 (8.9%)	3,180 (91.9%)	59.36	1	$p < .01$

* Not significant.

Table 4A: Demographic and Achievement-Related Characteristics of Latent Groups for the AB Test, University Sample

	Gender					Latent Group				
	Male		Female			Speeded		Non-Speeded		
					Effect					Effect
Variable	Mean (SD)	N	Mean (SD)	N	Size	Mean (SD)	N	Mean (SD)	N	Size
High School Variables										
Foreign Language										
Units	3.12 (1.13)	535	3.41 (1.15)	1,060	.25					
College Entrance Examination & Placement Test Scores										
ACT-English	23.73 (3.89)	422	24.95 (3.73)	916	.32					
ACT-Mathematics	23.97 (3.57)	422	22.93 (3.27)	916	.31					
English Placement										
Test	615.47 (84.47)	475	639.07 (81.89)	936	.29					
College Achievement Data										
First Semester GPA*	2.64 (.76)	527	2.82 (.77)	1,054	.24					
Cumulative GPA	2.72 (.74)	535	2.97 (.68)	1,062	.36	2.98 (.65)	314	2.87 (.73)	1,283	.15
Degree Credits	95.62 (37.47)	535	103.13 (36.83)	1,062	.20	106.06 (31.44)	314	99.28 (38.37)	1,283	.18
Transfer Credits	7.73 (12.18)	535	10.21 (12.29)	1,062	.20					
Failure Credits	1.19 (2.93)	535	.56 (2.11)	1,062	.26					
Mathematics GPA	2.09 (.99)	426	2.38 (1.00)	750	.29	2.42 (.96)	229	2.24 (1.01)	947	.18

* Grade Point Average

Table 4B. Demographic and Achievement-Related Characteristics Of Latent Groups for the BC Test, University Sample

	Gender					Latent Group				
	Male		Female			Speeded		Non-Speeded		
					Effect					Effect
Variable	Mean (SD)	N	Mean (SD)	N	Size	Mean (SD)	N	Mean (SD)	N	Size
High School Variables										
Rank in Class	39.96 (46.52)	1,534	32.56 (40.53)	1,289	.17					
Biology Units	1.35 (.58)	1,536	1.47 (.66)	1,290	.20					
Chemistry Units						1.28 (.50)	434	1.21 (.50)	2,392	.13
Physics Units	.99 (.45)	1,536	.83 (.48)	1,290	.34					
Foreign Language Units	3.42 (1.09)	1,535	3.69 (1.03)	1,287	.25					
College Entrance Examination & Placement Test Scores										
ACT-Composite						28.17 (2.81)	398	26.95 (3.29)	2,237	.38
ACT-English	25.84 (3.78)	1,417	26.48 (3.79)	1,218	.17	26.96 (3.53)	398	25.99 (3.83)	2,237	.26
ACT-Mathematics	28.30 (3.47)	1,417	26.68 (3.57)	1,218	.46	28.88 (2.91)	398	27.31 (3.67)	2,237	.44
ACT-Reading						28.25 (4.68)	398	27.24 (4.83)	2,237	.21
SAT-Mathematics	667.27 (64.73)	433	624.49 (73.84)	341	.62	677.50 (58.92)	144	641.78 (73.16)	630	.50
English Placement Test	656.97 (84.21)	1,385	673.88 (87.80)	1,178	.20	684.19 (85.17)	387	661.28 (86.03)	2,176	.27
French Placement Test	485.92 (112.65)	141	538.65 (117.28)	213	.46					
College Algebra Test	707.74 (96.43)	1,536	673.38 (94.32)	1,290	.36	735.52 (76.06)	434	684.17 (98.28)	2,392	.54
Trigonometry Test	715.31 (96.66)	1,536	675.53 (101.66)	1,290	.40	727.00 (86.67)	434	691.74 (102.38)	2,392	.35
College Achievement Data										
First Semester GPA	2.99 (.75)	1,534	3.11 (.70)	1,283	.17					
Cumulative GPA	2.97 (.69)	1,536	3.16 (.66)	1,290	.27					
Transfer Credits	9.70 (11.07)	1,536	12.67 (12.08)	1,290	.26					
Failure Credits	.87 (2.81)	1,536	.34 (1.79)	1,290	.22					
Mathematics GPA	2.58 (.96)	1,375	2.73 (.93)	1,000	,16	2.77 (.90)	374	2.62 (.96)	2,001	.16

* Grade Point Average

Table 5. Academic Majors of Students in the University Sample

	First Major			Second Major		
	Speeded N (%)	Non-Speeded N(%)	Total N	Speeded N (%)	Non-Speeded N(%)	Total N
AB Test Sample						
Undeclared	49 (13.1)	324 (86.9)	373	231 (19.4)	957 (80.6)	1,188
Humanities	62 (19.6)	254 (80.4)	316	26 (22.2)	91 (77.8)	117
Social Sciences	156 (22.5)	537 (77.5)	693	46 (19.7)	188 (80.3)	234
Biological Sciences	45 (23.8)	144 (76.2)	189	11 (21.2)	41 (78.8)	52
Physical Sciences	3 (9.7)	28 (90.3)	31	1 (10.0)	9 (90.0)	10
BC Test Sample						
Undeclared	71 (15.1)	398 (84.9)	469	289 (14.7)	1,678 (85.3)	1,967
Humanities	45 (14.3)	270 (85.7)	315	25 (14.1)	152 (85.9)	177
Social Sciences	149 (15.3)	823 (84.7)	972	60 (15.9)	318 (84.1)	378
Biological Sciences	61 (11.0)	494 (89.0)	555	21 (13.5)	134 (86.5)	155
Physical Sciences	108 (20.8)	410 (79.2)	518	40 (25.8)	115 (74.2)	155

Appendix A: WINBUGS Code Used for Mixture Rasch Model

```
 model
{
# Speededness Study:  Math 97/98 26 AB Test items
# First 18 items with equality constraints (3,000 examinees)
#
# Parameter Notation:
# theta = examinee ability parameter
# gmem = examinee group membership
# beta = non-normalized item difficulty parameter
# b = normalized item difficulty parameter
# pi = class mixing proportion
# mu = class mean ability parameter
{

  for (j in 1:N) {
   for (k in 1:T) {
          r[j,k]<-resp[j,k]
}}
  for (j in 1:G)
{
  alpht[j]<-alph[j]
}

 # Rasch model
   for (j in 1:N) {
       for (k in 1:T) {
           tt[j,k]<- exp(theta[j] - b[gmem[j],k])
           p[j,k]<-tt[j,k]/(1 + tt[j,k])
           r[j,k]~dbern(p[j,k])
           }
         theta[j] ~ dnorm(mut[gmem[j]],1)
         gmem[j] ~ dcat(pi[1:G])
        }
# Equality constraints
     beta[1,1]  <-  .07
     beta[2,1]  <-  .07
     beta[1,2] <- - .29
     beta[2,2] <- - .29
     beta[1,3] <- - .01
     beta[2,3] <- - .01
⋮

# Ordinal Constraints for items 19 to 26
```

```
         for (k in 19:T){
          beta[1,k]~dnorm(0,1.)         }

    for (k in 19:T){
          beta[2,k]~dnorm(0,1.) I(,beta[1,k]) }

    for (k in 1:T){
       for (j in 1:G){
        b[j,k]<-beta[j,k]-mean(beta[j,1:T])
        }
        bdiff[k]<-b[1,k]-b[2,k]
      }

      pi[1:2]~ ddirch(alpht[1:2])
      mut[1]~ dnorm(0.,1.)
      mut[2]~ dnorm(0.,1.)
}

list(N=3000, T=26 G=2,alph=c(100,300),
resp=structure(.Data=c(
1,1,1,0,1,1,1,1,0,1,1,1,1,1,1,0,1,1,1,0,1,1,1,1,1,1,

⋮

1,1,0,0,1,1,1,1,1,0,1,1,1,1,0,1,1,1,1,1,0,0,1,1,1,0), .Dim=c(3000,26)))
```

Figure 1: Sampling Histories

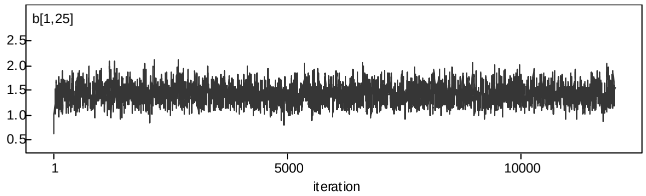

Figure 1a: Sampling History for $b_{1,25}$.

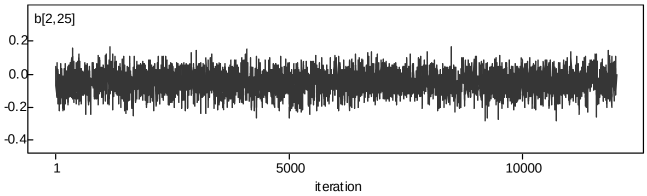

Figure 1b: Sampling History for $b_{2,25}$.

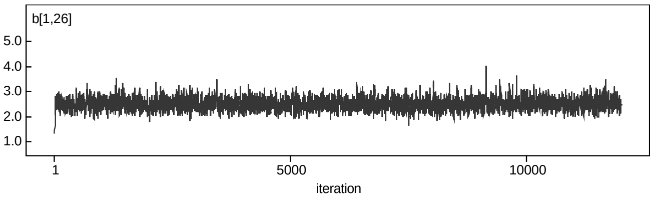

Figure 1c: Sampling History for $b_{1,26}$.

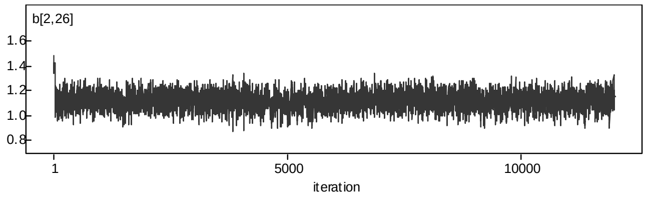

Figure 1d: Sampling History for $b_{2,26}$.

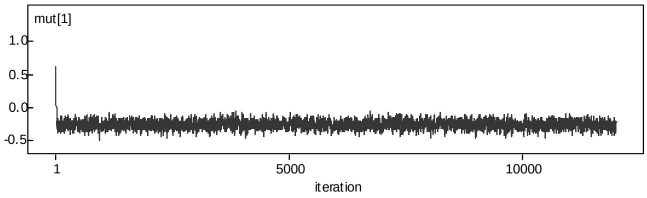

Figure 1e: Sampling History for $\mu(\theta_1)$.

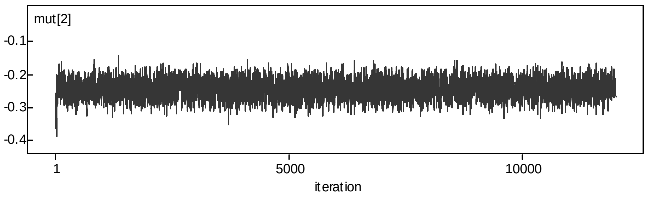

Figure 1f: Sampling History for $\mu(\theta_2)$.

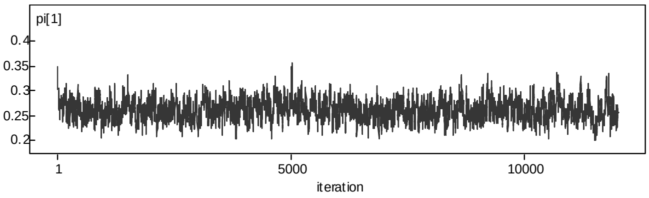

Figure 1g: Sampling History for π_1.

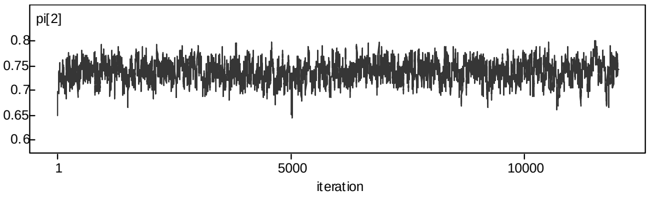

Figure 1h: Sampling History for π_2.

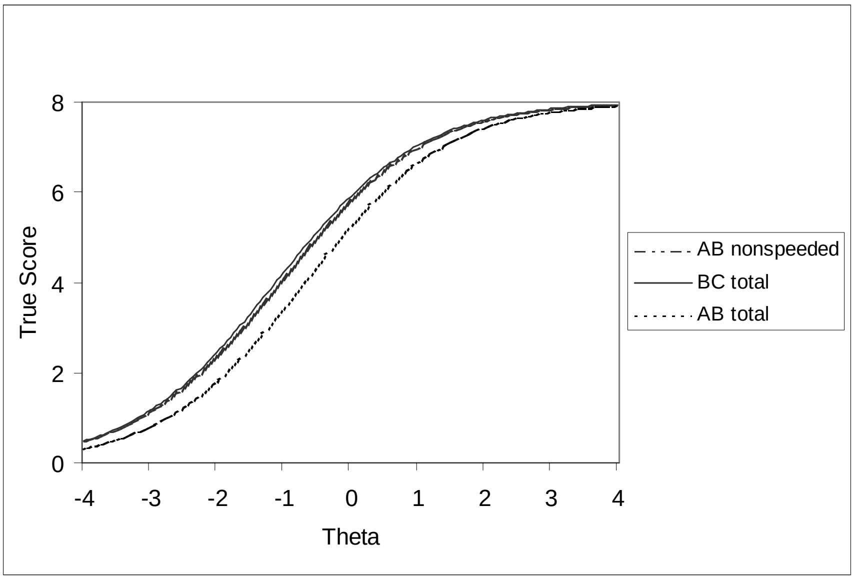

Figure 2: Test Characteristic Curves for Six End-of-Test Common Items.