Fall 10-20-2010

A Comparison of Linear and Nonlinear Factor Analysis in Examining the Effect of a Calculator Accommodation on Math Performance

Minji Kang Lee
University of Massachusetts Amherst, minjik@educ.umass.edu

Craig S. Wells
University of Massachusetts - Amherst, cswells@educ.umass.edu

Stephen G. Sireci
University of Massachusetts Amherst, sireci@acad.umass.edu

Running head: CALCULATOR ACCOMMODATION

A Comparison of Linear and Nonlinear Factor Analysis in Examining the Effect of a Calculator Accommodation on Math Performance

Minji K. Lee, Craig S. Wells, and Stephen G. Sireci

University of Massachusetts Amherst

Abstract

In a statewide achievement test, the use of a calculator is often allowed as an accommodation for students with disabilities (SWD). The purpose of the accommodation is to eliminate the disadvantage faced by SWDs relative to peers without disabilities with the same level of overall mathematics proficiency. However, it is important to determine if the accommodation has changed the construct and meaning of the test scores. One way of examining whether the construct has changed due to the accommodation is to compare performance on the assessment for SWDs to peers without disabilities, controlling for differences in proficiency. In this study, measurement invariance of a statewide 8th-grade mathematics test was explored to evaluate the comparability of the scores obtained by examinees who were accommodated with calculators and those who were not. Structural equation modeling (SEM) was used to assess measurement invariance at the content strand level. Also, nonlinear factor analysis (NFLA) was used to assess the invariance at the item level and the solution was compared to the SEM results. Results indicate the accommodation tended to make items in the content strand Number Sense and Operations easier. The relative strengths and limitations of using SEM and NLFA to evaluate test accommodations is discussed.

A Comparison of Linear and Nonlinear Factor Analysis in Examining the Effect of a Calculator Accommodation on Math Performance

Standardized tests are designed to promote fairness in testing by keeping the test content, administration conditions, and scoring processes consistent for all examinees. However, for examinees with disabilities, one or more features of the standardized conditions may make it more difficult for the student to demonstrate their true knowledge and skills. For this reason, many testing programs offer accommodations to the standard testing conditions to students with disabilities. The purpose of these accommodations is to remove any barrier to valid assessment of students' knowledge and skills. However, concerns over the degree to which the accommodation may change what is measured by the test, warrant research into the degree to which test accommodations affect the validity of the interpretations derived from test scores.

For students who are unable to calculate at any level of difficulty, one common nonstandard accommodation is the provision of a calculator. In order to receive this accommodation, the students must be virtually unable to perform calculation. Without this accommodation, the scores for these students may lead to underestimates of their true mathematics proficiency.

In this study, we examine whether a calculator accommodation changes the measurement properties of a large-scale, statewide mathematics assessment. The primary question motivating our analyses is, "Does this test measure the same ability in essentially the same way, regardless of whether the examinee is allowed to use a calculator?" If the answer to this question is no, then the mathematics assessment may not support valid inferences regarding the mathematics achievement of students with disabilities (SWD) who are given this accommodation. Since the

very purpose of the accommodation is to make the scores of disabled students comparable to students without disabilities, it is critically important to determine whether measurement invariance holds across both groups of students. If measurement invariance holds, then we can say the accommodation does not appear to influence the meaning of the test scores used for a particular intended purpose. If measurement invariance does not hold, then it may be necessary to modify the accommodation in some way. Thus, studies of measurement invariance are important for defending the validity of test accommodations, particularly when scores from accommodated and non-accommodated tests are aggregated together when test results are reported. Such aggregation is common in testing programs associated with statewide assessments under the No Child Left Behind legislation.

Method

Participants

Data from a large-scale, statewide, 8th-grade math assessment were gathered from roughly 70,000 examinees. Approximately 8% (6,140) of the examinees met the qualifications for a specific calculating disability and were allowed to use calculators during the test administration. The two groups of students differed dramatically in their observed score distributions. The total test score distributions for each group are presented in Figure 1. The no-calculator group exhibited a moderate ceiling effect, whereas the calculator group exhibited a positively-skewed distribution with a much lower average score. As Gotzmann (2001), Gierl, Gotzmann, and Boughton (2004), Gotzmann and Boughton (2004), and Sireci and Wells (2009) have noted, it is possible that the disparate proficiency distributions could influence the statistical analyses (e.g., significance levels and effect sizes). Therefore, it is important to prevent these distributional differences from affecting the analysis of measurement invariance. For this reason,

a stratified random sample of 6,140 no-calculator students was matched to the calculator group. That is, for each calculator student with a given sum score, one of the no-calculator students with the same sum score was drawn at random. In addition, both the no-calculator and calculator groups were split into random halves to allow for cross-validation.

INSERT FIGURE 1 ABOUT HERE

Instrument

The math assessment was comprised of 34 items. The items were reported to fall into the following content categories: 11 items (1) Patterns, Relations, and Algebra, 10 items in (2) Number Sense and Operations, 4 items in (3) Measurement, 3 items in (4) Geometry, and 7 items in (5) Data Analysis, Statistics, and Probability. Table 1 shows the content specification of this test.

Data Analysis

NOHARM procedure. In multi-group confirmatory factor analysis, it is possible to test for differences in factor loadings and intercepts between groups. However, this approach is not best when the indicators are dichotomously scored items. The factor loading of an indicator is the slope of the regression of the indicator on the factor/ability/construct measured by all of the indicators in common. If the indicator is dichotomously scored, then the expected score is the same as the probability of answering the item correctly. Because the relationship between the construct and the probability of answering correctly is non-linear, analyzing a set of dichotomous items requires a *nonlinear* factor model that keeps the expected score of each item between zero and one.

To address this problem, the item scores in each half of each group were analyzed with the computer program NOHARM (Fraser & McDonald, 2003), which implements a nonlinear

form of factor analysis appropriate for dichotomous items. The *i*th examinee's response to the *j*th item is assumed to be determined by an underlying continuous quantity ($X_{ij}^{*} = \lambda_j \theta_i + \varepsilon_{ij}$), where θ_i is the examinee's standing on the common factor measured by the test as a whole, λ_j is the item's loading on the common factor, and ε_{ij} is the *i*th examinee's unique behavior on the *j*th item that is unaccounted for by the examinee's standing on the common factor. If this quantity exceeds a certain threshold ($X_{ij}^{*} > \tau_j$), the examinee gives the correct response to the item. If this quantity does not exceed the threshold ($X_{ij}^{*} < \tau_j$), the examinee gives an incorrect response. Thus, NOHARM treats the continuous quantities underlying the item responses as indicators of the common factor θ. Both X_j^{*} and θ are scaled to have standard normal distributions.

The output of NOHARM includes the loading of each item's underlying continuous quantity on the common factor (λ_j) and each item's threshold (τ_j). NOHARM places the sign on τ_j in such a way that larger values correspond to easier items. The area under the curve of the standard normal distribution, from negative infinity to the threshold parameter, gives the percentage of examinees getting the item correct. This means that each τ_j can be treated as a *Z*-score of item difficulty, which helps with interpretation.

McDonald (1999) showed that this nonlinear factor analysis of dichotomous items is equivalent to item response theory (IRT). The traditional IRT parameterization says that an examinee with factor level θ has a probability of getting item *j* correct equal to

$$P(x_j = 1 \mid \theta) = \frac{1}{1 + \exp[-1.7 a_j (\theta - b_j)]},$$

where a_j is the discrimination parameter, and b_j is the difficulty parameter. McDonald showed that to convert λ_j to a_j, we can use the equation $a_j = \lambda_j / \sqrt{1-\lambda_j^2}$. It turns out that we need both τ_j and λ_j to obtain b_j. We can use the equation $b_j = -\tau_j / \lambda_j$. These conversions may be helpful to readers who are used to the traditional IRT notation but not the factor analysis notation. Note that a strength of NOHARM is its capacity to estimate IRT models with more than one construct dimension.

Differential item functioning (DIF) represents the situation where members of two groups differ in the probability of getting an item correct even when matched for the overall construct. In the context of NOHARM, DIF can occur for the following reasons: (1) the two groups have different thresholds (τ_j) that their underlying continuous quantities must exceed in order to get this item right, and (2) the two groups have different slopes (λ_j) in the regression of their underlying continuous quantities on the common factor (θ). In the traditional IRT parameterization, (1) means that the two groups have different difficulty parameters (b_j). Even if the item is equally discriminating in the two groups, a member of the group with the larger difficulty parameter will find the item harder than an equally proficient member of the other group. (2) means that the two groups differ in both their discrimination (a_j) and difficulty (b_j) parameters. Essentially, the item is a less reliable indicator of the construct in the group with the lower discrimination parameter.

The item parameters of the no-calculator and calculator groups were tested for DIF according to a procedure given by Lord (1980) and modified by McDonald (1999). Lord proposed plotting one group's IRT parameters for several items (e.g., the items' discrimination parameters) against the other groups' parameters. Any items lying far from the best-fitting

straight line are identified as showing DIF. McDonald modified this procedure by using the factor analysis parameterization of IRT instead of Lord's traditional parameterization. In McDonald's examples, the factor analysis parameterization of the IRT model tended to agree more across groups than the traditional parameterization, suggesting that it is easier to get accurate estimates of the factor analysis parameterization. This led McDonald to conjecture that the factor analysis parameterization used by NOHARM is superior for the purpose of conducting DIF analysis.

If the construct measured by the test has the same distribution in two groups, and if there is no DIF, then in a graph plotting the factor loadings in one group against the factor loadings in the other group, the points should lie on a straight line through the origin with a slope of 1. Under the same conditions, a graph plotting the thresholds in one group against the thresholds in the other group should also show a straight line through the origin with a slope of 1. If the item parameters as a whole do not show this pattern in one of the graphs, then we can conclude that many of the items show strong DIF. If the pattern does show up but an individual item lies very far from the line, then that particular item shows strong DIF. This visual approach was used to look for DIF. Potential cases of DIF were then verified in the cross-validation samples.

A statistical hypothesis-testing approach was also used to look for DIF. McDonald (1999) stated that the reciprocal of the root sample size gives a rough standard error for a factor loading estimated by NOHARM. In our case, SE ($\hat{\lambda}_C$) = SE ($\hat{\lambda}_{NC}$) = $\frac{1}{\sqrt{3070}}$ = .018.[1] The standard error of the difference between the two factor loadings can be computed as

$$\text{SD}\ (\hat{\lambda}_{NC} - \hat{\lambda}_C) = \sqrt{Var(\hat{\lambda}_{NC} - \hat{\lambda}_C)}$$

[1] The hat over the λ's are meant to emphasize that they stand for the estimates of the factor loadings, not the factor loadings themselves.

$$= \sqrt{Var(\hat{\lambda}_{NC}) + Var(\hat{\lambda}_C) - 2Cov(\hat{\lambda}_{NC} - \hat{\lambda}_C)}$$

$$= \sqrt{2Var(\hat{\lambda}_{NC})} = \sqrt{2} \times SE(\hat{\lambda}_{NC}).$$

This procedure was used to test for group discrepancies in factor loadings by constructing the 95% confidence interval, $\hat{\lambda}_{NC} - \hat{\lambda}_C \pm 1.96 \times \sqrt{2} \times SE(\hat{\lambda}_{NC}) = \hat{\lambda}_{NC} - \hat{\lambda}_C \pm .05$. Therefore, any differences in factor loadings, $\hat{\lambda}_{NC} - \hat{\lambda}_C$, exceeding |.05| will be statistically significant.

Since the threshold parameter is a simple function of the proportion getting the item correct (item difficulty), a test for the difference between two proportions was used to test for group discrepancies in thresholds. This procedure involves computing the standard deviation of the sampling distribution of the difference between two proportions, $\text{SE} = \sqrt{\frac{p(1-p)}{1/n_1 + 1/n_2}}$, where p is the pooled sample proportion, n_1 is the size of the no-calculator group, and n_2 is the size of calculator group. The test statistic is $\text{Z} = \frac{(p_1 - p_2)}{SE}$.

Note that the null hypothesis that the two groups are exactly equal in some parameter is false in real data (Cohen, 1994) and not really an interesting hypothesis; we care more about whether an estimated difference is large enough to be practically meaningful. So, the fact that a parameter difference is statistically significant is not necessarily a sign of meaningful DIF. If a difference is statistically significant, we should go on to consider whether the difference is large enough to be a cause for concern. Steinberg and Thissen (2006) suggested that a difference in the traditional IRT *b* parameter larger than .5 is practically meaningful. They do not make a suggestion for the *a* parameter. However, they do state that "[f]or many purposes, graphical displays of trace lines and/or their differences are easier to interpret than the parameters themselves" (p. 406). To convey the practical importance of differences in both difficulties and

discriminations, the test characteristic curves (TCCs) of the entire student groups were obtained for each of the five content strands and displayed in graphical form. The TCC gives the expected score on the content strand for a given level of the underlying construct. Therefore, the TCC gives a direct presentation of DIF effect size on the scale of the test (number correct). We can obtain the TCC of a given content strand by adding the item response functions of each item in the strand (Equation 1).

NOHARM Data Analysis.

In the first no-calculator sample, a unidimensional 3-parameter IRT model did not fit the data well. There were several Heywood cases (i.e., items with factor loadings estimated to equal 1). It is unclear why these items are Heywood cases; there is nothing distinctive about them in terms of difficulty or content strands. Since it is unlikely that any item is a perfect indicator of a construct, Heywood cases indicate that there is a problem. In this situation, the problem might be that many examinees obtained scores below 7 (Figure 1), which would be roughly the expected score if an examinee responded randomly to the multiple-choice items. This implies that low-ability examinees were systematically attracted to the incorrect responses. Whenever this happens, it might not be appropriate to include a lower-asymptote parameter, which assumes that examinees with very low ability always have some chance of responding correctly.

A unidimensional 2-parameter IRT model for the first no-calculator sample was tried instead. This model did not produce any Heywood cases. The Tanaka goodness-of-fit index was .987, indicating that the model fit the data well. A unidimensional 2-parameter IRT model was then fit to the data of the first calculator sample, and this was also successful (Tanaka GFI = .984). Good fits were also obtained in the cross-validation no-calculator sample (Tanaka GFI = .988) and in the cross-validation calculator sample (Tanaka GFI = .982).

Structural Equation Modeling.

Because the NOHARM approach to IRT does not provide statistical tests of any global hypothesis (e.g., measurement invariance between the two groups), the items were grouped into five parcels according to the five content strands and fit to an SEM multiple indicators and multiple causes (MIMIC) model with the computer program LISREL 8.80 using maximum likelihood estimation on covariance matrices. The grouping into parcels was necessary in order to create suitable indicators for SEM analysis. Since the unidimensional model is appropriate for the entire set of items, grouping the items into parcels on the basis of item-content domain specification is also appropriate. In the same way that IRT uses single items as indicators of a common factor, the structural equation modeling (SEM) approach to factor analysis uses continuous (or approximately continuous) test or parcel subscores as indicators of a common factor.

The score ranges for the five content strands in this data set are: 0-11 for Algebra, 0-10 for Number Sense, 0-3 for Measurement, 0-3 for Geometry, and 0-7 for Statistics.

A battery of test subscores shows a failure of measurement invariance across two groups if (1) the two populations have different intercepts in the regression of subscores on the common factor, or (2) the two populations have different slopes (factor loadings) in this regression. Lack of measurement invariance in the SEM framework is therefore analogous to DIF in the IRT framework. The purpose of a MIMIC factor model in SEM is to examine whether different groups have the same intercepts in the regression of subscores on the common factor. It may be helpful to consult Figure 8 while reading the following explanation. The MIMIC model can be written as

$$X_j = \alpha_j + \delta_j G + \lambda_j \theta + \varepsilon_j,$$

in which X_j denotes the *j*th parcel score, α_j denotes the intercept, δ_j denotes the difference in intercepts between the groups, *G* denotes the group dummy variable, λ_j denotes the factor loading, θ denotes factor score, and ε_j denotes the error. A MIMIC model uses a path from a grouping variable (i.e., calculator vs. no-calculator) to the common factor to account for the difference between groups in the average level of the factor. If a path going directly from the grouping variable to a subscore is necessary for good model fit, then the two groups have different intercepts in the regression of that subscore on the common factor. In other words, even if members of two groups have the same level of the factor, the member of the group with the higher intercept will have a higher expected subscore. In terms of the equation above, a significant path from the grouping variable to the *j*th parcel score means that δ_j is not equal to zero.

Results

If the construct measured by the test has the same distribution in two groups, and if there is no DIF, then in a graph plotting the factor loadings in one group against the factor loadings in the other group, the points should lie on a straight line through the origin with a slope of 1. Note that the two samples had identical observed-score distributions as a result of the stratified sampling. Figure 2 shows such a graph for the first no-calculator and calculator samples. Figure 3 shows the corresponding graph for the cross-validation samples.

INSERT FIGURE 2 ABOUT HERE

INSERT FIGURE 3 ABOUT HERE

The fit to a straight line through the origin with a slope of 1 looks fairly good overall in both the original and cross-validation samples. In the first samples, a linear regression of the calculator group's factor loadings on the no-calculator group's factor loadings gives an intercept of -.001 and a slope of .965. In the cross-validation samples, the linear regression gives an intercept of .012 and a slope of .983. This good fit to a straight line through the origin with a slope of 1 means that there is not much overall DIF with respect to the factor loadings (item discriminations in the IRT framework) between the no-calculator and calculator groups.

There are some signs of potentially noteworthy DIF with respect to factor loading in a few items. For instance, in both the first and cross-validation samples, items 10 and 28 seem to deviate rather far from the line. These items appear to be less discriminating in the calculator group.

INSERT FIGURE 4 ABOUT HERE

Figure 4 plots the thresholds in the first calculator sample against the thresholds in the first no-calculator sample. The area under the curve of the standard normal distribution, from negative infinity to the threshold parameter, gives the percentage of examinees getting the item correct. This means that larger thresholds correspond to easier items. Again, if the two groups have the same distribution of the ability measured by the test, and if there is no DIF, then the points should lie on a straight line through the origin with a slope of 1. The fit to the straight line, once again, looks fairly good overall, suggesting there is not much overall DIF with respect to thresholds, either. The cross-validation samples support this conclusion (Figure 5).

INSERT FIGURE 5 ABOUT HERE

In the first samples, a linear regression of the calculator group's thresholds on the no-calculator group's thresholds gives an intercept of -.029 and a slope of .892. In the cross-validation samples, the linear regression gives an intercept of -.043 and a slope of .831. The slopes are less than 1 because more items are on the right side of the line than on the left side; this means that the items tend to be slightly easier for the no-calculator group. A possible reason for this trend will be discussed later. In addition, it is clear that items 2 and 17 are unacceptably far from the line and show substantial DIF. These items are easier for the calculator group.

Table 2 presents the numerical results of the DIF analysis. Only items with a parameter showing statistically significant DIF in both the original and cross-validation samples are shown in Table 2. The only significant group differences in factor loadings replicated in the cross-validation samples were found for items 10 and 28. These items were already noted to be possible outliers in Figures 2 and 3. The other items do not show consistent evidence of strong DIF with respect to factor loadings.

There were many significant group differences in thresholds. Most of these were small differences favoring the no-calculator group. The few differences favoring the calculator group are worrisome because they tended to be much larger. Using the rule of thumb that flags a difference in the traditional IRT *b* parameter exceeding .5, we find that items 2, 6, and 17 showed practically meaningful DIF with respect to difficulty. As expected from Figures 4 and 5, these three items showed substantial differences favoring the calculator group.

Content Analysis and Hypotheses Regarding Sources of DIF

As discussed previously, items 2 and 17 tended to be easier for the calculator group. These items belonged to the content strand Number Sense and Operations. According to Table 2, there were four other items belonging to the content strand Number Sense and Operations that tended to be easier for the calculator group. A closer look at items 2 and 17 indicated that they focus on computation skills. Accordingly, students with accommodation may have had an advantage on these types of items. Item 6 belonged to the content strand Geometry. This item could have been reduced to a simpler computational problem since the formula for the Pythogorean theorem was provided on the reference sheet.

Items 10 and 28 showed evidence of DIF with respect to item factor loadings. These items discriminated among students with accommodation less effectively than among students taking the test under normal conditions. Item 10 belonged to Number Sense and Operations, and item 28 belonged to Patterns, Relations, and Algebra. The reasons for DIF are less clear for these items.

The substantial DIF for the Number Sense and Operations items may explain why the other items in the test tended to be somewhat easier for the no-calculator group. If a few Number Sense items show strong DIF with respect to thresholds favoring the calculator group, then the

no-calculator students must have been somewhat more able overall, because they were matched with calculator students according to sum scores. Thus, the no-calculator group would have found many of the items slightly easier.

Figure 6 shows the test characteristic curves (TCCs) for the five content strands for the whole student population.

INSERT FIGURE 6 ABOUT HERE

The TCCs of the no-calculator and calculator groups were very similar for each content strand. As expected from the analysis of the individual items, the difference in the TCC between the two groups was greatest for Number Sense and Operations (this difference was consistent across the two samples). On this content strand, the calculator group showed a noticeable advantage. The calculator group also showed a noticeable advantage on Geometry at the higher end of the ability range; this is perhaps consistent with the fact that one Geometry item was substantially easier for the calculator group. But even on these two content strands, the expected scores of the two groups never differed by more than a small fraction of a score point. For example, for Number Sense and Operations, the maximum difference between the two TCCs is only 0.07 score points. This suggests that the statistically significant cases of DIF in the other three content strands are not practically meaningful.

MIMIC Model

A MIMIC model was fit to test the global hypothesis that there is a failure of measurement invariance between the two groups in Number Sense and Operations due to the use of calculators. Appendix A contains the LISREL syntax for the MIMIC models.

Note that the MIMIC model does not test for the differences in factor loadings. By using a MIMIC model, we are assuming that the factor loadings describing the relationship between

the construct and its indicators are essentially the same for the two groups. In other words, we are assuming that there is only uniform DIF, which means that one group always has a higher expected score across the entire range of ability. Therefore if any differences in factor loadings exist, then MIMIC is not an appropriate model for testing measurement invariance. However, our DIF analysis did not find large differences in item factor loadings. Figure 6 confirms that the expected scores on the content strands increase with higher levels of the construct in similar ways across the two groups.

In all of the analyses below, the path from the mathematics achievement factor to Algebra was fixed to one in order to scale the factor. Algebra was chosen because its relationship to the math factor is strong (Figure 6).

First, a MIMIC model was fit without a path from the group variable to Number Sense (Figure 7).

INSERT FIGURE 7 ABOUT HERE

The model fit was acceptable: RMSEA = .048 with 90% C.I. (.041, .055), CFI = .99, SRMR = .021, and GFI = .99.

To determine whether the alternative hypothesis fits better, a MIMIC model was fit to the data adding a path from the group variable to Number Sense (Figure 8).

INSERT FIGURE 8 ABOUT HERE

The fit for this model to the data was improved: RMSEA = .023 with 90% C.I. (.015, .031), CFI = 1, SRMR = .0097, and GFI = 1), and the $\Delta \chi^2$ between the two models was statistically significant (χ^2=102.29, Δdf =1, $p < .05$). In addition, ΔCFI = .01 indicated that this improvement in fit was meaningful. Therefore, the model incorporating a lack of invariance for the Number Sense strand fit substantially better than the model assuming measurement invariance across all

strands. The negative path coefficient from the group indicator to the construct of math (γ_{11} = -.14, p = .006) and the positive path coefficient from the group indicator to Number Sense (γ_{21} = .43, p < .05) indicate that the calculator group performed worse on content strands *other* than Number Sense. These path coefficients are unstandardized. We can look at the differences between the two groups in units of standard deviations. The advantage of the no-calculator group over the calculator group in the math achievement factor was approximately .08 standard units, while the advantage of the calculator group in Number Sense was approximately .21 standard units. As stated above, because students in the no-calculator group were matched to those in calculator group on the basis of on sum scores, if the calculator group performed better on average in Number Sense, then the no-calculator group had to perform better in other areas to end up with matched sum scores[2].

Therefore, the MIMIC analysis strongly supports the hypothesis that the use of calculators has led to a failure of measurement invariance between the two groups in Number Sense and Operations.

Discussion

In this study, we examined whether the calculator accommodation changed the measurement properties of an 8th-grade mathematics test. Overall, there was very little difference in factor loadings between the no-calculator and calculator groups. This suggests that differences

[2] To ensure that this result was not obtained by chance, the analyses above were replicated with the cross-validation samples. A MIMIC model was fit without the path from the group variable to Number Sense. The result was similar to the one above. The model fit was generally good: RMSEA = .047 with 90% C.I. (.040, .054), CFI = .99, SRMR = .022, and GFI = .99. Then the MIMIC model including a path from the group variable to Number Sense was fit. The fit of this model was again impressive: RMSEA = .022 with 90% C.I. (.014, .030), CFI = 1, SRMR = .011, and GFI = 1). The $\Delta\chi^2$ between the two models was 100.24 and Δdf was 1 (p < .05). The change in CFI was again .01.

among students *within* either the no-calculator and calculator groups have similar meanings. That is, if students A and B come from the same group and differ by *x* score points, then this difference implies roughly the same difference in mathematics achievement no matter which group A and B come from. However, if students A and B come from different groups, then the interpretation of their score difference is complicated. When the examinee is allowed to use a calculator, the test no longer seems to measure one of the content strands in the way that it does for students without the calculator. Specifically, the accommodated group performed relatively better on average in Number Sense and Operations, so it appears that a calculator makes items of this type easier. As a result, the use of calculator as a form of accommodation may not allow valid comparisons of mathematics achievement in this content strand between students with learning disabilities and students taking the test under normal conditions. In other words, the failure of measurement invariance does not allow us to infer that the students from different groups with comparable sum scores have equal levels of the construct measured by the test. Comparisons across the students with disabilities granted the calculator accommodation and students taking the tests without a calculator should be limited to only the other content strands.

This study showed the failure of measurement invariance using two levels of analysis—item level and parcel level, using two statistical procedures (nonlinear factor analysis and structural equation modeling). Using nonlinear factor analysis to compare the test characteristic curves of the two groups presents a theoretically sound way to investigate measurement invariance. The use of the MIMIC model supplemented this analysis with tests of overall hypotheses. These two procedures provide a comprehensive means for investigating measurement invariance across standard and accommodated test administration conditions.

References

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49: 997-1003.*

Fraser, C., & McDonald, R. P. (2003). *NOHARM Version 3.0 user's guide.*

Gotzmann, A.J. (2001). The effect of large ability differences on type I error and power rates using SIBTEST and TESTGRAF DIF detection procedures. Unpublished Master's Thesis, University of Alberta, Edmonton, Alberta, Canada.

Gotzmann, A.J., & Boughton, K.A. (2004). *A comparison of Type I error and power rates for the Mantel-Haenszel and SIBTEST procedures when group differences are large and unbalanced.* Paper presented at the American Educational Research Association (AERA) in San Diego, California, April 2004.

Gierl, M.J., &Gotzmann, A., & Boughton, K.A. (2004). Performance of SIBTEST when the percentage of DIF items is large. *Applied Measurement in Education, 17,* 241-264.

Jöreskog, K.G. & Sörbom, D. (2007). *LISREL 8.80 for Windows* [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

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

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

R Development Core Team (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sireci, S.G., & Wells, C.S. (2009). *Evaluating the comparability of English and Spanish video accommodations for English language learners.* Washington, DC: Council of Chief State School Officers.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11*(4), 402-415.

Table 1

Test Specifications: The Statewide Grade 8 Mathematics Items

Content Strand	Percentage (+/- 5%)	Total Number of Points
Number Sense and Operations	26 %	14
Patterns, Relations, and Algebra	28 %	15
Geometry	13 %	7
Measurement	13 %	7
Data Analysis, Statistics and Probability	20 %	11

Table 2
DIF Analysis: The Statewide Grade 8 Mathematics Items

Item	Content Strand	τ_{NC}	τ_C	λ_{NC}	λ_C	$\tau_{NC} - \tau_C$	$\lambda_{NC} - \lambda_C$
2	Number Sense and Operations	-.366	.042	.482	.468	**-.408***	.014
		-.380	.057	.512	.550	**-.437***	-.038
3	Measurement	-.079	-.192	.492	.508	**.113**	-.016
		-.060	-.184	.484	.548	**.124**	**-.064**
6	Geometry	-.307	-.185	.112	.128	**-.122***	-.016
		-.274	-.185	.118	.184	**-.089***	**-.066**
8	Measurement	-1.046	-.885	.532	.565	**-.161**	-.033
		-1.023	-.912	.504	.597	**-.111***	**-.093**
10	Number Sense and Operations	.135	.162	.299	.226	-.027	**.073**
		.186	.104	.301	.246	**.082**	**.055**
11	Number Sense and Operations	-.445	-.316	.353	.350	**-.129**	.003
		-.509	-.373	.373	.394	**-.136**	-.021
14	Number Sense and Operations	-.253	-.157	.390	.436	**-.096**	-.046
		-.294	-.142	.411	.447	**-.152**	-.036
15	Patterns, Relations, and Algebra	-.222	-.409	.620	.532	**.187**	**.088**
		-.239	-.355	.602	.601	**.116**	.001
17	Number Sense and Operations	-.569	-.199	.467	.405	**-.370***	**.062**
		-.547	-.226	.451	.405	**-.321***	.046
19	Number Sense and Operations	-1.168	-1.060	.551	.416	**-.108**	**.135**
		-1.270	-1.064	.530	.495	**-.206**	.035
22	Number Sense and Operations	-.132	-.228	.481	.456	**.096**	.025
		-.138	-.228	.413	.460	**.090**	-.047
25	Patterns, Relations, and Algebra	-.231	-.296	.355	.353	**.065**	.002
		-.234	-.332	.408	.338	**.098**	**.070**
27	Number Sense and Operations	-.048	-.164	.532	.475	**.116**	**.057**
		-.057	-.150	.543	.510	**.093**	.033
28	Patterns, Relations, and Algebra	-.140	-.157	.543	.412	.017	**.131**

		-.144	-.162	.526	.450	.018	**.076**
31	Data Analysis, Statistics, and Probability	-.369	-.569	.557	.460	**.200***	**.097**
		-.397	-.548	.543	.548	**.151**	-.005
33	Patterns, Relations, and Algebra	.053	-.030	.401	.369	**.083**	.032
		0.068	-.016	.395	.397	**.084**	-.002

Note. τ stands for the threshold parameter and λ for the factor loading. The subscript *NC* stands for the no-calculator group and *C* for the calculator group. Statistically significant differences between the no-calculator and calculator groups at the .05 level are in bold. Threshold differences marked with asterisks correspond to differences in the traditional IRT *b* parameter that exceed .5.

Figure Captions

Figure 1. Histograms of the Grade 8 Mathematics sum scores for (a) the no-calculator group and (b) the calculator group.

Figure 2. An illustration of the MIMIC model.

Figure 3. Factor loadings in the first calculator group plotted against the factor loadings in the first no-calculator group. A line through the origin with a slope of 1 has been superimposed.

Figure 4. Factor loadings in the cross-validation calculator group plotted against the factor loadings in the cross-validation no-calculator group.

Figure 5. Thresholds in the first calculator group plotted against the thresholds in the first no-calculator group. A larger threshold indicates that the item is easier.

Figure 6. Test characteristic curves for the five content strands. "Algebra" stands for the content strand Patterns, Relations, and Algebra. "Measurement" stands for the content strand Measurement. "Geometry" stands for the content strand Geometry. "Statistics" stands for the content strand Data Analysis, Statistics, and Probability. "Number Sense" stands for the content strand Number Sense and Operations.

Figure 7. MIMIC model without a path from the group variable to Number Sense. The path coefficients are unstandardized, and the numbers in parentheses are standard errors. Calculator Group was dichotomously coded: 1 for the calculator group and 0 for the non-calculator group. The numbers in parentheses indicate the standard errors of the path coefficients.

Figure 8. MIMIC model with a path from the group variable to Number Sense. The path coefficients are unstandardized, and the numbers in parentheses are standard errors.

Figure 1

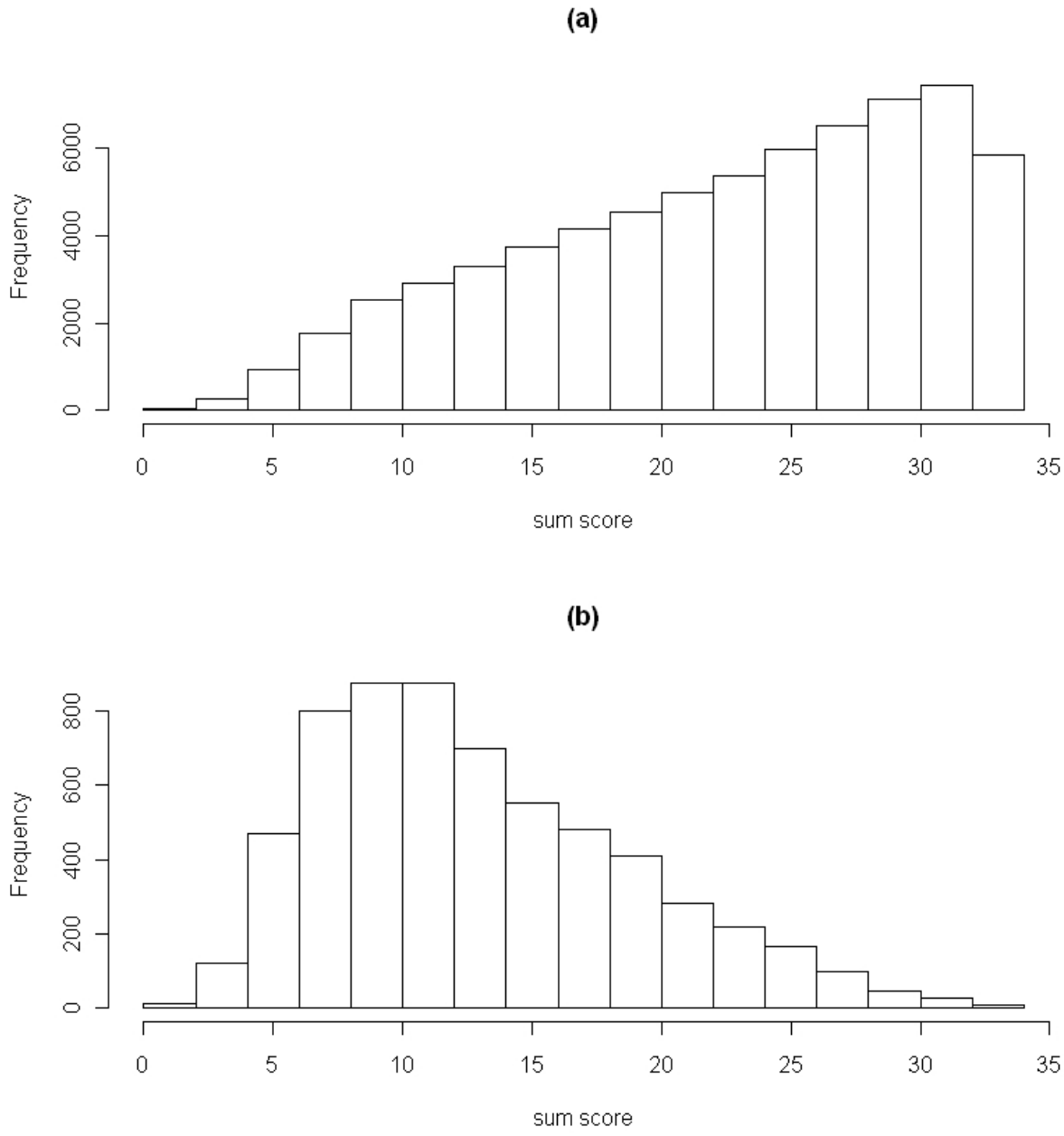

(a)
Frequency
0
2000
4000
6000
0
5
10
15
20
25
30
35
sum score
(b)
Frequency
0
200
400
600
800
0
5
10
15
20
25
30
35
sum score

Figure 2

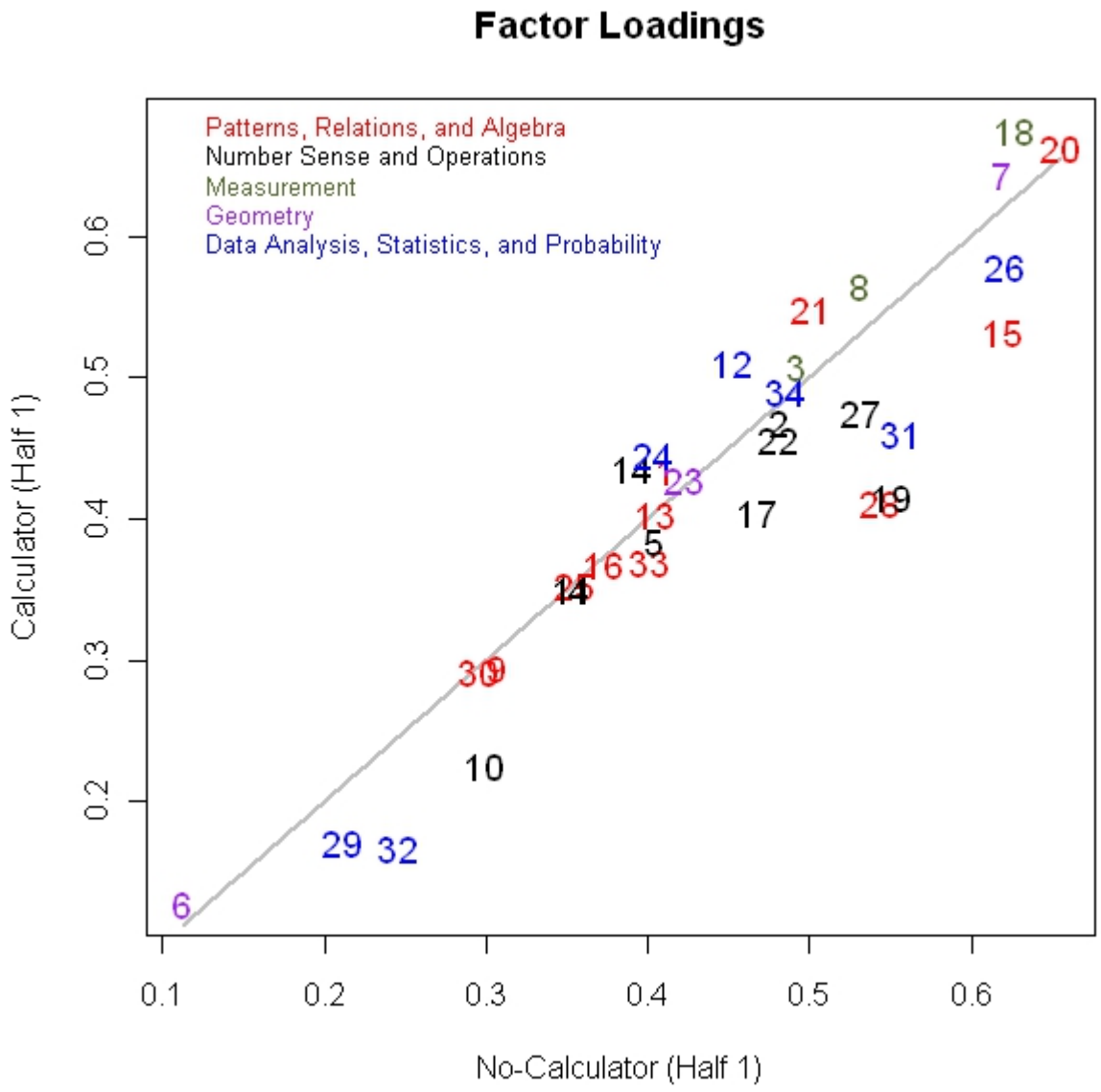

Factor Loadings
Patterns, Relations, and Algebra
Number Sense and Operations
Measurement
Geometry
Data Analysis, Statistics, and Probability
Calculator (Half 1)
No-Calculator (Half 1)
0.1
0.2
0.3
0.4
0.5
0.6

Figure 3

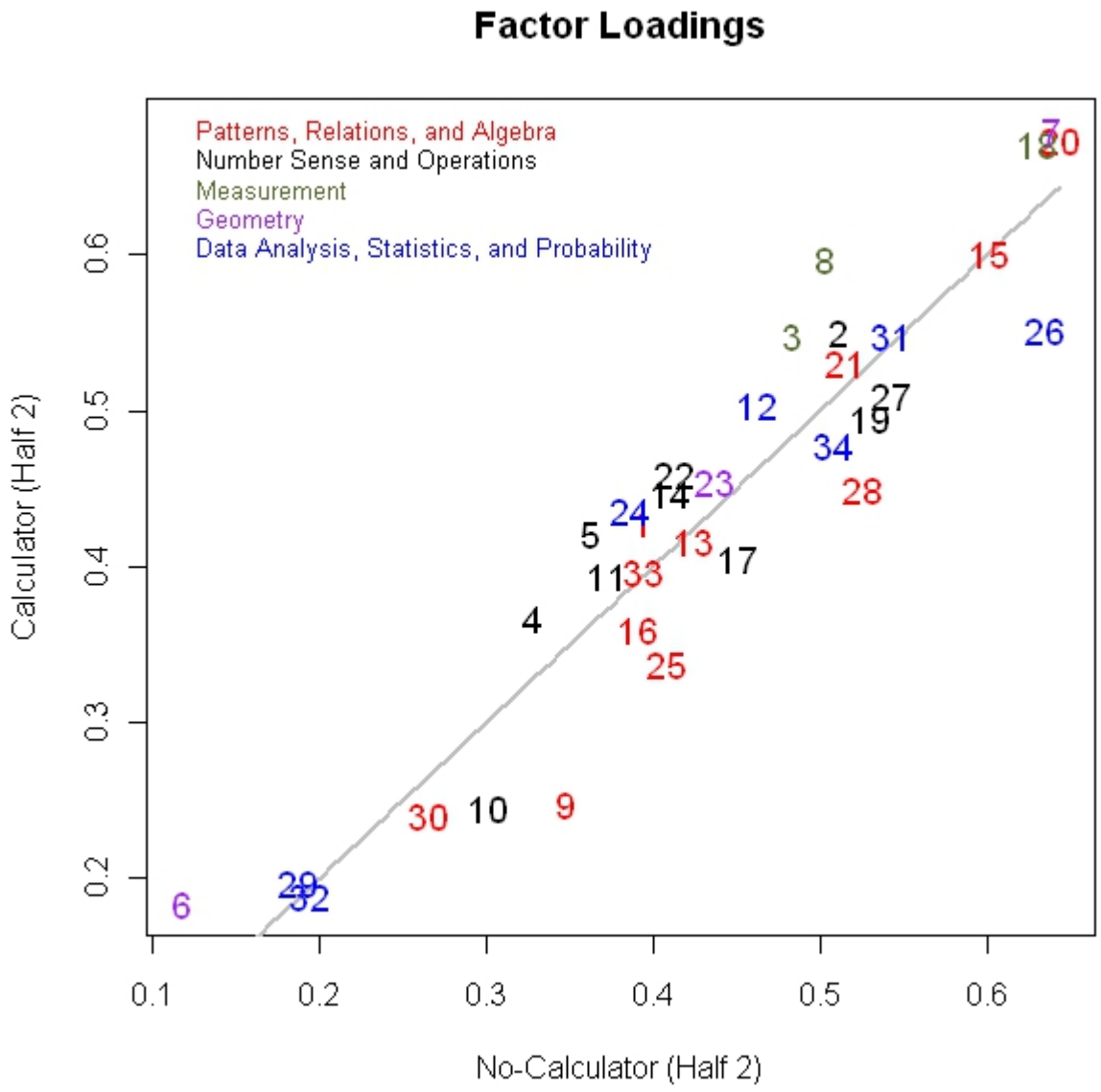

Factor Loadings
Patterns, Relations, and Algebra
Number Sense and Operations
Measurement
Geometry
Data Analysis, Statistics, and Probability
Calculator (Half 2)
No-Calculator (Half 2)
0.1
0.2
0.3
0.4
0.5
0.6

Figure 4

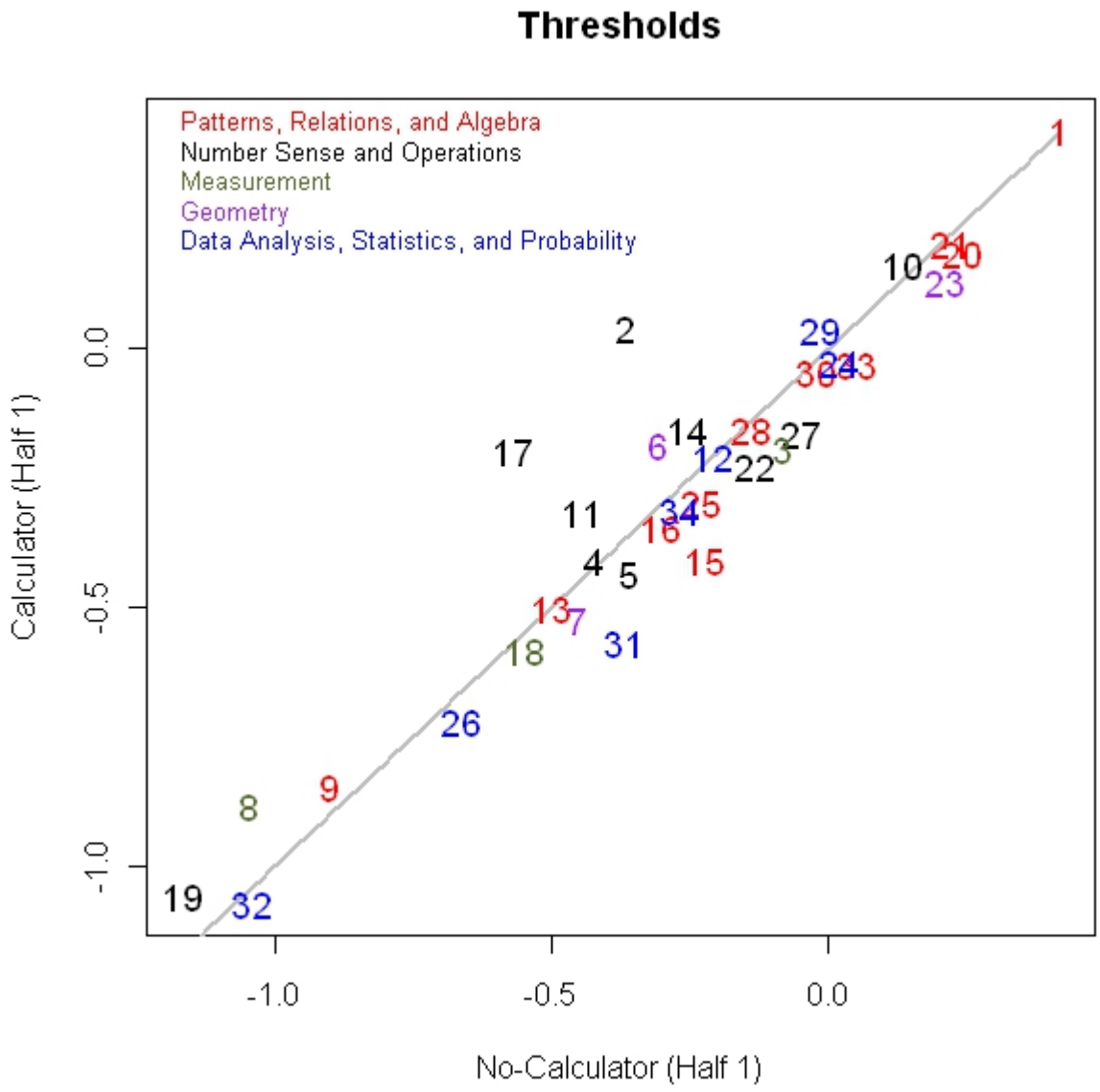

Thresholds
Patterns, Relations, and Algebra
Number Sense and Operations
Measurement
Geometry
Data Analysis, Statistics, and Probability
Calculator (Half 1)
No-Calculator (Half 1)
0.0
-0.5
-1.0
-1.0
-0.5
0.0

Figure 5

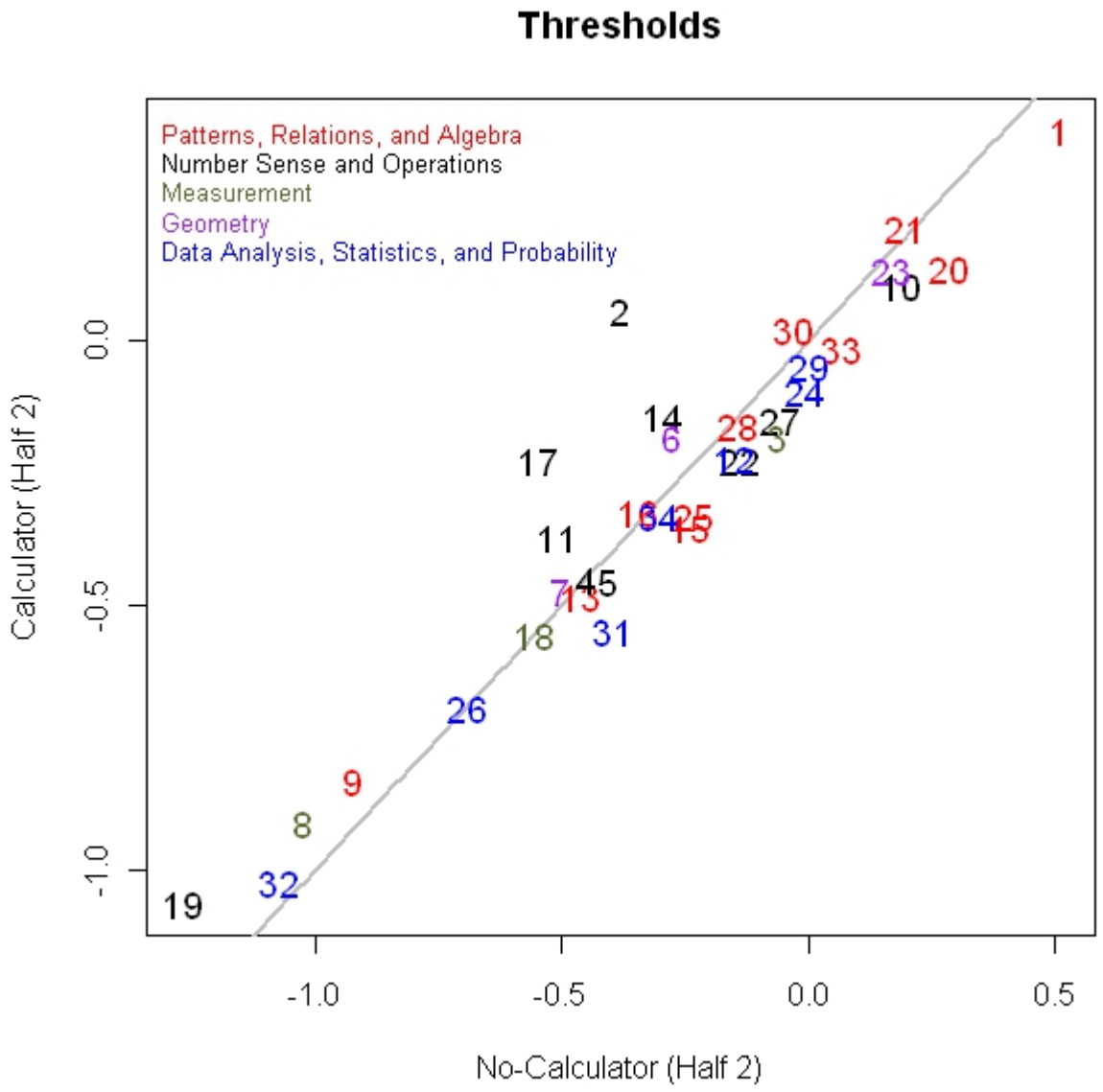

Thresholds
Patterns, Relations, and Algebra
Number Sense and Operations
Measurement
Geometry
Data Analysis, Statistics, and Probability
Calculator (Half 2)
No-Calculator (Half 2)
-1.0
-0.5
0.0
0.5

Figure 6

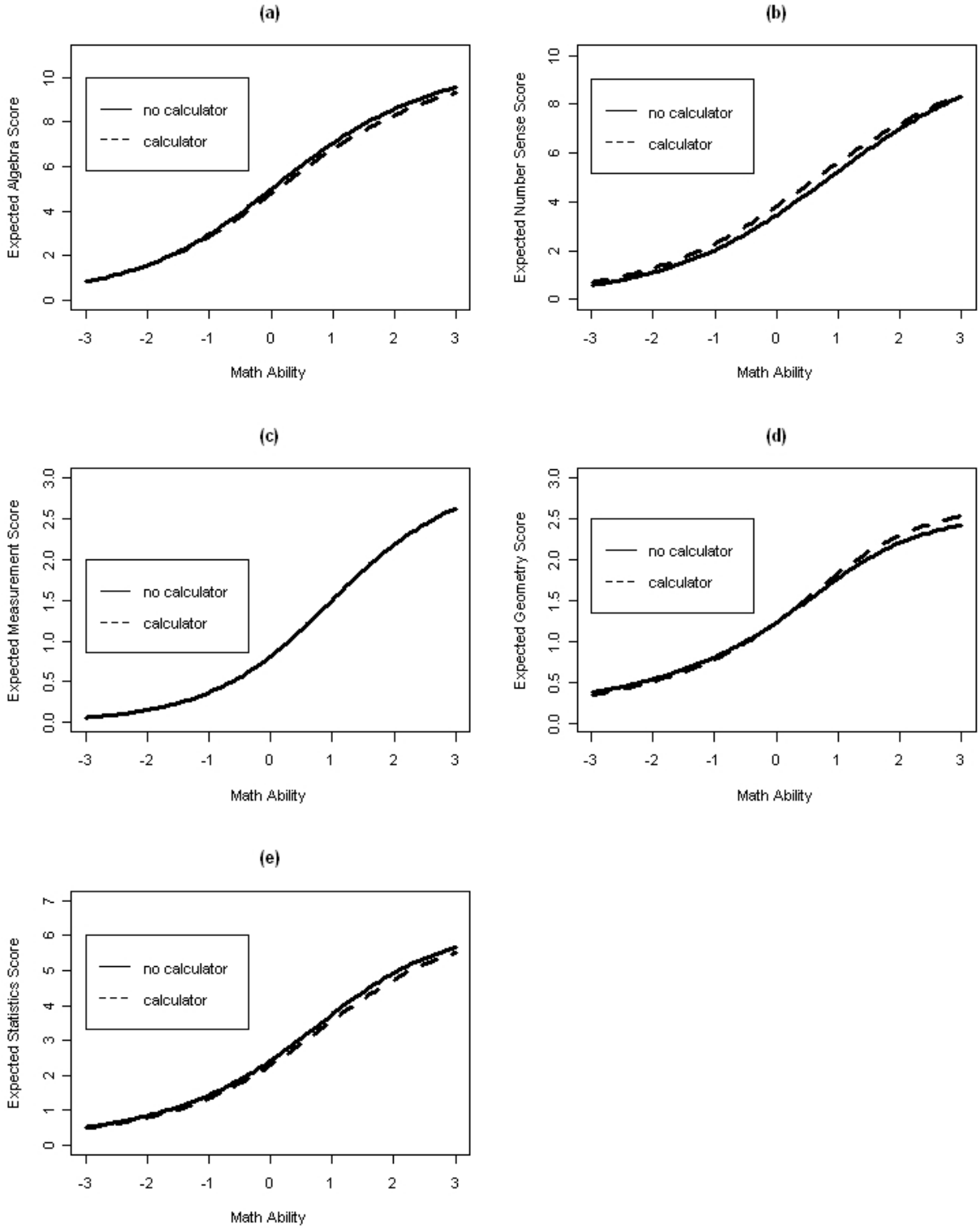
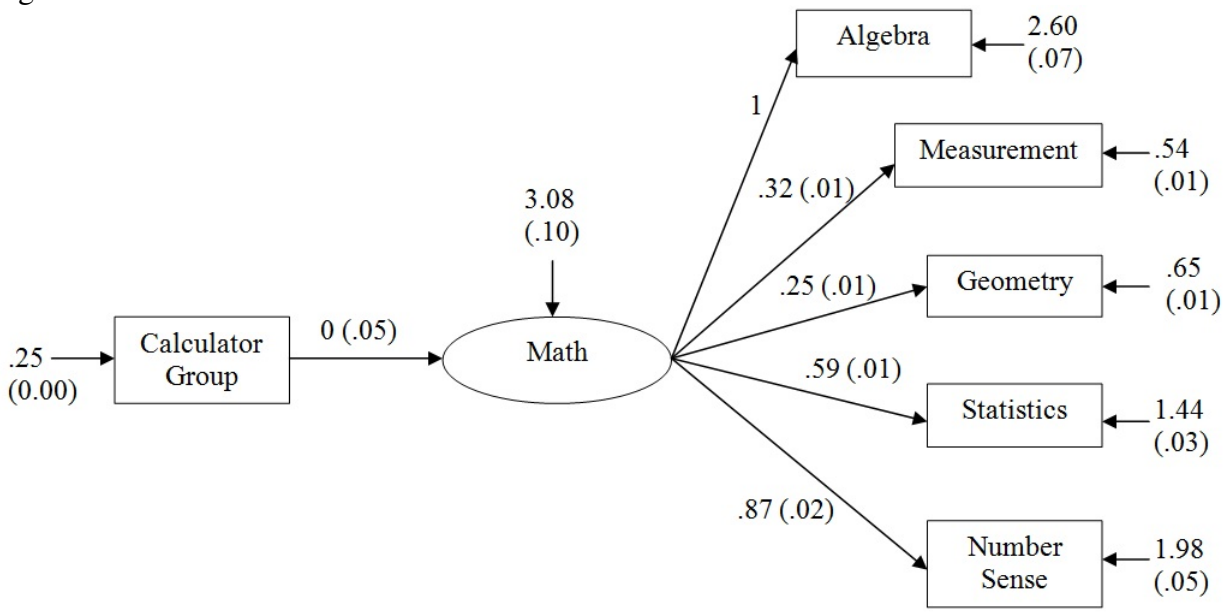
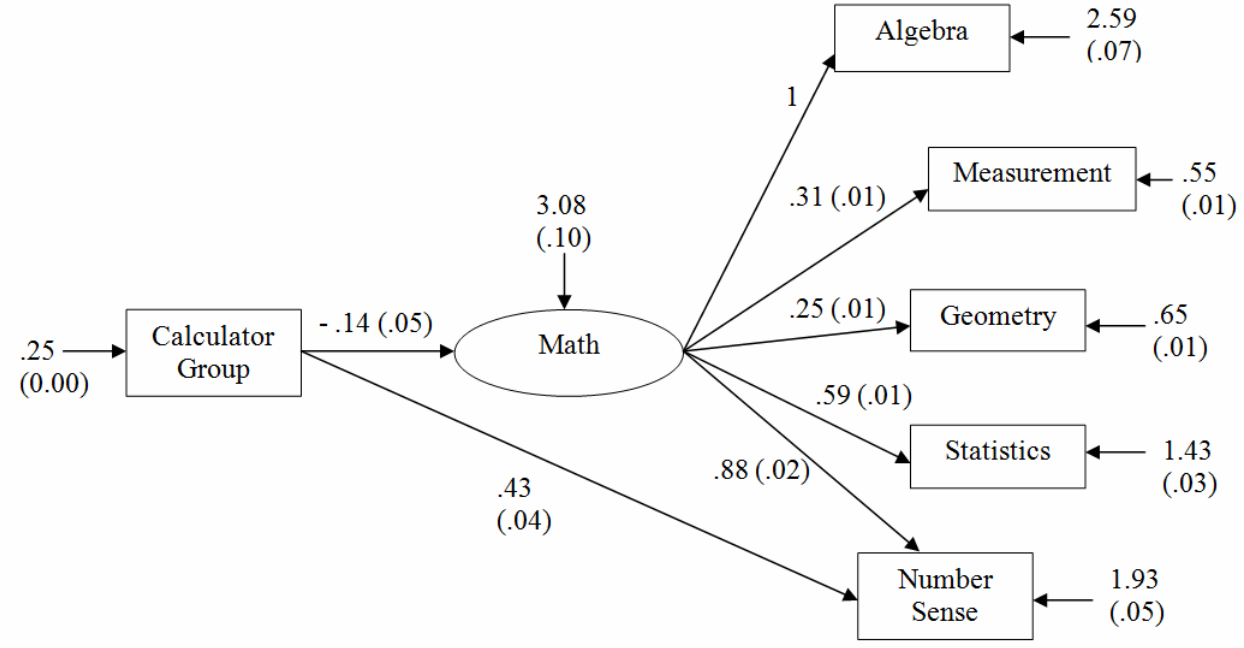

(a)
Expected Algebra Score
no calculator
calculator
Math Ability
(b)
Expected Number Sense Score
no calculator
calculator
Math Ability
(c)
Expected Measurement Score
no calculator
calculator
Math Ability
(d)
Expected Geometry Score
no calculator
calculator
Math Ability
(e)
Expected Statistics Score
no calculator
calculator
Math Ability

Figure 7



χ^2 = 135.55, df = 9, RMSEA = .048

Figure 8



x^2 = 33.26, df = 8, RMSEA = .023

Appendix A

LISREL Syntax of the MIMIX models fitted for the first half (Half1) and cross-validation (Half2) groups

LISREL syntax for the MIMIC model without a path from the group variable to Number Sense for Half1

```
Half1
DA NI=6 NO=6140 MA=CM
LA
Alg NumSen Meas Geom Stats CalcGP
CM SY
5.67718992
2.71464904  4.32624002
0.92411172  0.87003581  0.850659610
0.75125003  0.66858879  0.255003494  0.83864997
1.86187203  1.52325032  0.578226886  0.44896481  2.49831167
-.04455123  0.07664115 -0.004235217 -0.00301352 -0.02842482  0.250040723
MO NY=5 NX=1 BE=FU,FI NE=1 GA=FU,FI PS=SY,DI LY=FI TE=FI
LE
Math
FR GA(1,1)
FR LY(2,1) LY(3,1) LY(4,1) LY(5,1)
FR TE(1,1) TE(2,2) TE(3,3) TE(4,4) TE(5,5)
VA 1 LY(1,1)
PD
OU ME=ML TV SE MI RS
```

LISREL syntax for the MIMIC model with a path from the group variable to Number Sense for Half1

```
Half1
DA NI=6 NO=6140 MA=CM
LA
Alg NumSen Meas Geom Stats CalcGP
CM SY
5.67718992
2.71464904  4.32624002
0.92411172  0.87003581  0.850659610
0.75125003  0.66858879  0.255003494  0.83864997
1.86187203  1.52325032  0.578226886  0.44896481  2.49831167
-.04455123  0.07664115 -0.004235217 -0.00301352 -0.02842482  0.250040723
MO NY=5 NX=1 BE=FU,FI NE=2 GA=FU,FI PS=SY,DI LY=FI TE=FI
LE
Math NumSen
```

```
FR GA(1,1) GA(2,1)
FR BE(2,1)
FR LY(3,1) LY(4,1) LY(5,1)
FR TE(1,1) TE(3,3) TE(4,4) TE(5,5)
VA 1 LY(1,1) LY(2,2)
PD
OU ME=ML TV SE MI RS
```

LISREL syntax for the MIMIC model without a path from the group variable to Number Sense for Half2

```
Half2
DA NI=6 NO=6140 MA=CM
LA
Alg NumSen Meas Geom Stats CalcGP
CM SY
5.69198102
2.75695619  4.40075270
0.98160763  0.89872503  0.821543472
0.82414429  0.72103596  0.276924432 0.853324158
1.84342188  1.56057252  0.608241005 0.505543137  2.49367872
-0.04243362  0.07867731 -0.005701254 0.007493077 -0.03445187  0.250040723
MO NY=5 NX=1 BE=FU,FI NE=1 GA=FU,FI PS=SY,DI LY=FI TE=FI
LE
Math
FR GA(1,1)
FR LY(2,1) LY(3,1) LY(4,1) LY(5,1)
FR TE(1,1) TE(2,2) TE(3,3) TE(4,4) TE(5,5)
VA 1 LY(1,1)
PD
OU ME=ML TV SE MI RS
```

LISREL syntax for the MIMIC model with a path from the group variable to Number Sense for Half2

```
Half2
DA NI=6 NO=6140 MA=CM
LA
Alg NumSen Meas Geom Stats CalcGP
CM SY
5.69198102
2.75695619  4.40075270
0.98160763  0.89872503  0.821543472
0.82414429  0.72103596  0.276924432 0.853324158
1.84342188  1.56057252  0.608241005 0.505543137  2.49367872
-0.04243362  0.07867731 -0.005701254 0.007493077 -0.03445187  0.250040723
MO NY=5 NX=1 BE=FU,FI NE=2 GA=FU,FI PS=SY,DI LY=FI TE=FI
LE
Math NumSen
FR GA(1,1) GA(2,1)
```

```
FR BE(2,1)
FR LY(3,1) LY(4,1) LY(5,1)
FR TE(1,1) TE(3,3) TE(4,4) TE(5,5)
VA 1 LY(1,1) LY(2,2)
PD
OU ME=ML TV SE MI RS
```