An Analysis of a Ninth Grade Mathematics Intervention

Tracy Sweet

June 11, 2010

Abstract

Many ninth grade students are required to take Algebra I, and extra support is needed for underprepared students. A common solution is to double the amount of mathematics instruction but educators do not agree on how to use the extra time. One common intervention is to enroll these students in a year long Algebra I course, where they are taught Algebra during these double class periods. We compare this with offering a catch-up course on basic skills, Transition to Advanced Mathematics (TAM) taught during the first half of the year, followed by a typical Algebra I course during the second half of the year. Preliminary analyses suggest students in the TAM condition outperform the control condition on general math skills but students in each condition do not differ in performance on an algebra test even though TAM students have had only half of the exposure to Algebra.

1 Introduction

In recent years, the passing of No Child Left Behind (NCLB) has mandated states and districts to reexamine their educational practices. One goal of NCLB is to improve accountability of schools, to ensure that a high school diploma certifies that a student has not only satisfied various criteria but has the skills and knowledge necessary to be successful following school.

The National Assessment of Educational Progress (NAEP) suggests that students in 12th grade with a *basic* level of understanding in mathematics be able to perform arithmetic and basic geometry, understand sample statistics and probability, use and manipulate expressions and solve problems with linear equations represented in a variety of ways. These are the topics covered in most Algebra I courses. In 2005, the most recent administration of the survey assessment, 61% of all students tested at or above this *basic* level, but only 30% of black students test at or above this level.

To prepare students for post-secondary work, many districts require all students to take college preparatory mathematics; some districts only offer college preparatory level courses and some require these courses for graduation. For graduation in Maryland, students must earn a passing score on a series of state assessments which include an exam in Algebra I and is administered in ninth grade; thus, students take Algebra I regardless of prior coursework or achievement. The same is true for students in New York and California.

Unfortunately, many students are woefully unprepared, particularly in low income and urban districts. In a NAEP study of 11 urban school districts, only 53% of urban 8th graders tested at or above *basic* and in DC and Atlanta, only 31% achieved these levels. In fact, students are testing well below grade level. Neild and Balfanz (2006) found 54% of Philadelphia ninth grade students tested scored below a 7th grade level on a nationally normed math test.

The large gap between course standards and student achievement is especially disconcerting given that failing courses in ninth grade is a strong predictor of dropping out of high school (Neild et al. 2008). The ninth grade is so important that it is a central component of the Talent Development Model. Components of this reform model include creating a Ninth Grade Academy in which cohorts of students are taught by the same team of teachers. Students are supported by this small learning community and through catch-up courses in reading and mathematics and a seminar course to address social and study behaviors. Schools are not only restructured but the school day operates on a block schedule; students take four classes each day and each course is a semester long. Students in upper grades join similar small learning communities in the form of career academies.

There is some evidence that students benefit from attending Talent Development High Schools (TDHS) (Balfanz et al. 2004, Kemple et al. 2005). Kemple et al. (2005) found students in Philadelphia who attended Talent Development schools had higher achievement, rates of attendance, and Algebra I passing rates than similar students who did not attend Talent Development schools. One possible cause for student success in Algebra is the TDHS ninth grade mathematics curriculum, *Transition to Advanced Mathematics (TAM)* which precedes Algebra I. In several Baltimore city schools with dropout rates above 50%, Balfanz et al. (2004) found students enrolled in TAM courses had higher gains on standardized mathematics tests and felt they learned more mathematics than similar students not taking TAM courses.

Thus, researchers are interested in how the TAM curriculum, independent from the TDHS reform model, benefits under-prepared ninth grade students and whether it promotes success in Algebra I as most TAM students take Algebra I during the second semester of their ninth grade year. The principal investigators, Robert Balfanz and Ruth Neild, propose a group randomized trial to compare ninth grade students in a TAM/Algebra I sequence with students in a Algebra I sequence (Balfanz and Neild 2006). Before we describe the study in detail, we introduce the educational rationale behind each condition.

1.1 Transition to Advanced Mathematics

The purpose of Transition to Advanced Mathematics (TAM) is to improve conceptual understanding and skills needed to be successful in Algebra without being a repeat course of middle school mathematics. TAM is based on existing research of concepts and skills needed by underperforming ninth grade students, best teaching practices to these types of learners, feedback from classroom teachers about student misconceptions and knowledge gaps, and analysis of state/district standards. The National Research Council suggests that incoming ninth grades students struggle most with intermediate level topics, namely rational numbers and integers (Kilpatrick et al. 2001). Both of these topics are procedurally complicated in that whole number reasoning no longer applies, making them both conceptually challenging and also unintuitive (Stavy and Tirosh 2000). TAM attempts to address these issues in a relevant and rigorous way. Unit topics include rational numbers, integers, coordinate geometry, measurement, and functions.

Feedback from teachers of urban and underperforming students suggests that students need relevant topics and a variety of different activities that allow for movement and collaboration. In addition, teachers complain of lack of textbooks and materials and living circumstances that inhibit students carrying many books to and from school. Furthermore, teachers have little time for planning, and most new and inexperienced teachers lack the capacity to create elaborate activities and projects. As such, the TAM curriculum consists of bound, disposable student journals which

include all of the necessary printed material that students can easily carry back and forth to school. The teaching manual includes transparencies, cut-outs, grouping strategies for cooperative learning, hands-on activities and real-world examples. Furthermore, the TAM curriculum comes with a plethora of manipulatives and school supplies that are commonly unavailable in urban schools.

1.2 Stretch Algebra

As an alternative, many districts have created a stretch or double-dose algebra course for underprepared ninth graders. While these students still need some form of support to be successful in Algebra I, educators claim that students can gain the necessary concepts and skills while in the context of learning algebra. There is evidence that double-dosing can be successful among some ninth grade students. Nomi and Allensworth (2009) found that Chicago students in double-dose algebra courses increased their algebra test scores using regression discontinuity methods. They also found that double dosing was less effective for those students with the weakest math abilities.

2 Research Study

While there is evidence suggesting that each mode of instruction can be helpful to ninth grade students, this is the first study directly comparing the two methods (Balfanz and Neild 2006). Schools using the TAM curriculum are part of the TDHS reform model and evidence of student achievement could be due to the larger changes in the school and school environment. Furthermore, double-dose algebra has not been assessed on a national level and neither curriculum has been assessed in a randomized control study.

The purpose of this study is to directly compare a stretch algebra sequence with TAM followed by a traditional algebra course among ninth grade students who need additional curriculum support using a randomized control design. For the first year of the study, eight districts in six states were selected. Districts were required to volunteer an even number of participating schools and schools needed to have at least 75 underprepared first time ninth grade students. Within each district, schools were randomized to one of two treatments.

Students in the treatment group take TAM for 80-90 minutes each day for the first half of the school year followed by a traditional Algebra I course for 80-90 minutes each day during the second half of the school year (TAM/A1 condition). Students in the control condition take the district Algebra I course for 80-90 minutes each day for the entire school year (Stretch condition). All schools in the study have district-hired coaches to help them implement the curriculum as well as TDHS math facilitators to lead trainings and workshops. Teachers receive equal amounts of professional development in areas chosen by teachers or school leaders. Since TAM teachers receive supplies as part of the curriculum, stretch algebra teachers received gift cards to an educational supply company.

To measure student achievement in mathematics, students are tested in the fall, winter, and spring. In addition, students and teachers complete both beginning and end of year surveys. Other data include student and school level information provided by the district, teacher observations from the fall and spring, and facilitator reports about implementation. Research questions are given below.

3 Data

There are several sources of data; from the students directly we have test scores and survey responses, from the teachers we have survey responses, and from the district, student demographics and attendance, and these data vary in their state of use. We summarize only the data used in our analyses below (Table 1).

Table 1: Summary of Data

Student Level Covariates	N	Variable Type
Condition	3327	Binary
Race	2137	Discrete
Gender	2788	Binary
Grade	2520	Quantitative
Birth Year	2384	Quantitative
Free/Reduced Lunch	985	Binary
English Second Language	484	Binary
Fall CTBS	2399	Quantitative
Fall OH	2454	Quantitative
Winter CTBS	2208	Quantitative
Spring Alg TN	1930	Quantitative
Student Survey Responses	1683	Discrete

Teacher Level Covariates	N	Variable Type
Condition	83	Binary
Teacher Survey Responses	72	Mixed
Fall Teacher Observations	76	Mixed
Spring Teacher Observations	68	Mixed

4 Overall Goals

There are several outcome measures to investigate and we are interested in the effect of condition on each outcome measure. These include the California Test of Basic Skills (CTBS), which measures general math ability, given in the winter, the Algebra Terra Nova exam given in the spring, and student attitude measures. In addition to comparing across conditions, we are interested in how student level and teacher level variables affect each outcome measure.

5 Preliminary Analyses

5.1 EDA: Comparing Students Across Conditions

There is always concern with randomized trials that treatment groups will vary significantly across condition. Thus, we compare student demographics by condition (Table 2); more detailed analyses are included in Appendix A.1. Although differences at baseline may not affect outcome measures, we include them as we believe they may provide valuable information as to understanding why we do or do not see differences in achievement across condition. What is especially interesting is that missingness appears to vary by condition. We note that student covariates vary significantly by district as well (see Appendix A.2).

Table 2: Differences in Student Level Variables By Condition

Covariate	Statistical Test	p-value	Conclusion
Race w/missing data	χ^2	$p < 0.001$	TAM condition has higher proportion of black students, Stretch condition has higher proportion of hispanic students
Race w/o missing data	χ^2	$p < 1e-10$	TAM has higher proportion of missingness
Gender w/missing data	χ^2	$p < 1e-8$	TAM students have higher rates of missingness
Gender w/o missing data	χ^2	$p = 0.267$	Proportion of female students same across conditions
FRL w/missing data	χ^2	$p = 0.002$	TAM students have higher rates of missingness
FRL w/o missing data	χ^2	$p = 0.267$	No differences found between conditions
Fall CTBS	t-test	0.617	No differences found
Fall OH	t-test	0.097	No differences found

5.2 Teacher Surveys

Teachers were given brief surveys in the fall and spring during the study. The fall survey contains demographic and background items such as teaching experience and certification. The spring survey contains mostly items about teaching practices and beliefs. Unfortunately, there is a problem with the spring surveys as they were either not administered, not collected, or lost during the study. Although JHU facilitators attempted to re-administer the survey, less than half of the teachers responded. The level of non-response coupled with the time that passed (at least 4 months), we choose to omit these survey results from our analysis.

We compare various measures of teaching experience by condition (Table 3). Stretch condition teachers have more years of teaching experience but not significantly more experience teaching ninth grade mathematics or algebra. In other measures, teachers do not differ by condition. See Appendix A3 for more details.

Table 3: Differences in Teacher Level Variables by Condition Assessed by t-test

Variable	p-value	Conclusion of Significant t-test
Yrs Teaching	0.074	Slight evidence that stretch teachers have more experience
Yrs w/Ninth Grade Math	0.433	No difference
Yrs w/Algebra	0.446	No difference
Exp w/Block	0.356	No difference
Exp w/Coaching	0.775	No difference
Exp w/Doubledose	0.144	No difference
Proportion who Volunteered	0.103	No difference
Proportion w/ Math major	0.918	No difference
Proportion Certified	0.444	No difference

6 Student Surveys

While we believe outcome measures will be affected by pretest scores, we also believe student attitudes and teacher characteristics to be relevant. To measure these, we use responses to collections of survey items administered in the spring.

The survey consists of 35 questions scored on a four-point Likert scale. The full exploratory factor analysis is given in Appendix A.4 and we provide a short summary. Student survey items generally load on two different factors and many items are excluded as they do not load at all. Based on several measures of reliability, we find a optimal set of items for each factor (Table 4).

Although Cronbach's α is commonly used for reliability, Sijtsma (2009) suggests using the *glb* (greatest lower bound) as a better measure for reliability. Revelle and Zinbarg (2009) recommend ω_t which is computed using the uniqueness of each item (the unique variability of that item with respect to the other items). In addition, we compute Mokken scalability coefficients H for each set of items (van derk Ark 2007). H is a measure of covariance for discrete response items. A value of $H = 1$ implies a perfect Guttman scale and values of $0.3 < H < 0.4$, $0.4 < H < 0.5$, and $H > 0.5$

are considered weak, moderate, and strong scales. We name factors 1 and 2, *Student Attitudes* and *Teacher Practices* respectively (see Table 5 and Table 6 for corresponding survey questions).

Table 4: Reliability Measures for Optimal Student Survey Constructs Obtained From Exploratory Factor Analysis

	Construct 1 (Student Attitude)	Construct 2 (Teacher Practice)
α	0.90	0.76
glb	0.92	0.81
ω_t	0.88	0.71
H	0.43	0.30

Table 5: Student Survey Items: Student Attitude Factor

Survey Item	Question
14	I liked coming to math class
15	I paid attention in math class
16	I did my math homework
17	I felt that I could do almost all the work in math if I didn't give up
18	I felt bored in math class
19	I felt confused in math class
20	I worked hard in math class
21	I felt that math class was interesting
22	I studied for math tests and quizzes
23	I felt successful in math
24	I felt confident that I could do the math work
28	I am good at math
31	I like mathematics
32	When I try hard, I can usually understand math
34	Math problems usually make sense to me
35	This year I became a better math student than I used to be

Table 6: Student Survey Items: Teacher Practice Factor

Survey Item	Question
4	Students used objects or tools, such as rulers, protractors, or Algebra tiles.
5	The teacher asks students to explain how they got their answers
6	When I didn't understand something, my teacher tried to help me by asking me questions about my thinking.
7	Students are asked to show more than one way of solving a math problem
8	I worked on math problems during class time with other students in my class
9	II was asked to write a few sentences about how I solved a math problem
10	Students worked in small groups or with a partner
12	My teacher uses real-life examples to help us understand math
13	The teacher made sure that everyone understood before moving on to another topic

7 Teacher Observations

Teachers were observed in the fall and spring during the study year by two veteran teachers unaffiliated with the districts. Each observation lasted an entire 80 or 90 minute class period and various aspects of the class were recorded. Quantitative measures include the number of students present and number of students engaged during the beginning, middle, and end of the class, amount of time spent on house-keeping versus instruction, and the amount of time students spent working alone, in groups, and as a whole class. Teachers were also rated on a four point scale on their use of teaching practices, the extent that the class was student-centered and the level of student learning that was evident.

An exploratory factor analysis (see Appendix A.5 for details) indicate two factors for the fall observation data and two factors for the spring observation data. Factors and items are listed in Table 7. We also note that when the fall and spring observation data were grouped together or separately, items loaded in the same arrangements and fall items did not load with spring items.

Table 7: Optimal Teacher Observation Constructs Obtained From Exploratory Factor Analysis

Factor	Items	Description
Fall Factor 1	Best Practice, Student Learn, Student Center	Fall scores on teaching practices, student learning, student centeredness
Fall Factor 2	Engage1, Engage2, Engage3	Fall percentage of students engaged at the beginning, middle and end
Spring Factor 1	Best Practice, Student Learn, Student Center	Spring scores on teaching practices, student learning, student centeredness
Spring Factor 2	Engage1, Engage2, Engage3	Spring percentage of students engaged at the beginning, middle and end

Finally, we include the same reliability measures introduced in the previous section to justify our choice of these items for each factor Table 8. Note that the values for ω are inappropriate for the fall and spring factors which contain engagement levels. We suspect it is partly due to the small variation among the levels of engagement and the restricted range of [0,1].

Table 8: Reliability Measures for Optimal Teacher Observation Constructs Obtained From Exploratory Factor Analysis

	α	glb	ω	H
Fall1	0.94	0.94	0.93	0.94
Fall 2	0.76	0.83	-3.26	NA
Spring 1	0.93	0.94	0.92	0.91
Spring 2	0.89	0.90	-218	NA

8 Methodology

8.1 Missing Data and Multiple Imputation

Given the scale and length of our study, it unsurprising that we have missing data. There was movement by both students and teachers as well as administrative difficulties with collecting and reporting data. Missingness varies by covariate (Table 1); some missingness is due to school and district level issues and some is due to teacher and student issues. The purpose of this section is to highlight causes for missing data and justify the modifications made to compensate.

We first introduce some terminology common in the field (Schafer and Olsen (1998); Gelman and Hill (2006)). Missing data generally falls into one of three categories: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). MCAR data is data that if it were known and included in statistical models would not affect the results of the model. More formally, we can think of MCAR data as data being generated from the same distribution as the observable data. If we choose to exclude MCAR data, our overall inference and conclusions are unbiased; however, a smaller sample size increases standard errors.

MAR data is slightly different from MCAR data in that MAR data is assumed to be generated from the same distribution as the observed data only when conditioned on other covariates. For example, suppose student test scores tend to be missing for students who would have low scores because these students have poor attendance rates. These scores would not be considered MCAR. However, if we condition these scores on known attendance rates, we can assume that the missing test scores do not differ from the observed test scores. The final category, MNAR, encompasses all missing data that that is neither MAR or MCAR and is the most difficult type of missing data with which to work. MCAR and MAR data can be imputed whereas MNAR data is not imputable without making additional assumptions. There are methods to work with MNAR data but these are beyond the scope of this project. Instead, we constructed a data set that is a subsample of the full data set with missing data that is either MAR or MCAR.

We first discuss covariates that we choose to exclude from our analyses. First, we exclude Free and Reduced Lunch (FRL) status and English as a Second Language (ESL) status. These are measures that are missing at the district level; for example, only one district reported ESL status for students and only three districts reported FRL status for students. Districts cite privacy concerns as reasons for not disseminating such data. While these measures may be available on a district level, we do not believe district FRL or ESL status as an accurate estimate of FRL or ESL status of our study students. Recall that these students have been selected by the school as those students who are several years behind in mathematics and we believe these students to have higher rates of FRL and ESL than what is reported for the entire district. We also exclude the race variable. Although we might expect to find race effects, we have a large amount of missing race data. In fact, an entire district failed to report race and half the schools from another district are missing race. Furthermore, approximately half the students from two other districts are also missing race. We investigate the effects of race for the districts with minimal missing data (see Appendix A.6) and do not find evidence that race affects our outcome variables.

For our statistical inference, we therefore desire a subsample of our data that is maximally informative and includes as many students as possible but at the same time has missing data that is either MCAR or MAR. Because the intervention was designed for students who attend school somewhat regularly and who were enrolled in the treatment courses for the course of the year,

we are interested in students for whom we have evidence of their being in the study, namely the existence of their test scores. We first choose to exclude students who were missing both pretest scores (Fall CTBS or Fall OH) as these students were absent during the two days the test were administered and during the three rounds of make-up testing. For a student to be absent for these five or more days, we can assume that these students either did not attend school regularly or the student enrolled in the study several weeks after the start of the year (and missed the first half of the treatment).

While we could have included students with either a Winter CTBS or Spring Algebra TN posttest score, the Spring Algebra TN posttest score is the most important outcome measure from a policy perspective, and we chose to exclude students without a Spring Algebra test score. Moreover, including students with a midyear test score and without a spring test score would have increased our sample size by only 90 students.

In fact, given the sample of students who attend school regularly and remained in the study through the spring, spring test scores and survey responses are MAR. Given that the student has taken a pretest, the Spring Algebra TN score or student survey is a proxy indicator that that student was a part of the study for the course of the year. Our population of interest include only students exposure to the full treatment (regular attendance and attended the school during the entire year). A student who from this sample but that is missing a spring test score or a spring survey is assumed to be no different from the rest of the other students given their previous test scores. Thus, if we include students with a Spring Algebra TN score and a completed student survey, we have a sample that is from the same distribution as our focus population conditional on previous measures.

Based on this sampling procedure, we created a sample of 1233 students. We refer to this sample of students as our *standard* sample. Of these, 83 are missing one pretest score and 107 are missing the Winter CTBS test score. We choose to use 5 imputations (i.e. generate 5 sets of complete case data) for several reasons. Rubin (1987) suggests than 2-3 imputations offer adequate coverage probabilities of interval estimates, especially given the low amount of missing data. Schafer and Olsen (1998) suggest 3-5 imputations for most applications, and as storage is not a concern, we use 5 imputations.

There are also teachers for whom we are missing data: teachers who failed to return a fall survey and teachers who were not observed during the fall and/or the spring rounds of observations. There is reason to believe that these data may be MNAR, i.e. that these teachers may differ from teachers with complete or almost complete data cases. Teachers who did not complete a fall survey either did not attend the training sessions or were hired after the start of the school year and for whatever reason failed to complete a survey. While these teachers may not differ in ability, we lack the data to support either claim. Teachers missing observations were absent on the days of the observations, which suggests that these teachers may have higher rates of absenteeism than the teachers who were observed. By the same token, these teachers may not differ from the teachers who were observed and completed a teacher survey. We could also assume teacher data is MCAR. In our model selection procedures, we find teacher level variables are not relevant (Appendix A.7).

We impute our missing test scores in the following way: Since pretest scores are correlated, we impute missing pretest scores using a linear model, regressing one pretest score on the other. For the Winter CTBS test score, previous models (see Appendix A.7) suggested correlations with both pretest scores (Fall OH and Fall CTBS), Spring Algebra TN and condition. We use a linear model with these covariates to impute missing CTBS scores. See Appendix A.8 for regression models for

imputation.

Some final notes about our *standard sample*: First a summary of the pre-imputation data set is presented in Table 9. Note that, although treatment condition was not use to select students the standard sample is approximately balanced between conditions. We further examined differences between TAM and Stretch students on pretest scores: Table 10 gives the means, standard deviations, two-sample T statistics and p-values for the pre-imputation sample and the 5 imputation samples (ignoring clustering effects as described below). Fall OH pretest scores are significantly higher for the Stretch condition than the TAM condition. While this is unfortunate, we find a similar pattern with Stretch Fall OH scores being greater than the TAM in the full data set (see Appendix A.1). Any differences in condition may affect our analyses, so we may revisit this issue when we discuss our results.

Table 9: Summary of Pre-Imputation Data: Sample Sizes, Raw Score Means, and Standard Deviations by Condition

Variable	Stretch			TAM		
	N	$N_{missing}$	Mean (SD)	N	$N_{missing}$	Mean (SD)
Fall CTBS	627	41	10.42 (3.81)	606	15	10.19 (3.90)
Fall OH	627	11	22.56 (10.26)	606	16	21.33 (10.72)

Table 10: *Standard* Dataset n=1233: Comparing Raw Pretest Scores Across Condition by Dataset Using Two Sample t-test

	Fall CTBS				Fall OH			
Dataset	Stretch Mean (SD)	TAM Mean (SD)	T	$p-$ value	Stretch Mean (SD)	TAM Mean (SD)	T	$p-$ value
Pre	10.42 (3.81)	10.19 (3.90)	1.02	0.31	22.56 (10.26)	21.33 (10.72)	2.04	0.04
Imp 1	10.42 (3.83)	10.18 3.91)	1.08	0.28	22.47 (10.32)	21.68 (10.72)	2.00	0.05
Imp 2	10.42 (3.79)	10.17 (3.88)	1.14	0.25	22.28 (10.27)	21.18 (10.68)	2.18	0.03
Imp 3	10.38 (3.81)	10.18 (3.90)	0.91	0.36	22.56 (10.24)	21.32 (10.74)	2.07	0.04
Imp 4	10.41 (3.82)	10.15 (3.92)	1.14	0.25	22.51 (10.27)	21.24 (10.78)	2.11	0.03
Imp 5	10.35 (3.82)	10.20 (3.91)	0.68	0.50	22.37 (10.32)	21.31 (10.69)	1.78	0.08

Second, based on our selection procedure, different proportions of students from each school are retained in our sample. Figure 1 shows the distribution of these proportions. It is not surprising that schools vary in their levels of student participation in the study. Schools vary in their levels of student absenteeism, student mobility, and overall ability to implement such a study. Furthermore, we note that we are missing gender for 39 students in this new sample. Based on our preliminary models (see Appendix A.6), gender is not correlated with any combination of covariates, and for this reason we choose not to impute gender values or include them in our models.

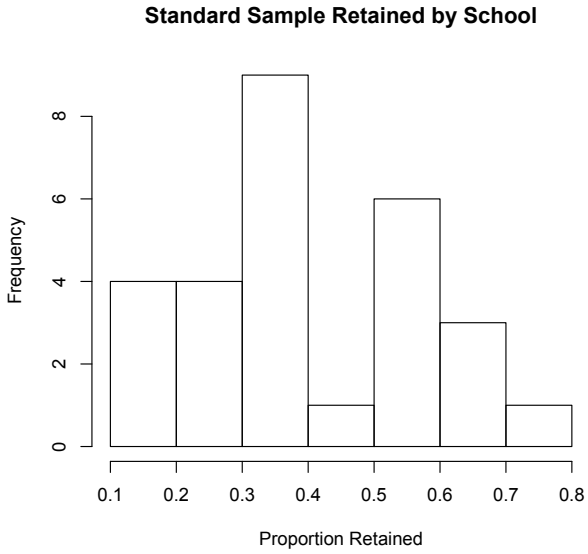


Figure 1: Proportion of Students by School Selected to be in *Standard* Dataset n=1233

8.2 Parameter Estimates from Multiple Imputation

For m imputations, we report the average parameter estimate,

$$\mu = \frac{1}{m}\sum_{i=1}^{m}\mu_i$$

where μ_i represents a parameter estimate for each imputation (Rubin 1987). The variance for each parameter estimate has two components, the variance of the parameter estimate for a given imputation (within-imputation variance) and the variance between imputations (between-imputation variance). Rubin (1987) suggests a weighted sum of these two components, namely,

$$Var(\mu) = \frac{1}{m}\sum_{i=1}^{m}Var(\mu_i) + (1+\frac{1}{m})Var(\mu^\star)$$

where $\mu^\star$ is the set of variance estimates across imputations. For all subsequent models, we report parameter estimates calculated in this way.

8.3 Motivation for a Nested Model

Because of the nested structure of the data, we have prior beliefs that outcome measures are more highly correlated for students with the same teacher and within the same school than between teachers or schools. We also believe that students and teachers within a district are also more correlated than between districts.

For each outcome variable, Figure 2 shows the variance in scores within each teacher (left) and within each school (right). Most within teacher variances are smaller than the overall variance for each outcome measure. Similarly, outcome measure variances tend to be smaller within schools than their respective overall variances. As we have some evidence of nesting, we choose to use a hierarchical linear modeling (HLM) structure for our data.

Ordinary least squares (OLS) assumptions are violated as observations are not independent of each other. HLM assumed that observations at the highest level of the model are independent of each other and that residuals at each level are uncorrelated. In addition, HLM follows the OLS assumptions of linearity, normality of residuals, and homoscedasticity.

8.4 Latent Variable Estimation

In Section 6, we introduce two constructs that we name Student Attitude and Teacher Practice based on student responses to two sets of items from the student survey. While each set of items reflect both Student Attitude and Teacher Practice, they are merely proxy measures of each and the actual variables, Student Attitude and Teacher Practice, are latent variables. Furthermore, we assume each of these sets of items is an adequate measure of the corresponding latent variable.

We consider two methods to estimate latent variables, using sum scores and fitting an Item Response Theory (IRT) model. Our items are polytomous with four categories. A sum score is calculated by simply adding the responses for each item. Sum scores are used mostly for convenience. We can also fit a polytomous IRT model to estimate the value of the latent variable given the set of student responses. In most cases, we would recommend an IRT model for latent variable estimation for several reasons. An IRT model produces information about each student's latent

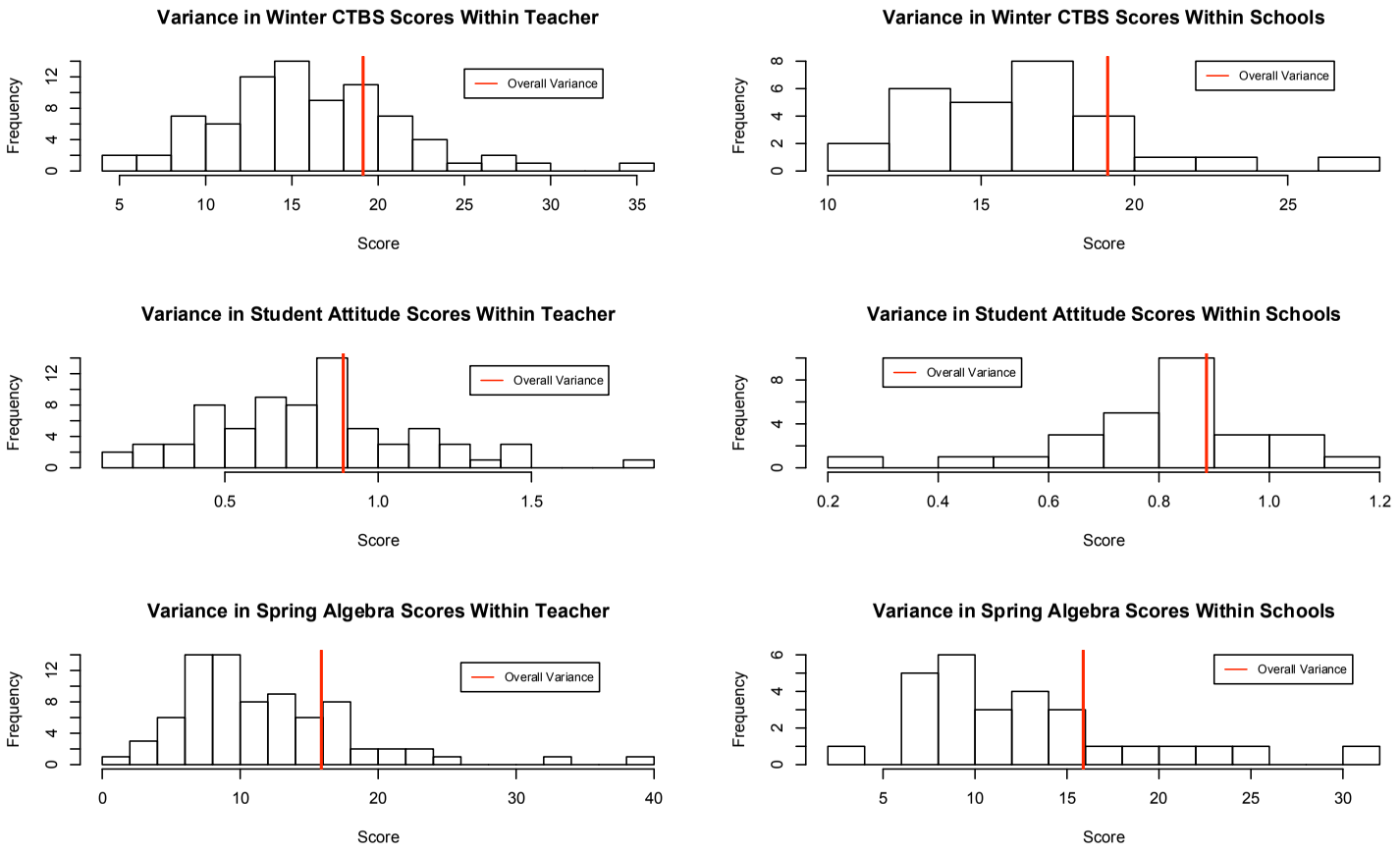


Figure 2: Variance in Outcome Measures Within Teacher and Within School

variable estimate but also provides information about each item. We can determine which items students tend to rate higher or lower and which items are better at separating or discriminating among students. In addition, a sum score treats a student responses to each item equally, whereas an IRT model incorporates items parameters in the latent variable estimation. For example, two students may have the same total response score but one student might have selected low responses to items that most the students selected high categories and the other might have selected low responses to items to items that most students selected low categories.

For items with polytomous responses, there are two popular models: The Graded Response Model(GRM; Samejima (1969)) and The General Partial Credit Model (GPCM; Muraki (1992), Masters (1982)). The probability of student i selecting category higher than k on item j as given by the GRM is

$$P(X_{ij} > k) = \frac{exp\big[a_j(\theta_i - b_{jk})\big]}{1 + exp\big[a_j(\theta_i - b_{jk})\big]}$$

where a_j is an item discrimination parameter, b_j is an item location parameter, and θ_i is the latent variable. Since the model specifies cumulative probabilities, there is an additional constraint that $b_{jk-1} < b_{jk}$ which guarantees that $P(X_{ij} > k) < P(X_{ij} > k-1)$. It is trivial that $P(X_{ij} = k) = P(X_{ij} > k-1) - P(X_{ij} > k)$.

For the GPCM, the probability of student i selecting category k on item j is given directly,

$$P(X_{ij} = k) = P_{jk}(\theta_i) = \frac{exp\big[\sum_{v=1}^{k} a_j(\theta_i - b_j + d_{jk})\big]}{\sum_{c=1}^{4} exp\big[\sum_{v=1}^{c} a_j(\theta_i - b_j + d_{jk})\big]}$$

where a_j is an item discrimination parameter, b_j is an item location parameter, d_{jk} is a threshold parameter for response k on item j, and θ_i is the latent variable.

We first compare model fits for the GRM and GPCM for each set of responses using AIC and BIC scores, and the GRM fits the data better (Table 11). We can fit these models using frequentist methods in R or using a Gibbs Sampler. We will revisit differences in methodology when we introduce our overall models in Section 9, but due to large differences in simulation time, we choose to use the GPCM to model these response data instead of the GRM. Although the GRM produces lower AIC and BIC scores, the θ_i estimate for each student i is highly correlated to those estimates given by the GPCM, $R^2 = 0.998$ and $R^2 = 0.997$ for Student Attitude and Teacher Practice respectively.

Table 11: AIC and BIC Scores for Latent Variable Estimation Using IRT Models: GRM and GPCM

	Model	AIC	BIC
Student Attitude	GRM	41140	41467
	GPCM	41315	41642
	Model	AIC	BIC
Teacher Practice	GRM	25864	26048
	GPCM	25880	26064

9 Results

For each response variable, we use variable selection procedures to obtain an optimal model, and we present the results separately.

9.1 Variable Selection

Using variable selection techniques, those variables that decrease the Akaike Criterion Information (AIC) score or Bayesian Criterion Information (BIC) score are included in our models. Variables from all three models are described in detail in Table 12. We also use these measures to select the number of random effects and the number/level of variance components for our models. For more detailed analysis of our variable selection process, see Appendix A.7. The only variables that were purposely excluded are the factors associated with teacher observations in the fall and spring.

Table 12: Description of Explanatory Variables included in Final HLM Models

Explanatory Variable	Description
Pretest CTBS	Intermediate math skills test given in the fall
Pretest OH	Algebra success predictor test given in the fall
Condition = Stretch	Indicator variable whether student is in the stretch condition
Gender = Male	Indicator variable whether student is male
Student Attitude Score	Estimate from IRT model based on student survey responses (see Student Surveys)
Teacher Practices Score	Estimate from IRT model based on student survey responses (see Student Surveys)

9.2 Winter CTBS Model

The selected model for Winter CTBS scores is a 2-level random intercept HLM with student and school levels.

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j}[FallCTBS]_{ij} + \beta_2[FallOH]_{ij} + \beta_3[Stretch]_j + \epsilon_{ij} \\ \beta_{0j} &= \gamma_0 + U_{0j} \end{aligned}$$

where i indexes students, j indexes schools, $\epsilon_{ij} \sim N(0, \sigma^2)$ and $U_{0j} \sim N(0, \tau^2)$.

We choose to fit this model using both frequentist and Bayesian methods. For maximum likelihood estimation, we fit our models in SAS, and for our Bayesian method of Gibbs sampling, we use a program called Just Another Gibbs Sampler (JAGS). Standardized regression coefficients and variance components for each method are shown in Figure 3. Additional tables and figures can be found in the Appendix A.9.1. Students with higher pretest scores (FallCTBS and FallOH) tend to have higher Winter CTBS scores and students in the Stretch condition do significantly worse than students in the TAM condition. Student level variance is much higher than school level variance, suggesting that students vary more within a school than average student performance varies across schools. In terms of methodology, SAS and JAGS produce similar parameter estimates.

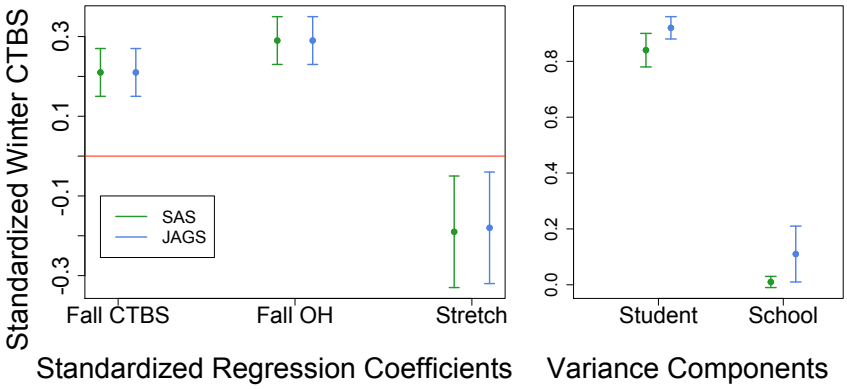


Figure 3: Winter CTBS Model: Standardized Regression Coefficients and Variance Components Estimated via SAS and JAGS with Error Bars (± 2SE)

9.3 Student Attitude Model

For Student Attitude as an outcome variable, we fit a 3-level random intercept HLM with student, teacher, and school levels. Also notice that in this model, there are two latent variables: α which is the student attitude measure and θ which is the teacher practice measure.

$$\begin{aligned}
\alpha_{ijk} &= \beta_{0jk} + \beta_1[FallOH]_{ijk} + \beta_{2j}[SpringAlgebraTN]_{ijk} + \\
&\quad \beta_3[Stretch]_k + \beta_4[\theta]_{ijk} + \epsilon_{ijk} \\
\beta_{0jk} &= \eta_{0k} + U_{0jk} \\
\eta_{0k} &= \delta_{00} + V_{0k}
\end{aligned}$$

where i indexes students, j indexes teachers, k indexes schools, $\epsilon_{ij} \sim N(0, \sigma^2)$, $U_{0jk} \sim N(0, \tau_1^2)$, and $V_{0k} \sim N(0, \tau_2^2)$.

In Section 8.5, we introduce two methods of estimating latent variables, using sum scores and fiting an IRT model. Thus, we could use either one of these techniques to first estimate these latent variables and then use these estimates as covariates in our HLM. These models are shown graphically in Figure 4 and Figure 5. Notice that in each of these models, we treat the latent variables as fixed and known when we fit the HLM. Another method for fitting this model is to estimate the latent variables and HLM parameters simultaneously using a Gibbs Sampler (Figure 6).

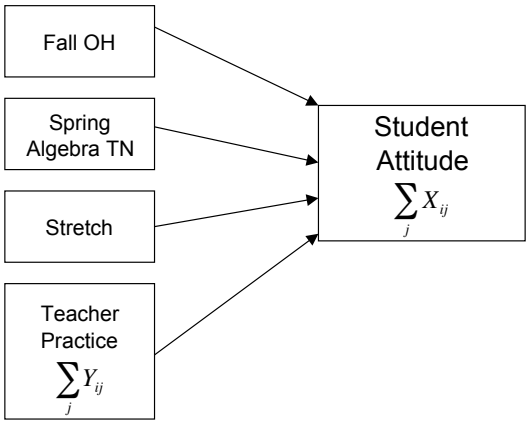


Figure 4: Student Attitude Model: Estimating Latent Variables Separately Using Sum Scores and Using Sum Scores in HLM

Regression coefficient estimates and variance components are shown in Figure 7. See Appendix A.9.2 for additional tables and figures. Students with higher Fall OH and Spring Algebra TN scores tend to have more positive student attitudes. Similarly, students who rate their teachers as having better teaching practices also have more positive student attitudes. There is not any evidence of an effect of condition on student attitude. Similar to the Winter CTBS model, we see more variance at the student level than at either the teacher or school levels. In fact, average student attitude varies very little from teacher to teacher and from school to school.

Unlike the Winter CTBS model, we do see differences in the three estimation methods. Using sum scores produces similar parameter estimates and standard errors as when IRT model estimates

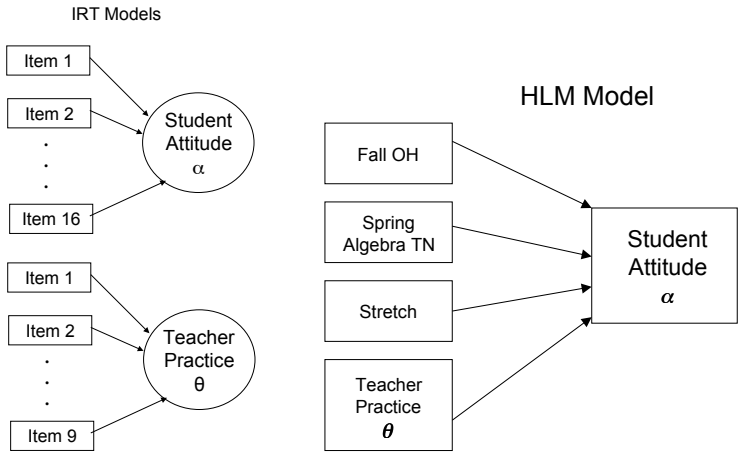


Figure 5: Student Attitude Model: Estimating Latent Variables Separately Using an IRT Model and Using IRT Model Estimates in HLM

are used, but parameter estimates are more dissimilar when we estimate latents and HLM parameters simultaneously in JAGS. In fact, there appears to be some amount of shrinkage towards zero in both the regression coefficients and variance components. We discuss possible reasons for these differences in Sections 11 and 12.3.

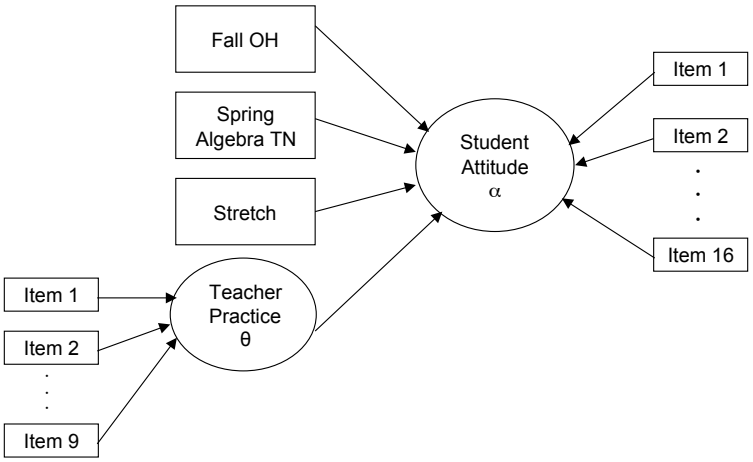


Figure 6: Student Attitude Model: Estimating Latent Variables and HLM Parameters Simultaneously

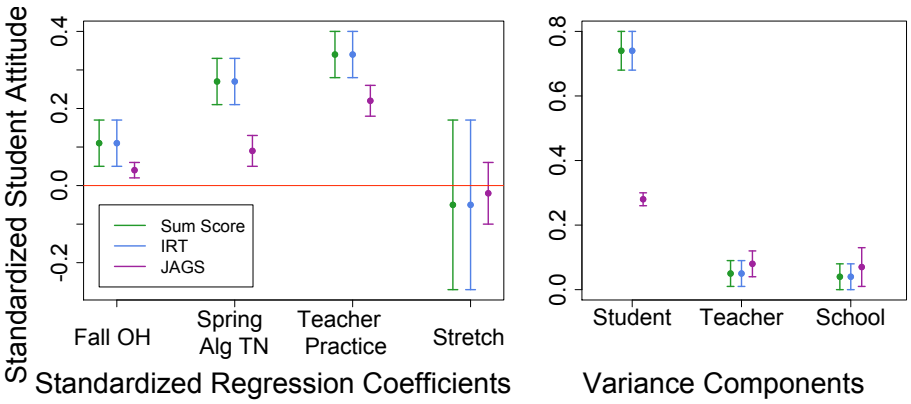


Figure 7: Student Attitude Model: Standardized Regression Coefficients and Variance Components With 3 Methods of Latent Variable Estimation, Sum Scores, IRT Models, and Simultaneous Estimation via JAGS with Error Bars (± 2SE)

9.4 Spring Algebra TN Model

Our final model is a 3-level random intercept HLM modeling the Spring Algebra TN score as the outcome variable. The levels for this model are student, teacher, and district. As in the Student Attitude Model, our model includes latent variables, Student Attitude and Teacher Practice. We present three methods of estimating our latent variables in the context of this model (Figure 8-Figure 10).

$$\begin{aligned} Y_{ijk} &= \beta_{0jk} + \beta_1[FallOH]_{ijk} + \beta_2[WinterCTBS]_{ijk} + \beta_3[Stretch]_{jk} \\ &\quad \beta_4[\theta]_{ijk} + \beta_5[\alpha]_{ijk} + \epsilon_{ijk} \\ \beta_{0jk} &= \eta_{0k} + U_{0jk} \\ \eta_{0k} &= \delta_{00} + V_{0k} \end{aligned}$$

where i indexes students, j indexes teachers, k indexes districts, $\epsilon_{ij} \sim N(0, \sigma^2)$, $U_{0jk} \sim N(0, \tau_1^2)$, and $V_{0k} \sim N(0, \tau_2^2)$.

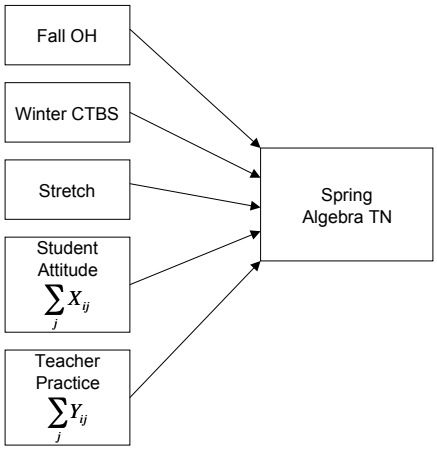


Figure 8: Spring Algebra TN Model: Estimating Latent Variables Separately Using Sum Scores and Using Sum Scores in HLM

Regression coefficient estimates and variance components are displayed in Figure 11. Additional information can be found in Appendix A.9.3. Students with higher Fall OH and Winter CTBS scores have higher Spring Algebra TN scores. There is also a positive relationship between Student Attitude and test score. While only marginally significant, students who rate their teachers as having poor Teacher Practices tend to score slightly better on the Spring Algebra TN test. Finally, there is not any evidence of differences between condition on algebra test scores.

In this model, there are fewer differences among the three estimation techniques. Again, using sum scores or IRT parameter estimates yield similar regression coefficients and variance components. The JAGS model for Spring Algebra TN no longer shows the same shrinkage trend that we saw in the Student Attitude model; rather, we see an increase in variance and a much higher regression coefficient of the variable Student Attitude.

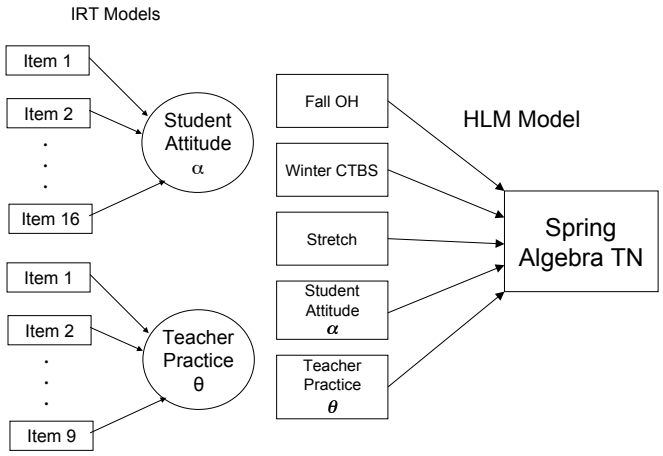


Figure 9: Spring Algebra TN Model: Estimating Latent Variables Separately Using an IRT Model and Using IRT Model Estimates in HLM

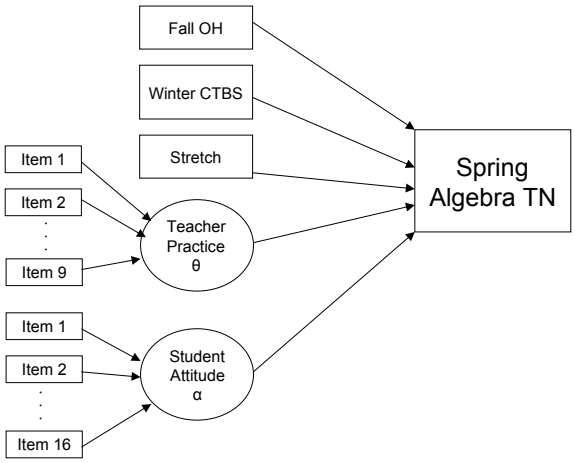


Figure 10: Spring Algebra TN Model: Estimating Latent Variables and HLM Parameters Simultaneously

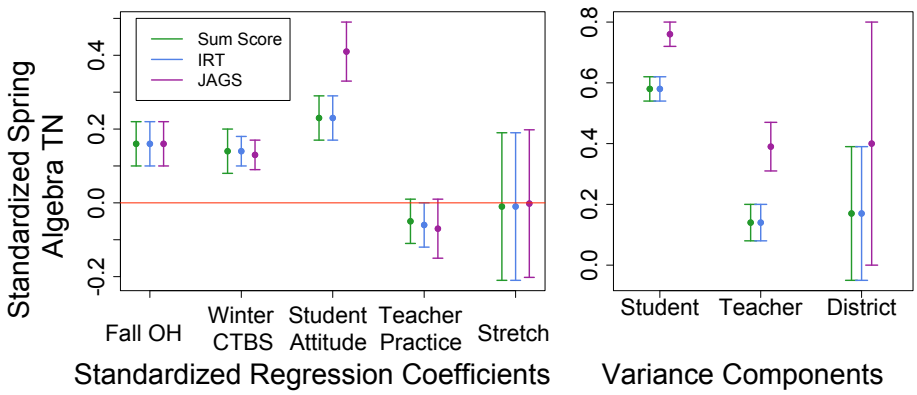


Figure 11: Student Attitude Model: Standardized Regression Coefficients and Variance Components With 3 Methods of Latent Variable Estimation, Sum Scores, IRT Models, and Simultaneous Estimation via JAGS with Error Bars (± 2SE)

10 Model Diagnostics

10.1 Goodness of Fit

We use goodness of fit statistics to assess our posterior samples from our JAGS models. We first introduce a goodness of fit statistic for our IRT models which was also used by (Johnson et al. (1999), Masters (1982)). For response y of student i to item j with parameters ϕ, we define the test statistic:

$$T_j(y|\phi) = \sum_{i=1}^{N} \frac{(y_{ij} - E(Y_{ij}|\phi))^2}{NV(Y_{ij}|\phi)}$$

Then, we define the *posterior predictive p-value* (Gelman et al. 1996) for item j as:

$$p \approx \frac{\#\{s : T_j(y|\phi_s) < T_j(y_s^\star|\phi_s) : s = 1, ...M\}}{M}$$

where $y_s^\star$ is data simulated from ϕ_s, the parameters from one step of our MCMC chain. A low p-value suggests a poor fit as the simulated data is a better fit for the model than the actual data.

As an example, we compare test statistics for our IRT models. Recall in Section 8.4, we introduced two IRT models for polytomous responses, the GRM and GPCM. We compare the p-values for each item fit using each model (Table 13).

We use a slightly adapted test statistic to assess our HLM model fits using JAGS. For outcome measure y of student i with parameters ϕ, define our test statistic:

$$H(y|\phi) = \sum_{i=1}^{N} \frac{(y_i - E(Y_i|\phi))^2}{NV(Y_i|\phi)}$$

then

$$p \approx \frac{\#\{s : H(y|\phi_s) < H(y_s^\star|\phi_s) : s = 1, ...M\}}{M}$$

For the Winter CTBS model (a 2-level HLM without an IRT component), we have a p-value of 0.975 which indicates that the model fit our data just as well as data simulated from our model.

For our Student Attitude and Spring Algebra TN models, we can assess the IRT and HLM portions of the model separately and together. We first test goodness of fit on each of the items for the IRT portion of the model. We then create a weighted average of test statistics to assess the overall model. Let S be the test statistic for the Student Attitude IRT and T be the test statistic for the Teacher Practice IRT. Define S and T as:

$$S_j(y|\phi) = \sum_{i=1}^{N} \frac{(y_{ij} - E(Y_{ij}|\phi))^2}{NV(Y_{ij}|\phi)}$$

where y is the set of responses to the Student Attitude items.

$$T_j(y|\phi) = \sum_{i=1}^{N} \frac{(y_{ij} - E(Y_{ij}|\phi))^2}{NV(Y_{ij}|\phi)}$$

where y is the set of responses to the Teacher Practice items.

Table 13: Goodness of Fit $p-$values for GRM and GPCM for Student Attitude and Teacher Practice Responses

	Item	GPCM	GRM
Student Attitude Responses	1	0.46	0.46
	2	0.49	0.52
	3	0.56	0.42
	4	0.30	0.14
	5	0.09	0.23
	6	0.23	0.29
	7	0.48	0.41
	8	0.52	0.46
	9	0.76	0.73
	10	0.58	0.64
	11	0.29	0.27
	12	0.48	0.41
	13	0.57	0.44
	14	0.52	0.62
	15	0.52	0.45
	16	0.43	0.40
Teacher Practice Responses	1	0.72	0.48
	2	0.74	0.59
	3	0.65	0.45
	4	0.59	0.54
	5	0.68	0.49
	6	0.84	0.64
	7	0.73	0.39
	8	0.69	0.43
	9	0.84	0.67

Then we construct a weighted average of test statistics in the following way: Let:

$$
\begin{aligned}
S_{avg} &= \frac{1}{16}\sum_{j=1}^{16} S_j(y|\phi_s) \\
S^{\star}_{avg} &= \frac{1}{16}\sum_{j=1}^{16} S_j(y^{\star}_s|\phi_s) \\
T_{avg} &= \frac{1}{9}\sum_{j=1}^{9} T_j(y|\phi_s) \\
T^{\star}_{avg} &= \frac{1}{9}\sum_{j=1}^{9} S_j(y^{\star}_s|\phi_s)
\end{aligned}
$$

Then,

$$p \approx \frac{\#\{s : S_{avg} + T_{avg} + H(y|\phi_s) < S^{\star}_{avg} + T^{\star}_{avg} + H(y^{\star}_s|\phi_s) : s = 1, ...M\}}{M}$$

We assess the IRT model portions of our Student Attitude model, and p-values are given in Table 14. There is no evidence of poor fit for any of the items in either the Student Attitude nor Teacher Practice IRT portion and in fact, the overall model has a p-value of 1. The same is true for the Spring Algebra TN model (Table 15). Based on these results, we believe that these models do not fail to fit our data; however, we question whether the test statistic H is an adequate measure of fit for linear models as it returns p-values near one for each HLM tested (Winter CTBS, Student Attitude, and Spring Alg TN).

Table 14: Goodness of Fit Tests for Student Attitude Model

	Item p-value							
Student Attitude IRT	1	2	3	4	5	6	7	8
	0.50	0.78	0.50	0.27	0.16	0.31	0.69	0.87
	9	10	11	12	13	14	15	16
	0.67	1.00	0.98	0.92	0.75	0.54	0.95	0.44

Teacher Practice IRT	1	2	3	4	5	6	7	8	9
	0.67	0.74	0.62	0.57	0.72	0.84	0.73	0.67	0.77

Overall Model	1.00

Table 15: Goodness of Fit Tests for Spring Algebra TN Model

	Item p-value							
Student Attitude IRT	1	2	3	4	5	6	7	8
	0.47	0.54	0.56	0.30	0.11	0.21	0.48	0.56
	9	10	11	12	13	14	15	16
	0.77	0.59	0.28	0.46	0.58	0.50	0.49	0.44

Teacher Practice IRT	1	2	3	4	5	6	7	8	9
	0.69	0.71	0.67	0.57	0.70	0.83	0.71	0.69	0.84

Overall Model	1.00

11 Sensitivity Analyses

11.1 Priors

In addition to model fit diagnostics, sensitivity analyses are also used in Bayesian analyses. We fit both the Student Attitude model and Spring Algebra TN model in JAGS using different prior specifications to determine the effect of prior choice on parameter estimates. We fit each model using less informative and more informative priors. Priors are shown in Table 16. See Appendix A.11 for JAGS code.

Table 16: Gibbs Sampler Priors Used for Sensitivity Analysis

Parameter	Flat Priors	Original Model Priors	Strong Priors
Discrimination	U[0,10]	U[0,4]	U[0, 4]
Location	N(0, 10)	N (0,1)	N(0, 0.5)
Threshold	N(0, 1)	N(0, 0.01)	N(0, 0.001)
Student Variance	U[0, 100]	U[0, 10]	U[0, 1]
Teacher Variance	U[0, 100]	U[0, 10]	U[0, 1]
School Variance	U[0, 100]	U[0, 10]	U[0, 1]
Regression Coefficient	N(0, 100)	N(0, 1)	N (0.2, 0.5)*

*Prior for Stretch Coefficient is N(0, 0.5)

We compare standardized regression coefficients and variance component estimates from each set of priors and each model (Figure 12, Figure 13). Exact parameter estimates are included in the Appendix A.10. For the Student Attitude model (Figure 12), we see that a stronger, more informative prior causes regression coefficients and variance components to decrease towards zero. Thus, some of the shrinkage we saw in Section 9.3 is due to our choice of prior despite our original efforts to choose an uninformative prior. Note also that flatter priors have larger standard errors.

For the Spring Algebra TN model (Figure 13), priors have a less consistent effect on parameter estimates. We see little effect of priors on regression coefficients on fixed covariates, and different effects on our latent variables. Priors do not appear to affect the coefficient for Teacher Practice but stronger priors do seem to increase the regression coefficient for Student Attitude and increase the associated standard error. Other than an atypically large Teacher variance estimate, priors do not appear to affect variance components.

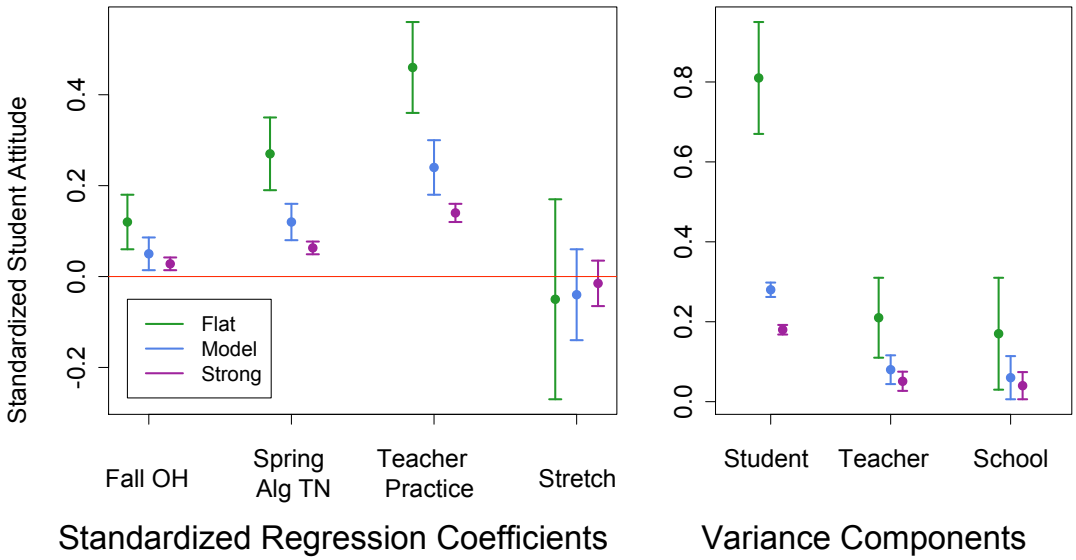


Figure 12: Student Attitude Model: Effect of Gibbs Sampler Priors on Parameter Estimates

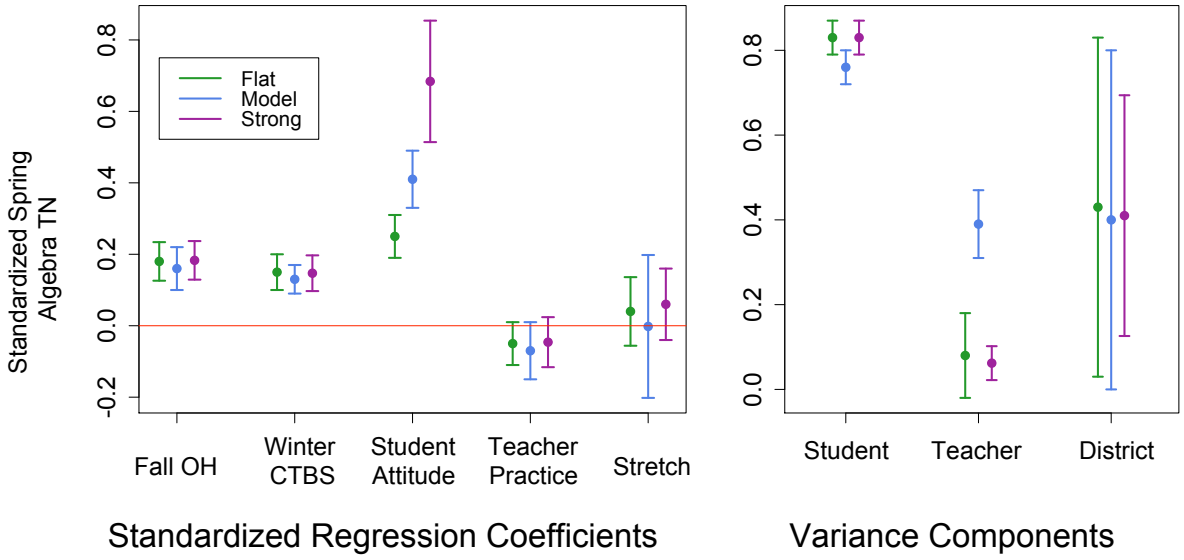


Figure 13: Spring Algebra TN Model: Effect of Gibbs Sampler Priors on Parameter Estimates

11.2 Imputation Techniques

We are also interested in the effect that the missing data imputation technique has on our parameter estimates. Recall that for our results, we use a regression model to impute missing data. We refer to this method of imputation as *Random Regression* as we sampled from our regression model using residual error. We consider two additional methods of imputation. We first use a method that is computationally simple, using overall means for missing values for each variable. For example, in our data set we are missing values for Fall OH, Fall CTBS and Winter CTBS. Rather than using a model to estimate missing values, we use the group mean. The overall mean Fall OH scores is used for all of the missing Fall OH score, the overall mean Fall CTBS score is used for all of the missing Fall CTBS scores, and the overall mean Winter CTBS is used for all of the missing Winter CTBS scores. We refer to this method of imputation as *Group Means*, and we note that this method only yields one data set.

The second method involves imputing missing data values using Bayesian techniques. We can use a Gibbs sampler to simultaneously model our missing data along with our model of interest. As JAGS is a Gibbs Sampling algorithm, we could theoretically use a JAGS model to impute missing data and fit our model. However, BUGS-type algorithms are recommended for missingness in the response variable only (Kynn 2006). We therefore create a JAGS model that simulates draws for missing data for Fall CTBS, Fall OH and Winter CTBS as outcome variables using regression models as well as parameter values for the primary model. See Appendix A.11 for model codes. We refer to this imputation technique as *JAGS*.

According to Gelman et al. (2004), the proper way to use a Gibbs sampler to model data with missing values is to use a two stage approach. Note that the model remains the same in each stage. In the first stage, the simulation is run and for each missing value, a random sample of n posterior draws are taken, creating n complete data sets. The second stage uses these n imputations and the simulation is rerun with these data. Posterior samples are then pooled to create posterior densities for each parameter. Due to time constraints, it was not possible to do this two-stage approach with our data however we would expect parameter estimates to be similar.

We compare parameter estimates using these methods of imputation for each model (Figure 14 and Figure 15). First, there appears to be no difference between using the Random Regression and JAGS imputation techniques for either model. There is some evidence of shrinkage of regression coefficients using group means in the Student Attitude Model (Figure 14). This is unsurprising as using group means for missing data shrinks slope parameters since including additional data point with a slope of zero influences the slope toward zero. The method of imputation appears to have no effect on variance components. For the Spring Algebra TN model (Figure 15), using group means for missing data significantly reduces the Student Attitude slope as well as the Teacher level variance component. We also note that the Student level variance is higher for using group means and it is possible that the increase in Student level variance drives the decrease in Teacher level variance.

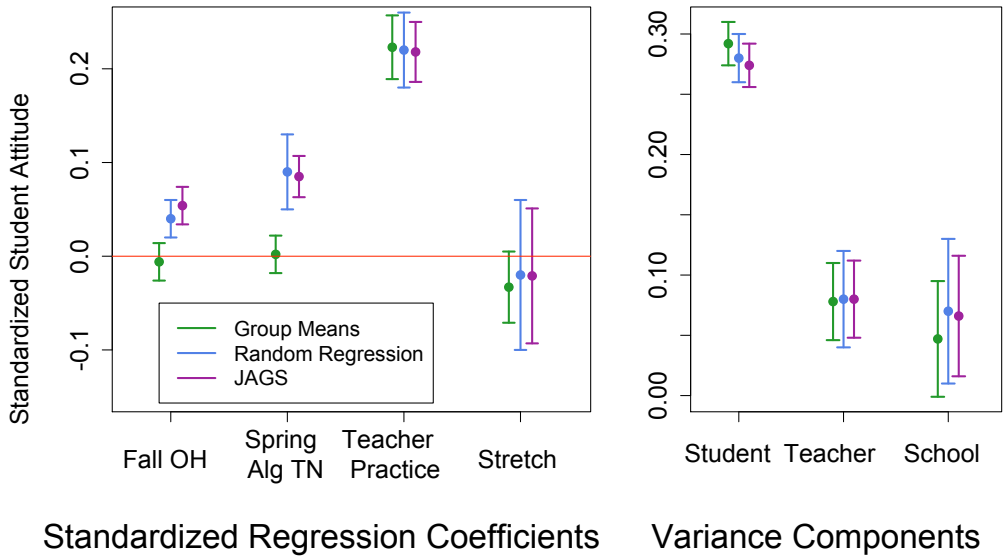


Figure 14: Student Attitude Model: Effects of Missing Data Imputation Techniques on Parameter Estimates

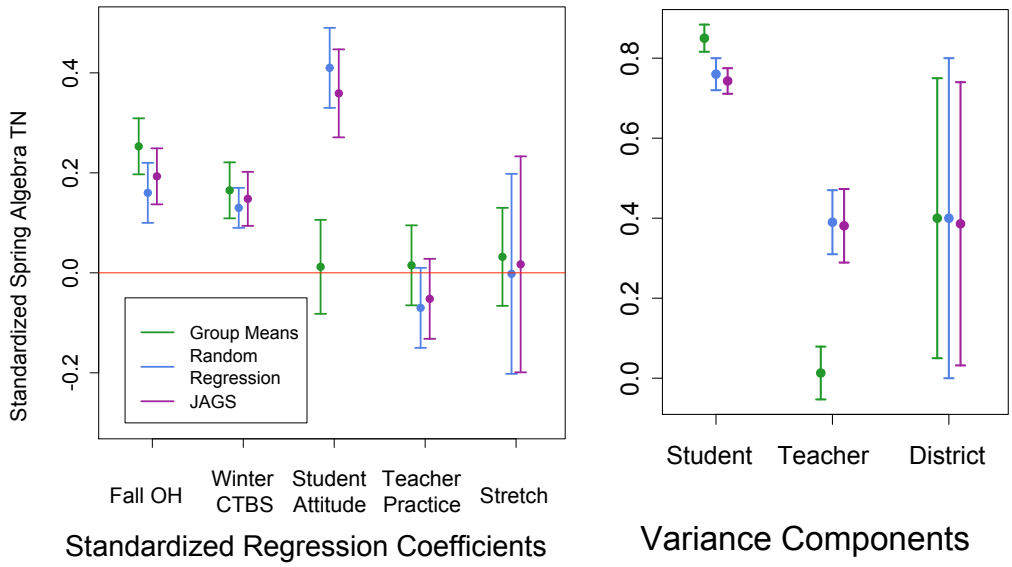


Figure 15: Spring Algebra TN Model: Effects of Missing Data Imputation Techniques on Parameter Estimates

12 Discussion

12.1 Research Study

Given our focus population of students who remained in the study for the entire time and who attend school somewhat regularly, we infer that students in the TAM condition outperform the Stretch condition on a test measuring general math ability through questions targeting intermediate math skills. While the Stretch condition allows teachers the additional time to focus on intermediate math skills, the TAM condition provides structured lessons on these topics. Thus, it is not surprising that TAM students would outperform the Stretch students.

The attitudes of these students, however, do not differ between conditions. There are several possible reasons for these null results. Recall that these surveys were administered in the spring when all students were enrolled in an Algebra I course. It is possible that student attitudes may have changed during the TAM course (as this is a goal of the course), but that these attitudes reverted back during the second semester. Another possible interpretation of course is that TAM does not affect student attitudes. We did find positive relationships between attitude and teacher practices, Fall OH scores, and Spring Alg TN scores. Students whose teachers have better practices have better attitudes towards their class. Students who began the study with higher propensity to learn algebra have better attitudes and students who end the study with higher algebra abilities have better attitudes.

Finally, we find that algebra performance does not differ between conditions. While these results may appear to be null, most educators would agree that we should expect students in the TAM condition to do worse than students in the Stretch condition as they had less instructional time in Algebra I. Recall that Stretch students took Algebra coursework throughout both semesters whereas the TAM students did not begin formal Algebra coursework until the second semester. These results are quite promising and suggest that a course like TAM facilitates students to "catch-up" to their classmates.

12.2 Latent Variable Estimation

We use three methods to estimate latent variable measures from survey responses: sum scores and IRT models, a separate model and as part of an HLM. Overall trends in regression coefficients are constant across estimation methods, however regression coefficients and variance components differed from models using sum scores in SAS more in the JAGS models using IRT-based latent variables than in the SAS models using IRT-based latent variables. This is less than surprising for two reasons. The first is that we are comparing not only two different methods of estimation but also two different models. The SAS models treat the latent variables as fixed and known whereas the JAGS models treat them as parameters to be estimated. The second reason is that the latent variable estimates using sum scores and a separate IRT model are highly correlated. In essence, the two SAS models are virtually identical whereas the JAGS model includes uncertainty for estimating the latent variables.

Under most circumstances, we do not expect sum scores to be identical to latent variable estimates (θ) produced by an IRT model. While these methods of estimation may produce correlated estimates, the purpose of using an IRT model is to take into account differences in the survey items. Items with varying location parameters are more likely to yield θ estimates that are less highly correlated with sum scores. We see this with our simulated data in Section 12.3. For our

particular data set, however, we find that all items have similar location parameter estimates and therefore produce highly correlated sum scores and θ estimates.

12.3 Simulations

In this section, we explore causes behind some of the trends in our results. There is a large effect of prior specification on the JAGS models parameter estimates when the outcome variable is modeled as a latent but there is little effect when the outcome variable is fixed. In addition, we saw some amount of shrinkage of regression coefficients towards zero when the outcome variable is modeled as a latent variable although it is possible that this is solely attributed to choice in prior. The difference between these two models (Student Attitude and Spring Algebra TN) is that the Student Attitude model incorporates measurement error in student responses as part of the outcome variable. To investigate the effects of priors and measurement error, we conduct several simulations.

Each simulated dataset contains 500 students and the number of survey items are 16 and 9 for Student Attitude and Teacher Practice. We create two datasets for each model, one with high measurement error (low discrimination item parameters) and one with low measurement error (high discrimination item parameters).

We fit each model using four methods. For the first two, we first estimate our latent variables using sum scores or an IRT model and then fit the HLM in SAS. For the second two, we use a JAGS model to estimate our latent variables and HLM in the same model. We use the set of priors used in the results section and the set of flat priors used in Section 11.1. As these data are simulated, the true regression coefficient and variance component is also shown.

The simulated Student Attitude model parameter estimates are given in Figure 16 and Figure 17. We see that the prior specification significantly affects parameter estimates with the flat prior increasing both standard errors and parameter estimates. There is a slight difference between estimates with low and high measurement error. Also, there is shrinkage of parameter estimates towards zero with the high measurement error data only, suggesting that both measurement error and prior specification affect parameter estimates.

There is little effect of measurement error on parameter estimates when the outcome variable is not a latent variable, and again, we see that there is no effect of prior specification on parameter estimates (Figure 18 and Figure 19). Thus, we conclude that including measurement error in the model is particularly important when the outcome variable is a latent variable. Similarly, these models are also highly sensitive to choice in prior.

These conclusions are particularly important as most outcome variables in educational interventions are in fact latent variables, regardless of whether they are modeled as latent or not. For this study, two of our outcome measures are test scores which we assume to be proxy measures for overall math ability and proficiency in algebra. In fact, both of these test scores are actually sum scores based on the number of correct responses on each test and these outcome variables could also be modeled as latent variables.

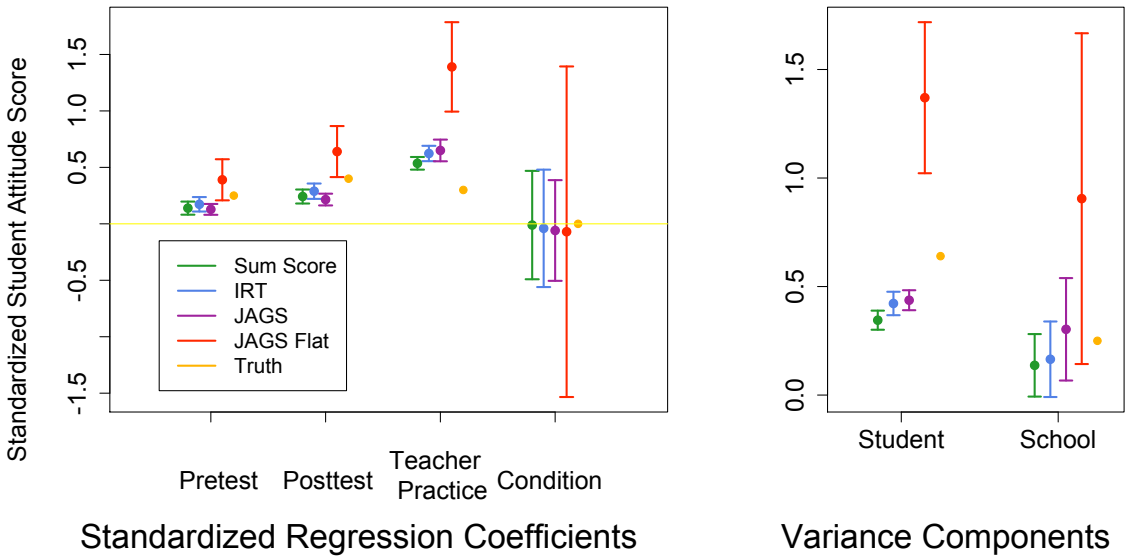


Figure 16: Student Attitude Model: Simulated Data with Low Measurement Error

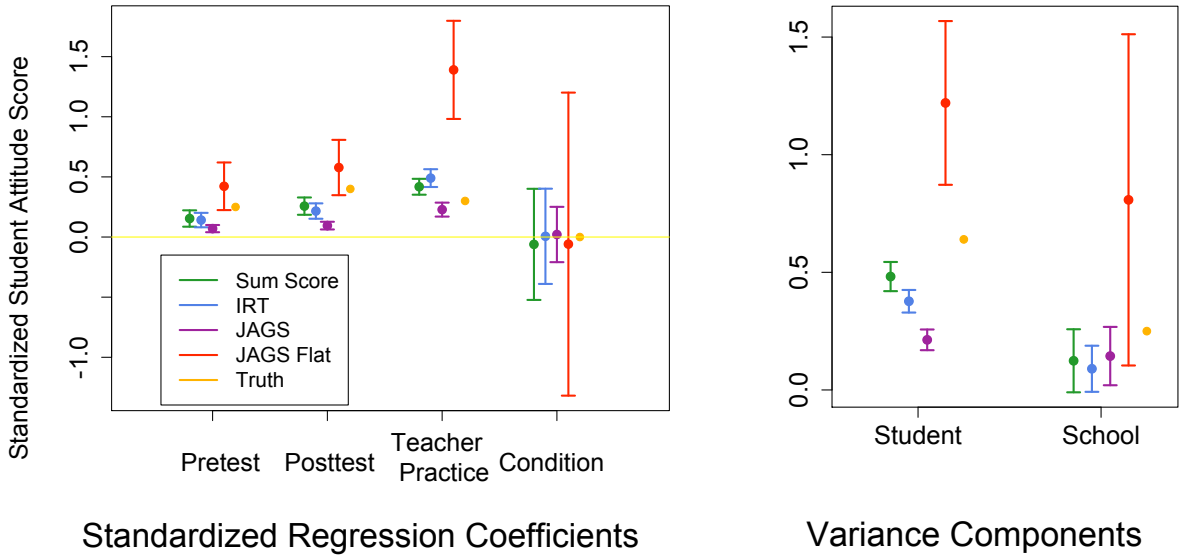


Figure 17: Student Attitude Model: Simulated Data with High Measurement Error

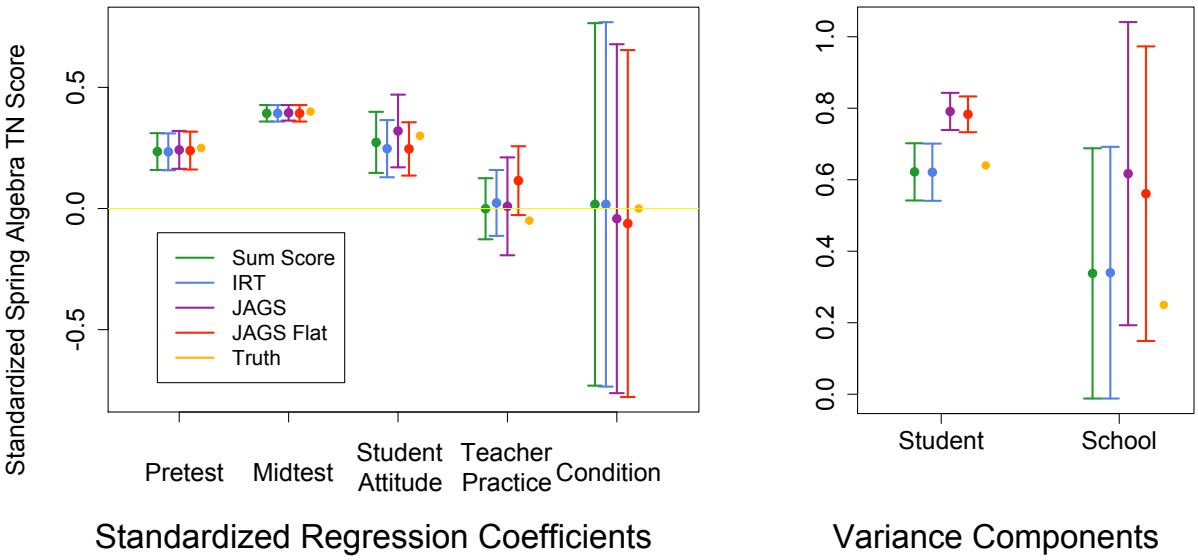


Figure 18: Spring Algebra TN Model: Simulated Data with Low Measurement Error

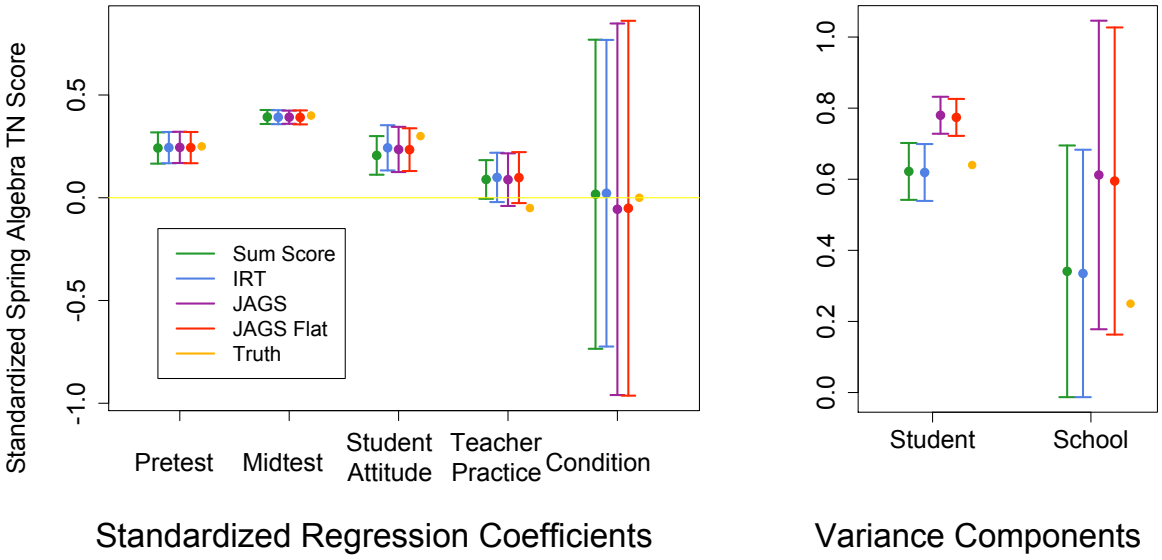


Figure 19: Spring Algebra TN Model: Simulated Data with Low Measurement Error

A Appendix

A.1 Differences in Student Baseline Variables by Condition

To determine differences in categorical student baseline variables by condition, we fit log linear models to the full data set (n=3327). We fit two models, one with the missing values as a separate variable and one with the missing values excluded. For each variable and each model, we show the residuals, Pearson χ^2 test statistic for model of independence and p-value (Table 17 -Table 19). In addition, we compare fall test scores (Table 20).

Table 17: Contingency Table for Students Race by Condition

Race	Stretch	TAM
White	233	255
Black	411	574
Hispanic	301	251
Asian	8	16
Native American	24	42
Multiracial	5	14
Other	1	1
NA	665	525

Model	χ^2 t.s.	p-value
w/missing	61.2	$p < 1e-10$
w/o missing	31.6	$p < 2e-5$

Table 18: Student Gender by Condition

	Stretch	TAM
Male	759	730
Female	689	610
NA	201	338

Model	χ^2 t.s.	p-value
w/missing	40.3	$p < 1e-8$
w/o missing	1.2	$p = 0.267$

Table 19: Free and Reduced Lunch Status by Condition

	Stretch	TAM
No Free and Reduced Lunch	132	178
Free and Reduced Lunch	313	362
NA	1204	1138

Model	χ^2 t.s.	p-value
w/missing	12.0	$p = 0.002$
w/o missing	1.2	$p = 0.267$

Table 20: Fall Pretest Scores by Condition

Variable	Stretch Mean (n)	TAM Mean (n)	2-Sample t-test p-value
Fall CTBS	6.64 (1205)	6.59 (1194)	0.617
Orleans Hanna	21.27 (1247)	20.55 (1207)	0.097

A.2 Differences in Student Baseline Variables by District

We include similar analyses by district. We include counts and Pearson χ^2 test statistics for the model of independence and p-values for log linear models that include and do not include missing data. In addition, we fit an ANOVA model for each fall pretest score by district.

Table 21: Student Race by District

District	White	Black	Hispanic	Asian	Nat. Am.	Multiracial	Other	NA
0	51	19	366	3	44	0	0	217
1	31	103	68	0	2	3	0	33
2	0	0	0	0	0	0	0	607
3	246	66	48	11	16	0	0	2
4	1	209	1	0	0	0	0	180
5	71	202	26	8	2	0	2	12
6	1	194	10	0	0	0	0	136
7	87	192	34	2	2	16	0	3

Model	χ^2 t.s.	p-value
w/missing	3905.2	$p \approx 0$
w/o missing	1785.5	$p \approx 0$

Table 22: Student Gender by District, n=2788

	Male	Female	NA
0	267	224	209
1	122	90	28
2	269	228	110
3	231	156	2
4	100	131	160
5	187	134	2
6	151	162	28
7	162	174	0

Model	χ^2 t.s.	p-value
w/missing	602	$p \approx 0$
w/o missing	27.4	$p = 0.0003$

Table 23: Free and Reduced Lunch Status by District, n=985

	No FRL	FRL	NA
0	142	324	234
1	0	0	240
2	0	0	607
3	133	194	62
4	0	0	391
5	0	0	323
6	35	157	149
7	0	0	336

Model	χ^2 t.s.	p-value
w/missing	2371.4	$p \approx 0$
w/o missing	29.7	$p = 0.0001$

Table 24: ANOVA Analysis Pretest Scores by District

District	Mean CTBS	Mean OH
0	6.9	20.8
1	6.9	24.2
2	5.8	16.1
3	6.5	21.1
4	5.5	19.0
5	7.0	21.2
6	7.1	23.8
7	7.3	23.7

Test	F statistic	p-value
Fall CTBS	13.41	0.00025
Orleans Hanna	24.416	$< 1e-6$

A.3 Differences in Teacher Baseline Variables by Condition

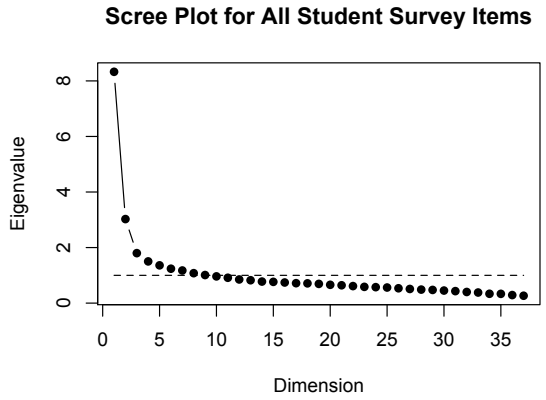
We compare teacher level variables by condition. Self-reported responses by condition are compared using a 2-sample t-test in Table 25.

Table 25: Differences in Teacher Level Covariates by Condition

Covariate	Stretch Mean	TAM Mean	p-value
Yrs Teaching	9.40	5.44	0.046
Yrs w/Ninth Grade Math	5.49	3.84	0.235
Yrs w/Algebra	5.61	3.79	0.207
Experience w/Block Schedule	0.70	0.69	0.909
Experience w/Coaching	0.35	0.28	0.533
Experience w/Doubledose	0.38	0.50	0.287
Proportion who Volunteered	0.58	0.38	0.090
Proportion w/ Math major	0.64	0.66	0.894
Proportion Certified	0.67	0.61	0.587

A.4 Student Survey Exploratory Factor Analysis

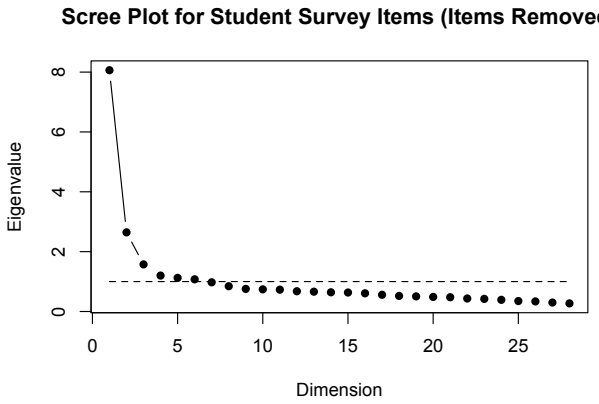
We first fit a scree plot to determine the number of factors. Based on the scree plot, we fit a 3 and



5 factor model.

We include the results of the 3-factor model fits with no rotation and oblique rotation (Table 26, Table 27) however there is some evidence that a 2-factor model may fit better.

We remove items that do not load well and create another scree plot in which a 2-factor model is suggested. We fit a 2-factor model and include factor loadings for the no rotation and oblique rotation models (Table 28, Table 29).



We group the items into two factors and determine a optimal set of items for each factor based on reliability measures. We name the factors in the following way: Student Attitudes (Items 14-24, 26, 28, 31-35, 37) and Teacher Practices (Items 4-10, 12, 13). Based solely on these results, it appears that the optimal item set for Student Attitudes include Items 14-21, 28, 31, 32, and 34, however researchers at Johns Hopkins found evidence for including item 35. As this does not drastically affect reliability measures, we include this item. In addition, we also consider Item 37 however research team members did not find evidence for this Item in their analysis so we do not include this Item. For Teacher Practices, we have so few items that we prefer to keep all 9 items rather than use a more reliable set of fewer items.

Table 26: Factor Loadings for 3-Factor Model on Student Survey Data: No Rotation and Items with Loadings < 0.25 on all Factors Excluded

Item	Factor1	Factor2	Factor3
14	0.67	0.17	-0.01
15	0.63	0.09	0.45
16	0.54	0.05	0.36
17	0.57	-0.04	0.00
20	0.65	0.01	0.39
21	0.68	0.16	-0.02
23	0.79	-0.13	0.02
24	0.71	-0.18	-0.03
28	-0.70	0.36	0.19
31	-0.69	0.16	0.18
34	-0.67	0.25	0.24
35	-0.62	0.03	0.07
2	0.11	0.31	-0.02
3	0.17	0.30	-0.07
4	0.18	0.44	-0.17
5	0.28	0.38	-0.13
6	0.38	0.45	-0.11
7	0.21	0.43	-0.14
8	0.31	0.36	-0.18
9	0.14	0.41	-0.18
10	0.27	0.35	-0.17
11	0.08	0.30	0.01
12	0.32	0.43	-0.17
13	0.44	0.34	-0.03
18	-0.49	-0.12	0.06
19	-0.48	0.24	0.19
22	0.47	0.19	0.22
26	-0.35	-0.06	-0.30
32	-0.49	0.15	0.17
33	-0.47	-0.01	0.16
37	0.37	-0.19	0.04

Table 27: Factor Loadings for 3-Factor Model on Student Survey Data: Oblique Rotation and Items with Loadings < 0.25 on all Factors Excluded

Item	Factor1	Factor2	Factor3
19	0.65	0.05	0.11
23	-0.59	0.02	0.27
24	-0.62	-0.02	0.18
28	0.88	0.13	0.06
31	0.72	-0.07	0.03
32	0.57	-0.04	0.08
34	0.82	-0.01	0.12
4	0.09	0.56	-0.10
5	0.01	0.51	-0.03
6	0.01	0.58	0.05
7	0.09	0.55	-0.05
8	-0.07	0.51	-0.08
9	0.09	0.53	-0.13
12	-0.01	0.58	-0.05
15	0.02	0.00	0.80
16	-0.02	-0.01	0.64
20	-0.10	-0.05	0.72
26	-0.05	0.02	-0.50
2	0.16	0.35	0.05
3	0.06	0.37	0.00
10	-0.04	0.49	-0.09
11	0.18	0.31	0.08
13	-0.05	0.45	0.14
14	-0.31	0.31	0.23
17	-0.40	0.08	0.18
18	0.28	-0.25	-0.10
21	-0.34	0.31	0.22
22	0.02	0.18	0.46
25	-0.02	0.13	-0.35
33	0.43	-0.19	0.05
35	0.48	-0.14	-0.11
36	0.41	0.03	0.13
37	-0.36	-0.13	0.15

Table 28: Factor Loadings for 2-Factor Model on Student Survey Data: No Rotation and Items with Loadings < 0.25 on all Factors Excluded

Item	Factor1	Factor2
14	0.68	0.20
15	0.60	0.05
16	0.52	0.02
17	0.58	-0.05
20	0.63	-0.03
21	0.70	0.18
23	0.79	-0.15
24	0.72	-0.19
28	-0.69	0.35
31	-0.69	0.13
34	-0.66	0.24
35	-0.62	0.03
4	0.18	0.42
5	0.28	0.37
6	0.38	0.46
7	0.22	0.44
8	0.31	0.35
9	0.14	0.42
10	0.27	0.33
12	0.32	0.45
13	0.44	0.36
18	-0.48	-0.15
19	-0.49	0.20
22	0.46	0.17
26	-0.34	-0.02
32	-0.50	0.16
33	-0.48	-0.02
37	0.38	-0.18

Table 29: Factor Loadings for 2-Factor Model on Student Survey Data: Oblique Rotation and Items with Loadings < 0.25 on all Factors Excluded

Item	Factor1	Factor2
17	-0.53	0.09
19	0.58	0.11
20	-0.55	0.12
23	-0.79	0.03
24	-0.77	-0.04
28	0.87	0.22
31	0.69	-0.02
32	0.55	0.05
34	0.75	0.11
35	0.55	-0.13
4	0.18	0.53
6	0.04	0.61
7	0.17	0.55
9	0.22	0.52
12	0.08	0.59
13	-0.09	0.52
5	0.06	0.49
8	0.02	0.48
10	0.04	0.44
14	-0.42	0.40
15	-0.47	0.20
16	-0.43	0.15
18	0.29	-0.29
21	-0.45	0.38
22	-0.25	0.31
26	0.27	-0.11
33	0.39	-0.14
37	-0.46	-0.11

Table 30: Student Attitude Reliability Measures $\alpha = 0.84$, $glb = 0.85$, $\omega_t = 0.84$, H=0.26 and Leave One Item Out Reliability Measures

item	$\alpha_{[-i]}$	$glb_{[-i]}$	$\omega_{t,[-i]}$	$H_{[-i]}$
14	0.83	0.84	0.82	0.24
15	0.83	0.86	0.83	0.25
16	0.83	0.86	0.83	0.25
17	0.83	0.84	0.82	0.25
18	0.84	0.85	0.83	026
19	0.84	0.87	0.83	0.25
20	0.83	0.86	0.82	0.25
21	0.83	0.84	0.82	0.24
22	0.84	0.87	0.83	0.26
23	0.82	0.81	0.81	0.24
24	0.83	0.85	0.82	0.24
26	0.87	0.87	0.86	0.31
28	0.83	0.81	0.82	0.24
31	0.83	0.87	0.82	0.24
32	0.84	0.87	0.83	0.25
33	0.84	0.88	0.83	0.26
34	0.83	0.88	0.82	0.25
35	0.88	0.87	0.86	0.33
37	0.84	0.85	0.83	0.26

Table 31: Subsets of Student Attitude Items and Reliability Measures

Items	α	glb	ω_t	H
***	0.89	0.91	0.87	0.43
***, 34	0.90	0.93	0.87	0.43
***, 34, 35	0.86	0.88	0.85	0.32
***, 34, 35, 37	0.86	0.89	0.85	0.31
***, 33, 34, 35, 37	0.87	0.87	0.85	0.31

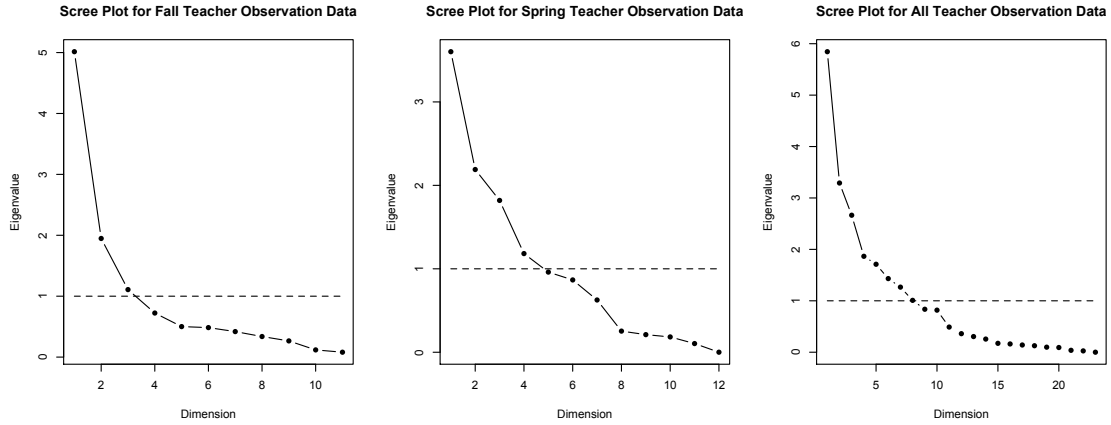
***Items 14-24, 28, 31, 32

Table 32: Teacher Practices Reliability Measures $\alpha = 0.76$, $glb = 0.81$, $\omega_t = 0.708$, H=0.304

item	$\alpha_{[-i]}$	$glb_{[-i]}$	$\omega_{t,[-i]}$	$H_{[-i]}$
4	0.740	0.80	0.69	0.30
5	0.738	0.80	0.69	0.30
6	0.727	0.79	0.66	0.29
7	0.739	0.79	0.68	0.30
8	0.739	0.77	0.68	0.31
9	0.743	0.79	0.69	0.31
10	0.729	0.75	0.69	0.31
12	0.744	0.79	0.67	0.30
13	0.743	0.79	0.68	0.31

A.5 Teacher Observation Exploratory Factor Analysis

To determine a possible number of factors, we plot scree plots. We include scree plots for the fall and spring data separately and one for the combined data. The observation data suggest a 2 or 3-factor model for the fall and a 2, 4 or 5-factor model spring data separately and a 4 or 8 factor model for the combined data.



We include factor loadings for 2-factor model for the fall observation data (Table 33). We see that the levels of engagement and housekeeping (eng1, eng2, eng3, timeinstruct, timehouse) load together and that teacher ratings (bestpractice, stdlearn, stdcenter) also load together.

Table 33: Exploratory Factor Analysis on Fall Teacher Observation Data: 2-Factor Model with No Rotation and Oblique Rotation

	No Rotation		Oblique Rotation	
Item	Factor1	Factor2	Factor1	Factor2
fall.eng1	0.42	0.56	-0.07	0.72
fall.eng2	0.55	0.71	-0.07	0.92
fall.eng3	0.71	0.32	0.35	0.57
fall.timeinstruct	0.61	0.30	0.28	0.51
fall.timehouse	-0.54	-0.45	-0.11	-0.65
fall.pctwhole	-0.25	0.51	-0.57	0.44
fall.pctgroup	0.41	-0.21	0.48	-0.08
fall.room	0.21	0.15	0.07	0.23
fall.bestpractice	0.94	-0.03	0.79	0.29
fall.stdlearn	0.91	-0.10	0.81	0.20
fall.stdcenter	0.93	-0.25	0.93	0.06

For the spring observation data, we include the factor loadings for a 4-factor model with no rotation and oblique rotation (Table 34, Table 35). These models suggest that levels of engagement (eng1, eng2, eng3) load together and that teacher ratings (bestpractice, stdlearn, stdcenter) also load together. Evidence that other variables load together appears contradictory and we use only these two factors from the spring data.

Table 34: Exploratory Factor Analysis on Spring Teacher Observation Data: 4-Factor Model with No Rotation

Item	Factor1	Factor2	Factor3	Factor4
spr.eng1	0.75	0.18	0.19	-0.36
spr.eng2	0.84	0.16	-0.04	-0.31
spr.eng3	0.78	0.02	0.17	-0.42
spr.timeinstruct	0.19	0.05	0.02	0.12
spr.timehouse	-0.19	-0.18	0.05	-0.12
spr.pctwhole	0.02	-0.64	-0.76	0.01
spr.pctgroup	0.00	1.00	-0.04	0.00
spr.pctalone	0.00	-0.54	0.84	0.00
spr.room	-0.04	0.21	-0.38	0.06
spr.bestpractice	0.49	0.39	0.17	0.65
spr.stdlearn	0.48	0.21	0.28	0.64
spr.stdcenter	0.31	0.42	0.29	0.64

For the both sets of observation data combined, we include the factor loadings for a 8-factor model with no rotation (Table 36). While there is clear evidence that there are not 8 factors, the extra variables make finding simple structure difficult. Instead, we focus only on the data that loaded together in the fall and spring factor models. We fit a 4-factor model on these data (Table 37) which suggests that the loading structure is the same when the fall and spring data are modeled separately or together.

Thus, we have two constructs for the fall data and two constructs for the spring data, but we calculate reliability measures to determine an optimal set for each factor (Table 38). Based on these reliability estimates and on the spring factors chosen, we choose to use a second fall factor containing only the levels of student engagement.

Table 35: Exploratory Factor Analysis on Spring Teacher Observation Data: 4-Factor Model with Oblique Rotation

Item	Factor1	Factor2	Factor3	Factor4
spr.eng1	-0.01	0.85	-0.04	-0.12
spr.eng2	0.06	0.89	0.15	0.09
spr.eng3	-0.07	0.91	-0.10	0.00
spr.timeinstruct	0.22	0.08	0.01	0.07
spr.timehouse	-0.21	-0.09	-0.14	0.00
spr.pctwhole	-0.02	0.00	0.31	0.98
spr.pctgroup	0.00	0.02	0.57	-0.74
spr.pctalone	0.04	-0.01	-1.01	-0.11
spr.room	0.03	-0.07	0.44	0.08
spr.bestpractice	0.91	0.03	0.06	0.02
spr.stdlearn	0.89	0.03	-0.13	0.08
spr.stdcenter	0.82	-0.11	-0.03	-0.15

Table 36: Exploratory Factor Analysis on Full Combined Teacher Observation Data: 8-Factor Model with Oblique Rotation

Item	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8
spr.eng1	-0.06	0.79	0.23	0.16	-0.08	-0.33	-0.20	-0.19
spr.eng2	-0.06	0.86	0.17	-0.06	-0.03	-0.21	-0.09	0.166
spr.eng3	-0.16	0.78	0.05	0.14	-0.12	-0.27	-0.16	0.19
spr.timeinstruct	0.13	0.19	0.08	0.04	-0.12	0.36	-0.05	0.23
spr.timehouse	-0.12	-0.16	-0.19	0.11	0.01	0.11	-0.20	-0.05
spr.pctwhole	0.07	0.03	-0.65	-0.75	-0.01	0.00	0.01	-0.01
spr.pctgroup	-0.03	0.00	1.00	-0.04	0.02	0.00	0.00	0.00
spr.pctalone	-0.03	0.00	-0.58	0.81	0.00	0.00	0.00	0.00
spr.room	0.06	-0.03	0.20	-0.35	-0.01	0.08	-0.11	-0.27
spr.bestpractice	0.37	0.42	0.40	0.31	-0.20	0.19	0.44	-0.08
spr.stdlearn	0.44	0.35	0.23	0.39	-0.16	0.15	0.48	0.03
spr.stdcenter	0.43	0.21	0.45	0.33	-0.39	0.17	0.41	-0.11
fall.eng1	0.23	0.32	-0.13	0.19	0.47	0.29	-0.18	-0.44
fall.eng2	0.33	0.32	0.02	0.05	0.63	0.38	0.12	-0.01
fall.eng3	0.63	0.21	-0.17	-0.05	0.39	0.33	0.00	0.38
fall.timeinstruct	0.41	0.22	-0.04	0.26	0.23	0.32	-0.27	-0.03
fall.timehouse	-0.43	-0.24	0.01	-0.19	-0.61	-0.36	0.10	0.24
fall.pctwhole	-0.46	0.19	-0.20	-0.17	0.45	-0.12	0.31	0.17
fall.pctgroup	0.45	0.09	0.19	0.14	-0.55	0.49	-0.34	0.03
fall.room	-0.11	0.07	0.05	-0.12	-0.15	0.62	-0.06	0.03
fall.bestpractice	0.85	0.05	0.04	0.16	0.19	0.04	0.00	-0.02
fall.stdlearn	0.80	0.11	-0.04	0.19	0.12	-0.08	0.01	0.06
fall.stdcenter	0.97	-0.07	0.09	0.13	0.00	-0.12	-0.02	0.00

Table 37: Exploratory Factor Analysis on Selected Teacher Observation Data: 4-Factor Model with Oblique Rotation

Item	Factor1	Factor2	Factor3	Factor4
fall.eng1	-0.04	0.68	0.08	-0.06
fall.eng2	-0.22	1.06	-0.04	0.09
fall.eng3	0.33	0.57	0.05	-0.10
fall.timeinstruct	0.45	0.16	0.09	0.01
fall.timehouse	-0.11	-0.65	0.09	0.07
fall.bestpractice	0.87	0.10	-0.04	0.02
fall.stdlearn	0.92	0.00	0.04	0.00
fall.stdcenter	1.01	-0.20	-0.06	0.10
spr.bestpractice	-0.05	0.17	0.06	0.87
spr.stdlearn	0.09	0.13	0.06	0.75
spr.stdcenter	0.10	-0.12	-0.09	0.97
spr.eng1	0.01	-0.14	0.82	0.02
spr.eng2	-0.13	0.13	0.88	0.04
spr.eng3	0.06	-0.09	0.91	-0.08

Table 38: Reliability Measures for Fall and Spring Teacher Observation Factors

Construct	Items	α	glb	ω_t	H
Fall Construct 1	bestpractice, stdlearn, stdcenter	0.94	0.94	0.93	0.94
Fall Construct 2a	eng1, eng2, eng3, timeinstruct, timehouse	-1.33	0.71	0.99	NA
Fall Construct 2b	eng1, eng2, eng3, timeinstruct	0.04	0.87	0.99	NA
Fall Construct 2c	eng1, eng2, eng3	0.76	0.83	-3.26	NA
Spring Construct 1	bestpractice, stdlearn, stdcenter	0.93	0.94	0.92	0.92
Spring Construct 2	eng1, eng2, eng3	0.89	0.90	-218.9	NA

A.6 Preliminary Models

We first fit models to determine whether to include race in our models. We fit a two level HLM for each outcome variable and include only those districts with mostly complete cases for race. Table 39 - Table 41 show regression coefficients for each model and we see that race is not significant. While AIC and BIC scores do not necessarily select out race, there is evidence that race is not relevant for these models.

Table 39: Winter CTBS as Response in 2 Level HLM with 4 Districts

Variable	Estimate (SE)
Fall OH	0.24 (0.03)
Fall CTBS	0.50 (0.03)
Stretch	-0.27 (0.08)
Race 0	0.52 (0.71)
Race 1	0.29 (0.71)
Race 2	0.33 (0.72)
Race 3	0.25 (0.73)
Race 4	0.59 (0.73)
Race 5	0.48 (0.73)

AIC = 1915.9, BIC=1917.2 vs. AIC = 1921.8, BIC=1923.1 for excluding race

Table 40: Student Attitude as Response in 2 Level HLM with 4 Districts

Variable	Estimate (SE)
Fall OH	0.14 (0.04)
Spring Alg TN	0.29 (0.04)
Stretch	0.09 (0.16)
Race 0	-0.34 (0.82)
Race 1	-0.04 (0.81)
Race 2	-0.14 (0.82)
Race 3	0.01 (0.86)
Race 4	0.16 (0.87)
Race 5	-0.17 (0.84)
Teacher Practice	0.37 (0.04)

AIC = 1624.7, BIC=1626.0 vs. AIC = 1633.9.0, BIC=1635.2 for excluding race

We now investigate models involving gender. We use our standard data set of 1233 students to fit a model to predict gender. We fit a logistic regression to predict gender based on the other student level covariates and we fail to find any correlations between gender and student level variables (Table 42).

Furthermore, we use these data to fit models for each outcome variable that include and exclude gender. AIC and BIC scores do not prefer including gender in each model as they do not decrease

Table 41: Spring Algebra TN as Response in 2 Level HLM with 4 Districts

Variable	Estimate (SE)
Fall OH	0.19 (0.04)
Winter CTBS	0.19 (0.04)
Stretch	0.21 (0.17)
Race 0	-0.86 (0.80)
Race 1	-0.78 (0.80)
Race 2	-0.43 (0.81)
Race 3	-0.83 (0.85)
Race 4	-0.56 (0.85)
Race 5	-0.84 (0.83)
Student Attitude	0.28 (0.04)
Teacher Practice	-0.13 (0.04)

AIC = 1528.4, BIC=1529.6 vs. AIC = 1534.0, BIC=1539.3 for excluding race

Table 42: *Standard Dataset* n=1233: Regression Coefficients for Model Predicting Gender

Variable	Estimate (SE)
Fall CTBS	-0.05 (0.07)
Fall OH	0.08 (0.06)
Winter CTBS	0.004 (0.06)
Spring Algebra TN	-0.10 (0.06)
Student Attitude	0.03 (0.06)
Teacher Practice	0.05 (0.06)

by more than 2 by including gender (Table 43 - Table 45). As gender is not correlated with other variables, we choose not to include gender in our final models.

Table 43: Winter CTBS as Response in 2 Level HLM with Standard Data Set

Variable	Estimate (SE)
Fall OH	0.30 (0.03)
Fall CTBS	0.28 (0.03)
Stretch	-0.18 (0.07)
Male	0.10 (0.05)

AIC = 3212.3, BIC=3214.8 vs. AIC = 3212.1, BIC=3214.6 for excluding gender

Table 44: Student Attitude as Response in 2 Level HLM with 4 Districts

Variable	Estimate (SE)
Fall OH	0.11 (0.03)
Spring Alg TN	0.27 (0.03)
Stretch	-0.05 (0.11)
Male	-0.01 (0.05)
Teacher Practice	0.33 (0.03)

AIC = 3141.4, BIC=3143.9 vs. AIC = 3137.3, BIC=3139.9 for excluding gender

Table 45: Spring Algebra TN as Response in 2 Level HLM with 4 Districts

Variable	Estimate (SE)
Fall OH	0.17 (0.03)
Winter CTBS	0.15 (0.03)
Stretch	-0.02 (0.18)
Male	-0.11 (0.05)
Student Attitude	0.24 (0.03)
Teacher Practice	-0.07 (0.03)

AIC = 2976.6, BIC=2979.1 vs. AIC = 2977.7, BIC=2980.2 for excluding gender

A.7 Variable Selection

With the entire data set, we select models by comparing AIC and BIC scores using a 2-level HLM (student/teacher). We use a somewhat greedy search technique in that we begin with all covariates in the model and retain only variables whose removal/addition does not significantly increase/decrease the AIC/BIC score by 2 or more. We consider two-way interactions only. We do not impute missing data in these models; rather we use complete cases only and include missing values as separate factors for ordinal variables only.

We begin with all variables in the model as fixed effects, and then we determine random effects in the model. We consider a random intercept model and a random intercept and slope model for test scores. We highlight the random effects models that minimizes AIC and BIC, and then we begin searching for interactions. We consider two-way interactions among student level variables only. If we find a better model by including interactions, we would highlight that model in pink. We then continue with the highlighted model by removing variables one at a time as fixed effects. Any variable that does not increase the AIC/BIC is removed. We then fit the model with these variables removed we repeat the process of removing fixed effects one at time again until there aren't any variables whose removal increases the AIC or BIC score. We highlight iterations of this process in pink.

We first outline variable selection for the Winter CTBS model. We see that the random intercept only model is best (Table 46). The next section looks at interactions and we see that including interactions fail to decrease the AIC/BIC so the highlighted pink model is still optimal. The next part of the table shows scores for eliminating student and teacher level variables. The next highlighted model is the model that removes all variables whose removal did not significantly increase the AIC/BIC. The models that follow are all models that removed those variables (frl, race, yrs exp, block, coach, vol) and one additional variable. We use the symbol + to represent the variables that we have already removed. The process is repeated with the variables that have not been removed. Based on these results, we see that all variables except "dd" (doubledose experience) can be removed. The final model is a random intercept model that includes Fall CTBS, Fall OH, condition, and dd.

For the Student Attitude model, we follow the same variable selection techniques (Table 47). Note that in this model selection procedure, race is selected to remain in the model based on AIC and BIC scores, but since race was not significant according to a t-test for regression coefficients and was missing for at least one full district and half of another, we decide not to include race in this model.

For the Spring Algebra TN model, we follow the same variable selection procedure (Table 48). Notice in this model that the model selected included an interaction between the fall and winter CTBS scores. With our standard data set, we use the student level variables selected and perform additional selection procedures to refine our models. With the Winter CTBS model, our variable selection procedure included one teacher level variable, experience with a double-dose course. We use our standard data set to compare AIC/BIC scores for including this variable for each imputation. The difference in AIC/BIC scores average 3.8 and 3.4 in favor of not including this teacher level variable, so we exclude it. For the Student Attitude model, there is also one teacher level variable, the number of years having taught math. Again, we compare AIC/BIC scores for including this variable for each imputation and they average 5.7 and 5.7 in favor of not including this teacher level variable.

For the Spring Algebra TN model, there are several models to compare (Table 49). Given this data set, there is some evidence of an interaction between the Fall and Winter CTBS scores on the outcome variable but for simplicity, we choose to exclude this interaction. Finally, we use this data set to determine the number of levels in each model (Table 50).

Table 46: Variable Selection Process for Winter CTBS Model

Random Effect	Interactions	Fixed Effect Removed	AIC	BIC
Fall CTBS			1386.3	1385.9
Fall OH			1389	1388.7
Intercept			1368.8	1368.6
Intercept	Fall CTBS*Stretch		1372.4	1372.2
Intercept	Fall CTBS*speced		1372.1	1371.9
Intercept	Fall CTBS*frl		1372.5	1372.3
Intercept	Fall CTBS*female		1369.6	1369.5
Intercept	Fall CTBS*race		1374.4	1374.3
Intercept	Fall CTBS*ohraw1		1374	1373.8
Intercept		FallOH	1477.5	1477.3
Intercept		FallCTBS	1387.1	1387
Intercept		speced	1369.2	1369.1
Intercept		frl	1362.2	1362.1
Intercept		female	1369.3	1369.2
Intercept		race	1366.7	1366.6
Teacher Level Variables				
Intercept		yrs exp	1365.9	1365.8
Intercept		yrs math	1368.4	1368.3
Intercept		yrs alg	1368.1	1367.9
Intercept		blockexp	1365.7	1365.6
Intercept		coach	1366.4	1366.3
Intercept		doubledose	1368.4	1368.3
Intercept		volunteer	1366.8	1366.6
Intercept		major	1367.9	1367.7
Intercept		cert	1367.1	1367
Intercept		frl, race, yrs exp, block, coach, vol	1343.3	1343.2
Intercept		+speced	1339.5	1339.4
Intercept		+female	1343.8	1343.7
Intercept		+yrs math	1344.1	1344
Intercept		+yrs alg	1343	1342.9
Intercept		+doubledose	1344.6	1344.5
Intercept		+major	1343.9	1343.8
Intercept		+cert	1341.3	1341.2
Intercept		frl, race, yrs exp, block, coach, vol, speced, yrs alg, cert	1334	1333.8
Intercept		+female	1333.6	1333.5
Intercept		+yrs math	1331.6	1331.5
Intercept		+dd	1336.5	1336.4
Intercept		+major	1334.3	1334.2

Table 47: Variable Selection Process for Student Attitude Model

Random Effect	Interactions	Fixed Effect Removed	AIC	BIC
Fall CTBS			1584.1	1583.7
Fall OH			1589.6	1589.2
Winter CTBS			1577.2	1576.8
Spring Algebra TN			1579.5	1579.2
Intercept			1571.6	1571.4
Intercept	FallCTBS *WinterCTBS		1575.2	1575
Intercept	FallCTBS *SprAlgTN		1572.7	1572.6
Intercept	FallCTBS *FallOH		1576.5	1576.3
Intercept	WinterCTBS*SprAlgTN		1570.8	1570.6
Intercept	WinterCTBS*FallOH		1572.6	1572.4
Intercept	FallOH*SprAlgTN		1574.4	1574.2
Intercept		FallCTBS	1568.5	1568.3
Intercept		FallOH	1575.2	1575
Intercept		WinterCTBS	1571.3	1571.1
Intercept		SpringAlgTN	1600.7	1600.6
Intercept		speced	1572.7	1572.6
Intercept		frl	1569.5	1569.4
Intercept		female	1568.1	1567.9
Intercept		race	1576.2	1576
Intercept		yrs exp	1569.5	1569.3
Intercept		yrs math	1572.2	1572.1
Intercept		yrs alg	1571.8	1571.6
Intercept		blockexp	1571.1	1570.9
Intercept		coach	1570.5	1570.3
Intercept		doubledose	1570.8	1570.7
Intercept		volunteer	1570.6	1570.4
Intercept		major	1571.5	1571.3
Intercept		cert	1570.9	1570.7
Intercept		teacher practice	1621.7	1621.5
Intercept		FallCTBS, Winter CTBS, frl, female, yrs exp, block, coach, doubledose, volunteer, major, cert	1547.1	1546.9
Intercept		+FallOH	1553.8	1553.7
Intercept		+SpringAlgTN	1581.1	1581
Intercept		+speced	1547.8	1547.7
Intercept		+race	1551.7	1551.6
Intercept		+yrs math	1549.2	1549.1
Intercept		+yrs alg	1549	1548.8
Intercept		+tptheta	1605.1	1604.9

Table 48: Variable Selection Process for Spring Alg TN Model

Random Effect	Interactions	Fixed Effect Removed	AIC	BIC
FallOH			1533.8	1533.5
FallCTBS			1525.9	1525.5
WinterCTBS			1537.5	1537.2
Intercept			1524.5	1524.3
Intercept	FallCTBS*FallOH		1523.5	1523.4
Intercept	FallCTBS*WinterCTBS		1519.2	1519
Intercept	FallCTBS*WinterCTBS, FallCTBS*Student Attitude		1518.4	1518.3
Intercept	FallCTBS*WinterCTBS	FallOH	1524.4	1524.2
Intercept	FallCTBS*WinterCTBS	speced	1517	1516.8
Intercept	FallCTBS*WinterCTBS	frl	1517.6	1517.4
Intercept	FallCTBS*WinterCTBS	female	1522.5	1522.3
Intercept	FallCTBS*WinterCTBS	race	1533.7	1533.6
Intercept	FallCTBS*WinterCTBS	yrs exp	1517.3	1517.1
Intercept	FallCTBS*WinterCTBS	yrs math	1520.3	1520.2
Intercept	FallCTBS*WinterCTBS	yrs alg	1520.1	1519.9
Intercept	FallCTBS*WinterCTBS	blockexp	1519.8	1519.7
Intercept	FallCTBS*WinterCTBS	coach	1518.6	1518.5
Intercept	FallCTBS*WinterCTBS	doubledose	1519.8	1519.6
Intercept	FallCTBS*WinterCTBS	volunteer	1518.2	1518.1
Intercept	FallCTBS*WinterCTBS	major	1518.5	1518.3
Intercept	FallCTBS*WinterCTBS	cert	1519.6	1519.4
Intercept	FallCTBS*WinterCTBS	student attitude	1551.6	1551.5
Intercept	FallCTBS*WinterCTBS	teacher practice	1518.6	1518.4
Intercept	FallCTBS*WinterCTBS	speced, yrs exp, frl	1510.5	1510.3
Intercept	FallCTBS*WinterCTBS	+female	1514.5	1514.3
Intercept	FallCTBS*WinterCTBS	+yrs math	1511.3	1511.1
Intercept	FallCTBS*WinterCTBS	+yrs alg	1511	1510.8
Intercept	FallCTBS*WinterCTBS	+blockexp	1510.7	1510.5
Intercept	FallCTBS*WinterCTBS	+coach	1509.5	1509.3
Intercept	FallCTBS*WinterCTBS	+doubledose	1511.7	1511.5
Intercept	FallCTBS*WinterCTBS	+volunteer	1509.2	1509.1
Intercept	FallCTBS*WinterCTBS	+major	1509.5	1509.3
Intercept	FallCTBS*WinterCTBS	+cert	1510.2	1510
Intercept	FallCTBS*WinterCTBS	+teacher practice	1509.9	1509.7
Intercept	FallCTBS*WinterCTBS	speced, yrs exp, yrs math, yrs alg, frl, teacher practice, blockexp, coach, volunteer, major, cert	1503.2	1503.1
Intercept	FallCTBS*WinterCTBS	+FallOH	1510.7	1510.6
Intercept	FallCTBS*WinterCTBS	+female	1507.3	1507.1
Intercept	FallCTBS*WinterCTBS	+race	1517.9	1517.7
Intercept	FallCTBS*WinterCTBS	+doubledose	1503.5	1503.4
Intercept	FallCTBS*WinterCTBS	+student attitude	1531	1530.9

Table 49: Spring Algebra TN Model Variable Selection Using Standard Imputation Data Sets

Model	Interaction	AIC	BIC
Fall CTBS,		2992.7	2997.3
Winter CTBS,		2989.7	2994.4
Fall OH, Stretch,	FallCTBS*WinterCTBS	2991.9	2996.6
SAttitude,		2995.3	3000
TPractice		2987.5	2992.2
Fall CTBS,		3004.6	3009.2
Winter CTBS,		3001.3	3005.9
Fall OH, Stretch,	No Interaction	3004.5	3009.1
SAttitude,		3004.6	3009.2
TPractice		3001.3	3006
		3000.2	3004.8
Winter CTBS,		2997.0	3001.7
Fall OH, Stretch,	FallCTBS*WinterCTBS	3000.1	3004.7
SAttitude,		3000.3	3005
TPractice		2997.1	3001.8

Table 50: AIC/BIC Scores for 2 and 3 Level Random Intercept Models for Each Response Variables

Winter CTBS Model: AIC/BIC Scores for 2 and 3 Level Random Intercept Models

Imputation	2 Level (Teacher)		2 Level (School)		3 Level (Teacher $\subset$ School)		3 Level (Teacher $\subset$ District)	
1	3319.9	3324.6	3318.7	3321.4	3320.7	3324.7	3319.9	3320.1
2	3317.2	3321.9	3315.4	3318.1	3315.4	3318.1	3319.2	3319.5
3	3316.4	3321.0	3314.1	3316.8	3316.1	3320.1	3316.4	3316.5
4	3310.8	3315.4	3308.8	3311.5	3310.8	3314.8	3310.8	3310.9
5	3316.4	3321.1	3314.9	3317.6	3314.9	3317.6	3318.4	3318.7

Student Attitude Model: AIC/BIC Scores for 2 and 3 Level Random Intercept Models

Imputation	2 Level (Teacher)		2 Level (School)		3 Level (Teacher $\subset$ School)		3 Level (Teacher $\subset$ District)	
1	3225.5	3230.1	3238.2	3240.9	3223.8	3227.8	3224.8	3225.0
2	3223.7	3228.4	3236.9	3239.5	3222.2	3226.2	3223.1	3223.3
3	3225.2	3229.8	3238.2	3240.9	3223.6	3227.6	3224.5	3224.8
4	3225.5	3230.1	3238.3	3241.0	3223.9	3227.9	3224.9	3225.1
5	3223.8	3228.4	3239.4	3239.4	3222.2	3226.2	3223.2	3223.4

Spring Algebra TN Model: AIC/BIC Scores for 2 and 3 Level Random Intercept Models

Imputation	2 Level (Teacher)		2 Level (School)		3 Level (Teacher $\subset$ School)		3 Level (Teacher $\subset$ District)	
1	3000.7	3005.3	3061.9	3064.6	2993.2	2997.2	2984.1	2984.4
2	2997.5	3002.1	3057.0	3059.7	2989.8	2993.8	2981.1	2981.3
3	3000.5	30005.2	3061.4	3064.1	2992.9	2996.9	2983.9	2984.1
4	3000.8	3005.4	3062.6	3065.3	2993.3	2997.3	2984.0	2984.3
5	2997.6	3002.3	3058.0	3060.7	2990.1	2994.1	2981.0	2981.2

A.8 Missing Data Imputation Regression Model

We use regression models to create imputations for missing data. As we are missing Fall OH, Fall CTBS, and Winter CTBS scores only, we describe in detail how these missing values were imputed.

To impute Fall OH scores, we fit the following linear model to determine estimates for β_0 and β_1.

$$FallOH_i = \beta_0 + \beta_1 FallCTBS_i$$

To generate missing data, we sampled in the following way:

$$FallOH_i^{mis} \sim N(\beta_0 + \beta_1 FallCTBS_i, \sigma^2)$$

where σ^2 is the residual variance. The same procedure was used to impute Fall CTBS scores.

For Winter CTBS scores, we fit the following linear model to determine estimates for β_0, ...β_4.

$$WinterCTBS_i = \beta_0 + \beta_1 FallCTBS_i + \beta_2 FallOH_i + \beta_3 Stretch_i + \beta_4 SpringAlgTN_i$$

To generate missing data, we sampled in the following way:

$$WinterCTBS_i^{mis} \sim N(beta_0 + \beta_1 FallCTBS_i + \beta_2 FallOH_i + \beta_3 Stretch_i + \beta_4 SpringAlgTN_i, \sigma^2)$$

where σ^2 is the residual variance.

A.9 Statistical Models

In this section, we present additional tables and figures presenting parameter estimates for each outcome variable.

A.9.1 Winter CTBS

Table 51 gives the parameter estimates for the Winter CTBS model for each fitting method.

Table 51: Winter CTBS Score Model: Regression Coefficients and Variance Components for Two Methods of Estimation SAS and JAGS

Covariate	SAS Estimate (SE)	JAGS Estimate (SE)
Pretest CTBS	0.21 (0.03)	0.21 (0.03)
Pretest OH	0.29 (0.03)	0.29 (0.03)
Condition	-0.19 (0.06)	-0.18 (0.07)
Student Variance	0.84 (0.03)	0.92 (0.02)
School Variance	0.01 (0.01)	0.11 (0.05)

A.9.2 Student Attitude

Table 52: Student Attitude Model: Regression Coefficients and Variance Components for Each Method of Latent Variable Estimation (Sum Scores, IRT, JAGS)

			Sum Scores Estimate (SE)	IRT Estimate (SE)	JAGS Estimate (SE)
Regression Coefficients	Variable	Fall OH	0.11 (0.03)	0.11 (0.03)	0.04 (0.01)
		Spring Alg	0.27 (0.03)	0.27 (0.03)	0.09 (0.02)
		Teacher Practice	0.34 (0.03)	0.34 (0.03)	0.22 (0.02)
		Condition = Stretch	-0.05 (0.11)	-0.05 (0.11)	-0.02 (0.04)
Variance Components	Level	Student	0.74 (0.03)	0.74 (0.03)	0.28 (0.01)
		Teacher	0.05 (0.02)	0.05 (0.02)	0.08 (0.02)
		School	0.04 (0.02)	0.04 (0.02)	0.07 (0.03)

A.9.3 Spring Algebra TN

Table 53: Spring Algebra TN Model: Regression Coefficients and Variance Components for Each Method of Latent Variable Estimation (Sum Scores, IRT, JAGS)

			Sum Scores Estimate (SE)	IRT Estimate (SE)	JAGS Estimate (SE)
Regression Coefficients	Variable	Fall OH	0.16 (0.03)	0.16 (0.03)	0.16 (0.03)
		Winter CTBS	0.14 (0.03)	0.14 (0.02)	0.13 (0.02)
		Student Attitude	0.23 (0.03)	0.23 (0.03)	0.41 (0.04)
		Teacher Practice	-0.05 (0.03)	-0.06 (0.03)	-0.07 (0.04)
		Condition = Stretch	-0.01 (0.10)	-0.01 (0.10)	-0.002 (0.10)
Variance Components	Level	Student	0.58 (0.02)	0.58 (0.02)	0.76 (0.02)
		Teacher	0.14 (0.03)	0.14 (0.03)	0.39 (0.04)
		District	0.17 (0.011)	0.17 (0.11)	0.40 (0.20)

A.10 Sensitivity Analyses

Table 54: Effects of Gibbs Sampler Priors on Parameter Estimates

Variable	Standardized Regression Coefficients		
	Flat Priors	Regular Priors	Strong Priors
Fall OH	0.12 (0.03)	0.05 (0.02)	0.06 (0.01)
Spring Alg TN	0.27 (0.04)	0.12 (0.03)	0.13 (0.02)
Teacher Practice	0.46 (0.05)	0.24 (0.03)	0.24 (0.02)
Stretch	-0.05 (0.11)	-0.04 (0.05)	-0.03 (0.05)
	Variance Components		
	Flat Priors	Regular Priors	Strong Priors
Student	0.81 (0.07)	0.28 (0.01)	0.39 (0.02)
Teacher	0.21 (0.05)	0.08 (0.02)	0.11 (0.02)
School	0.17 (0.07)	0.06 (0.03)	0.09 (0.03)

A.11 JAGS Codes

Winter CTBS Model

```
model{
for(i in 1:students){
mathrw2[i]~dnorm(mu[i],tau1)
mu[i]<- b0[schoolnumber[i]]+ b1*mathrw1[i] + b2*ohraw1[i] +b3*condition[i]
}

tau1 <- pow(sigma, (-2))
sigma~dunif(0, 1)
b1~dnorm(0, 1)
b2~dnorm(0, 1)
b3~dnorm(0, 1)

for(k in 1:schools){
b0[k]~dnorm(eta, tau2)}

eta~dnorm(0, 0.1)
tau2 <-pow(U1.tau, (-2))
U1.tau~dunif(0, 1)
}
```

Student Attitude Model

```
model{
#######SATHETA#######
for(i in 1:students){
for(j in 1:itemssa){
snum[i,j,1]<- exp(sa.a[j]*(satheta[i]-sa.c[j]+sa.d[j,1]))
snum[i,j,2]<- exp(sa.a[j]*(satheta[i]-sa.c[j]+sa.d[j,1]) + sa.a[j]*(satheta[i]-sa.c[j]+sa.d[j,2]))
snum[i,j,3]<- exp(sa.a[j]*(satheta[i]-sa.c[j]+sa.d[j,1]) + sa.a[j]*(satheta[i]-sa.c[j]+sa.d[j,2]) +
 sa.a[j]*(satheta[i]-sa.c[j]-sa.d[j,1]-sa.d[j,2]))
sdenom[i,j] <- 1+sum(snum[i,j,1:3])

sp.star[i,j,1] <- 1/sdenom[i,j]
sp.star[i,j,2] <- snum[i,j,1]/sdenom[i,j]
sp.star[i,j,3] <- snum[i,j,2]/sdenom[i,j]
sp.star[i,j,4] <- snum[i,j,3]/sdenom[i,j]

responsesa[i,j]~dcat(sp.star[i,j,1:4])
}

#####TPTHETA########
for(j in 1:itemstp){

tnum[i,j,1]<- exp(tp.a[j]*(tptheta[i]-tp.c[j]+tp.d[j,1]))
tnum[i,j,2]<- exp(tp.a[j]*(tptheta[i]-tp.c[j]+tp.d[j,1]) + tp.a[j]*(tptheta[i]-tp.c[j]+tp.d[j,2]))
tnum[i,j,3]<- exp(tp.a[j]*(tptheta[i]-tp.c[j]+tp.d[j,1]) + tp.a[j]*(tptheta[i]-tp.c[j]+tp.d[j,2]) +
tp.a[j]*(tptheta[i]-tp.c[j]-tp.d[j,1]-tp.d[j,2]))
tdenom[i,j] <- 1+sum(tnum[i,j,1:3])

tp.star[i,j,1] <- 1/tdenom[i,j]
```

```
tp.star[i,j,2] <- tnum[i,j,1]/tdenom[i,j]
tp.star[i,j,3] <- tnum[i,j,2]/tdenom[i,j]
tp.star[i,j,4] <- tnum[i,j,3]/tdenom[i,j]

responsetp[i,j]~dcat(tp.star[i,j,1:4])
}

tptheta[i]~dnorm(0, 1.0)
satheta[i]~dnorm(mu[i], tau1)
######HLM Model########
mu[i]<- b1*tptheta[i] +b2*mathrw3[i] + b3*ohraw1[i] +b4*condition[i] + b0[newtch[i], schoolnumber[i]]
int[i]<-b0[newtch[i], schoolnumber[i]]
}

########PRIORS#######
for(j in 1:itemssa){
sa.a[j]~dunif(0,4)
sa.c[j]~dnorm(0, 1)
sa.d[j,1]~dnorm(0, 100)
sa.d[j,2]~dnorm(0, 100)}

for(j in 1:itemstp){
tp.a[j]~dunif(0,4)
tp.c[j]~dnorm(0, 1)
tp.d[j,1]~dnorm(0, 100)
tp.d[j,2]~dnorm(0, 100)}

for(k in 1:schools){
for(j in 1:numofteachers[k]){
b0[j,k] ~dnorm(eta[k], tau2)
}
eta[k]~dnorm(delta, tau3)
}

delta~dnorm(0, 5.0)

tau1 <- pow(sigma, (-2))
sigma~dunif(0, 10)
tau2 <- pow(U1.tau, (-2))
U1.tau~dunif(0, 10)
tau3 <- pow(U2.tau, (-2))
U2.tau~dunif(0, 10)

b1~dnorm(0, 1)
b2~dnorm(0, 1)
b3~dnorm(0, 1)
b4~dnorm(0, 1)
}
```

Spring Algebra TN Model

```
model{
#######SATHETA#########
for(i in 1:students){
for(j in 1:itemssa){
snum[i,j,1]<- exp(sa.a[j]*(satheta[i]-sa.c[j]+sa.d[j,1]))
snum[i,j,2]<- exp(sa.a[j]*(satheta[i]-sa.c[j]+sa.d[j,1]) + sa.a[j]*(satheta[i]-sa.c[j]+sa.d[j,2]))
snum[i,j,3]<- exp(sa.a[j]*(satheta[i]-sa.c[j]+sa.d[j,1]) + sa.a[j]*(satheta[i]-sa.c[j]+sa.d[j,2]) +
sa.a[j]*(satheta[i]-sa.c[j]-sa.d[j,1]-sa.d[j,2]))
sdenom[i,j] <- 1+sum(snum[i,j,1:3])

sp.star[i,j,1] <- 1/sdenom[i,j]
sp.star[i,j,2] <- snum[i,j,1]/sdenom[i,j]
sp.star[i,j,3] <- snum[i,j,2]/sdenom[i,j]
sp.star[i,j,4] <- snum[i,j,3]/sdenom[i,j]

responsesa[i,j]~dcat(sp.star[i,j,1:4])
}

#####TPTHETA########
for(j in 1:itemstp){
tnum[i,j,1]<- exp(tp.a[j]*(tptheta[i]-tp.c[j]+tp.d[j,1]))
tnum[i,j,2]<- exp(tp.a[j]*(tptheta[i]-tp.c[j]+tp.d[j,1]) + tp.a[j]*(tptheta[i]-tp.c[j]+tp.d[j,2]))
tnum[i,j,3]<- exp(tp.a[j]*(tptheta[i]-tp.c[j]+tp.d[j,1]) + tp.a[j]*(tptheta[i]-tp.c[j]+tp.d[j,2]) +
tp.a[j]*(tptheta[i]-tp.c[j]-tp.d[j,1]-tp.d[j,2]))
tdenom[i,j] <- 1+sum(tnum[i,j,1:3])

tp.star[i,j,1] <- 1/tdenom[i,j]
tp.star[i,j,2] <- tnum[i,j,1]/tdenom[i,j]
tp.star[i,j,3] <- tnum[i,j,2]/tdenom[i,j]
tp.star[i,j,4] <- tnum[i,j,3]/tdenom[i,j]

responsetp[i,j]~dcat(tp.star[i,j,1:4])
}

tptheta[i]~dnorm(0, 1.0)
satheta[i]~dnorm(0, 1.0)

######HLM Model#######
mathrw3[i]~dnorm(mu[i], tau1)
mu[i] <- b0[newtch2[i], district[i]] + b1*satheta[i]+ b2*tptheta[i] + b3*mathrw2[i] + b4*ohraw1[i]
+b5*condition[i]
int[i]<-b0[newtch2[i], district[i]]
}

########Priors#######
for(j in 1:itemssa){
sa.a[j]~dlnorm(0,1)
sa.c[j]~dnorm(0, 1)
sa.d[j,1]~dnorm(0, 100)
sa.d[j,2]~dnorm(0, 100)}

for(j in 1:itemstp){
tp.a[j]~dlnorm(0,1)
tp.c[j]~dnorm(0, 1)
```

```
tp.d[j,1]~dnorm(0, 100)
tp.d[j,2]~dnorm(0, 100)}

for(k in 1:8){
for(j in 1:teachersindis[k]){
b0[j,k] ~dnorm(eta[k], tau2)
}
eta[k]~dnorm(delta, tau3)
}
delta~dnorm(0, 5.0)

tau1 <- pow(sigma, (-2))
sigma~dunif(0, 100)
tau2 <- pow(U1.tau, (-2))
U1.tau~dunif(0, 100)
tau3 <- pow(U2.tau, (-2))
U2.tau~dunif(0, 100)

b1~dnorm(0, 1)
b2~dnorm(0, 1)
b3~dnorm(0, 1)
b4~dnorm(0, 1)
b5~dnorm(0, 1)
}
```

Missing Data Imputation in Student Attitude Model

```
model{
####Model for FallCTBS Missing#####
for(i in 1:students){
mathrw1[i]~dnorm(mu.rw1[i], tau.rw1)
mu.rw1[i] <- b0.rw1[schoolnumber[i]]
}
for(k in 1:28){
b0.rw1[k]~dnorm(0, 1)
}
tau.rw1 <- pow(sigma.rw1, (-2))
sigma.rw1~dunif(0, 1)

####Model for FallOH Missing####
for(i in 1:students){
ohraw1[i]~dnorm(mu.oh1[i], tau.oh1)
mu.oh1[i]<- b0.oh1[schoolnumber[i]]+ b1.oh1*mathrw1[i]
}
for(k in 1:28){
b0.oh1[k]~dnorm(0, 1)
}
b1.oh1~dnorm(0,1)
tau.oh1 <- pow(sigma.oh1, (-2))
sigma.oh1~dunif(0, 1)

###Model for WinterCTBS Missing#####
for(i in 1:students){
mathrw2[i]~dnorm(mu.rw2[i], tau.rw2)
mu.rw2[i]<- b0.rw2[schoolnumber[i]]+ b1.rw2*ohraw1[i]+b2.rw2*mathrw1[i]+b3*condition[i]
```

```
}
for(k in 1:28){
b0.rw2[k]~dnorm(0, 1)
}
b1.rw2~dnorm(0,1)
b2.rw2~dnorm(0,1)
b3.rw2~dnorm(0,1)
tau.rw2 <- pow(sigma.rw2, (-2))
sigma.rw2~dunif(0, 1)

####Full Model######
for(i in 1:students){
for(j in 1:itemssa){
snum[i,j,1]<- exp(sa.a[j]*(satheta[i]-sa.c[j]+sa.d[j,1]))
snum[i,j,2]<- exp(sa.a[j]*(satheta[i]-sa.c[j]+sa.d[j,1]) + sa.a[j]*(satheta[i]-sa.c[j]+sa.d[j,2]))
snum[i,j,3]<- exp(sa.a[j]*(satheta[i]-sa.c[j]+sa.d[j,1]) + sa.a[j]*(satheta[i]-sa.c[j]+sa.d[j,2]) +
sa.a[j]*(satheta[i]-sa.c[j]-sa.d[j,1]-sa.d[j,2]))
sdenom[i,j] <- 1+sum(snum[i,j,1:3])

sp.star[i,j,1] <- 1/sdenom[i,j]
sp.star[i,j,2] <- snum[i,j,1]/sdenom[i,j]
sp.star[i,j,3] <- snum[i,j,2]/sdenom[i,j]
sp.star[i,j,4] <- snum[i,j,3]/sdenom[i,j]
responsesa[i,j]~dcat(sp.star[i,j,1:4])}

#####TPTHETA########
for(j in 1:itemstp){
tnum[i,j,1]<- exp(tp.a[j]*(tptheta[i]-tp.c[j]+tp.d[j,1]))
tnum[i,j,2]<- exp(tp.a[j]*(tptheta[i]-tp.c[j]+tp.d[j,1]) + tp.a[j]*(tptheta[i]-tp.c[j]+tp.d[j,2]))
tnum[i,j,3]<- exp(tp.a[j]*(tptheta[i]-tp.c[j]+tp.d[j,1]) + tp.a[j]*(tptheta[i]-tp.c[j]+tp.d[j,2]) +
tp.a[j]*(tptheta[i]-tp.c[j]-tp.d[j,1]-tp.d[j,2]))
tdenom[i,j] <- 1+sum(tnum[i,j,1:3])
tp.star[i,j,1] <- 1/tdenom[i,j]
tp.star[i,j,2] <- tnum[i,j,1]/tdenom[i,j]
tp.star[i,j,3] <- tnum[i,j,2]/tdenom[i,j]
tp.star[i,j,4] <- tnum[i,j,3]/tdenom[i,j]
responsetp[i,j]~dcat(tp.star[i,j,1:4])}

tptheta[i]~dnorm(0, 1.0)
satheta[i]~dnorm(mu[i], tau1)
mu[i]<- b1*tptheta[i] +b2*mathrw3[i] + b3*ohraw1[i] +b4*condition[i] + b0[newtch[i], schoolnumber[i]]}

########Priors#######
for(j in 1:itemssa){
sa.a[j]~dunif(0,4)
sa.c[j]~dnorm(0, 1)
sa.d[j,1]~dnorm(0, 100)
sa.d[j,2]~dnorm(0, 100)}

for(j in 1:itemstp){
tp.a[j]~dunif(0,4)
tp.c[j]~dnorm(0, 1)
tp.d[j,1]~dnorm(0, 100)
tp.d[j,2]~dnorm(0, 100)}
```

```
for(k in 1:schools){
for(j in 1:numofteachers[k]){
b0[j,k] ~dnorm(eta[k], tau2)
}
eta[k]~dnorm(delta, tau3)
}
delta~dnorm(0, 5.0)

tau1 <- pow(sigma, (-2))
sigma~dunif(0, 10)
tau2 <- pow(U1.tau, (-2))
U1.tau~dunif(0, 10)
tau3 <- pow(U2.tau, (-2))
U2.tau~dunif(0, 10)

b1~dnorm(0, 1)
b2~dnorm(0, 1)
b3~dnorm(0, 1)
b4~dnorm(0, 1)
}
```

Missing Data Imputation in Spring Algebra TN Model

```
model{
####Model for FallCTBS Missing#####
for(i in 1:students){
mathrw1[i]~dnorm(mu.rw1[i], tau.rw1)
mu.rw1[i] <- b0.rw1[schoolnumber[i]]
}
for(k in 1:28){
b0.rw1[k]~dnorm(0, 1)
}
tau.rw1 <- pow(sigma.rw1, (-2))
sigma.rw1~dunif(0, 1)

####Model for FallOH Missing####
for(i in 1:students){
ohraw1[i]~dnorm(mu.oh1[i], tau.oh1)
mu.oh1[i]<- b0.oh1[schoolnumber[i]]+ b1.oh1*mathrw1[i]
}
for(k in 1:28){
b0.oh1[k]~dnorm(0, 1)
}
b1.oh1~dnorm(0,1)
tau.oh1 <- pow(sigma.oh1, (-2))
sigma.oh1~dunif(0, 1)

###Model for WinterCTBS Missing#####
for(i in 1:students){
mathrw2[i]~dnorm(mu.rw2[i], tau.rw2)
mu.rw2[i]<- b0.rw2[schoolnumber[i]]+ b1.rw2*ohraw1[i]+b2.rw2*mathrw1[i]+b3*condition[i]
}
for(k in 1:28){
b0.rw2[k]~dnorm(0, 1)
}
```

```
b1.rw2~dnorm(0,1)
b2.rw2~dnorm(0,1)
b3.rw2~dnorm(0,1)
tau.rw2 <- pow(sigma.rw2, (-2))
sigma.rw2~dunif(0, 1)

#####Actual Model######
for(i in 1:students){
for(j in 1:itemssa){
snum[i,j,1]<- exp(sa.a[j]*(satheta[i]-sa.c[j]+sa.d[j,1]))
snum[i,j,2]<- exp(sa.a[j]*(satheta[i]-sa.c[j]+sa.d[j,1]) + sa.a[j]*(satheta[i]-sa.c[j]+sa.d[j,2]))
snum[i,j,3]<- exp(sa.a[j]*(satheta[i]-sa.c[j]+sa.d[j,1]) + sa.a[j]*(satheta[i]-sa.c[j]+sa.d[j,2]) +
sa.a[j]*(satheta[i]-sa.c[j]-sa.d[j,1]-sa.d[j,2]))
sdenom[i,j] <- 1+sum(snum[i,j,1:3])

sp.star[i,j,1] <- 1/sdenom[i,j]
sp.star[i,j,2] <- snum[i,j,1]/sdenom[i,j]
sp.star[i,j,3] <- snum[i,j,2]/sdenom[i,j]
sp.star[i,j,4] <- snum[i,j,3]/sdenom[i,j]
responsesa[i,j]~dcat(sp.star[i,j,1:4])
}
#####TPTHETA########
for(j in 1:itemstp){
tnum[i,j,1]<- exp(tp.a[j]*(tptheta[i]-tp.c[j]+tp.d[j,1]))
tnum[i,j,2]<- exp(tp.a[j]*(tptheta[i]-tp.c[j]+tp.d[j,1]) + tp.a[j]*(tptheta[i]-tp.c[j]+tp.d[j,2]))
tnum[i,j,3]<- exp(tp.a[j]*(tptheta[i]-tp.c[j]+tp.d[j,1]) + tp.a[j]*(tptheta[i]-tp.c[j]+tp.d[j,2]) +
tp.a[j]*(tptheta[i]-tp.c[j]-tp.d[j,1]-tp.d[j,2]))
tdenom[i,j] <- 1+sum(tnum[i,j,1:3])

tp.star[i,j,1] <- 1/tdenom[i,j]
tp.star[i,j,2] <- tnum[i,j,1]/tdenom[i,j]
tp.star[i,j,3] <- tnum[i,j,2]/tdenom[i,j]
tp.star[i,j,4] <- tnum[i,j,3]/tdenom[i,j]
responsetp[i,j]~dcat(tp.star[i,j,1:4])
}
tptheta[i]~dnorm(0, 1.0)
satheta[i]~dnorm(0, 1.0)
mathrw3[i]~dnorm(mu[i], tau1)
mu[i] <- b0[newtch2[i], district[i]] + b1*satheta[i]+ b2*tptheta[i] + b3*mathrw2[i] + b4*ohraw1[i]
+b5*condition[i]
int[i]<-b0[newtch2[i], district[i]]
}

########Priors#######
for(j in 1:itemssa){
sa.a[j]~dlnorm(0,1)
sa.c[j]~dnorm(0, 1)
sa.d[j,1]~dnorm(0, 100)
sa.d[j,2]~dnorm(0, 100)}

for(j in 1:itemstp){
tp.a[j]~dlnorm(0,1)
tp.c[j]~dnorm(0, 1)
tp.d[j,1]~dnorm(0, 100)
tp.d[j,2]~dnorm(0, 100)}
```

```
for(k in 1:8){
for(j in 1:teachersindis[k]){
b0[j,k] ~dnorm(eta[k], tau2)
}
eta[k]~dnorm(delta, tau3)
}
delta~dnorm(0, 5.0)

tau1 <- pow(sigma, (-2))
sigma~dunif(0, 100)
tau2 <- pow(U1.tau, (-2))
U1.tau~dunif(0, 100)
tau3 <- pow(U2.tau, (-2))
U2.tau~dunif(0, 100)

b1~dnorm(0, 1)
b2~dnorm(0, 1)
b3~dnorm(0, 1)
b4~dnorm(0, 1)
b5~dnorm(0, 1)
}
```

References

Balfanz, R., Letgers, N., and Jordan, W. (2004). Catching up: the impact of the talent development ninth grade instructional interventions in reading and mathematics in high poverty high schools. Technical Report 69, Center for Research on the Education of Students Placed at Risk, Johns Hopkins University.

Balfanz, R. and Neild, R. (2006). Successful transitions to algebra 1: A randomized control trial of two theories of ninth grade algebra instruction. Proposal for Grant Number 101198, funded by NCER. Center for the Social Organization of Schools, Johns Hopkins University.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian data analysis*. Chapman and Hall/CRC.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., Meng, X., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807.

Johnson, M., Cohen, W., and Junker, B. (1999). Measuring appropriability in research and development with item response models. Technical report, Carnegie Mellon University.

Kemple, J., Herlihy, C., and Smith, T. (2005). Making progress toward graduation: evidence from talent development high school model. Technical report, MDRC.

Kilpatrick, J., Swafford, J., and Findell, B., editors (2001). *In Adding it up: Helping Children Learn Mathematics*. The National Academies Press.

Kynn, M. (2006). Missing data in winbugs. `psc.maths.lancs.ac.uk/.../missingdata/MissingBayesian.artic`

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

Muraki, E. (1992). A generalized partial credit model: application of an em algorithm. *Applied Psychological Measurement*, 16:159–176.

Neild, R. and Balfanz, R. (2006). An extreme degree of difficulty: the educational demographics of urban neighborhood high schools. *Journal for Education for Students Placed at Risk*, 11:123–141.

Neild, R., Stoner-Eby, S., and Furstenberg, F. (2008). Connecting entrance and departure. *Education and Urban Society*, 40(5):543–569.

Nomi, T. and Allensworth, E. (2009). 'double-dose' algebra as an alternative strategy to remediation: effects on students' academic outcomes. *Journal of Research on Educational Effectiveness*, 2:111–148.

Revelle, W. and Zinbarg, R. (2009). Coefficients alpha, beta, omega, and the glb: comments on sijtsma. *Psychometrika*, 74:145–154.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. New York: J.Wiley & Sons.

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

Schafer, J. and Olsen, M. K. (1998). Multiple imputation for multivariate missing data problems: A data analyst's perspective. *Multivariate Behavioral Research*, 33:545–571.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of cronbach's alpha. *Psychometrika*, 74:107–120.

Stavy, R. and Tirosh, D. (2000). *How students (mis-)understand science and mathematics.* Teacher's College Press.

van derk Ark, L. (2007). Mokken scale analysis in r. *Journal of Statistical Software*, 20.