Increased Math Achievement in Grade 3 and 6 Students Participating in JUMP Math's 2011-12 National Book Fund Program.

Beverley Murray, Ph.D.

(revised February 16, 2015)[1]

Executive Summary:

JUMP Math characterizes its approach to math instruction as *guided discovery*, a combination of direct instruction, discovery learning, and varied practice.[2] Complex math problems are taught by decomposing them into incremental steps and advocating mastery of simpler concepts before advancement to more complex concepts. Scaffolding of math problems is widely used to assist with independent learning. The program also promotes the importance of building student confidence and the notion that all students are capable of learning mathematics with appropriate supports.[3] Components of the program include professional development; *Teacher Resources* composed of lesson plans, quizzes/tests, and answer keys; *SMART Lesson Materials* for use with interactive white boards; and student *Assessment & Practice* books.

To evaluate the growth of students using the JUMP Math program, math achievement was assessed in both the fall and spring for students in non-blended grade 3 and 6 classrooms participating in JUMP's 2011-12 National Book Fund (NBF) program. A total of 326 students in eighteen classrooms completed the math computation subtest of the *Wide Range Achievement Test, Fourth Edition* (WRAT-4) in October 2011 and May 2012. Average student growth in math achievement was 1.8 times that of the WRAT-4 standardization sample, and mean standard scores in the spring (mean (SD) = 100.9 (11.8), N = 326) were significantly higher than mean standard scores in the fall (mean (SD) = 96.8 (10.9), N = 326, t (paired) = -7.96, $p < 0.001$). The corresponding percentile rank of students increased from the 42nd percentile in the fall to the 53rd percentile in the spring. The number of students scoring 'above average' or higher increased by 112% in the spring (72 students) compared to the fall (34 students). The number of students scoring 'below average' decreased by 25% in the spring (64 students) compared to the fall (85 students). We cannot know for certain whether the increased growth in math achievement relative to the WRAT-4 was due solely to the JUMP Math program because this study did not employ randomized control and treatment groups. By using a standardized test with alternate forms, however, we reduced the potential impact of several confounding variables making it likely that the JUMP Math program played a significant role.

[1] An earlier version of this report first appeared September 11, 2013.

[2] http://www.jumpmath.org/

[3] Mighton, J. (2004). The myth of ability: nurturing mathematical talent in every child. Toronto: House of Anansi Press.

Background:

Every year, JUMP Math's National Book Fund Program awards free JUMP Math resources to classrooms across Canada. This program is funded primarily through a grant from TD Bank, augmented by internally generated funding from JUMP Math. To be considered for the award, school principals and teachers must submit an application in which they describe their community and the needs of their students. Priority for awards is given to schools serving high-need communities where student achievement in mathematics is below national standards. In the 2011-12 school year, JUMP Math's National Book Fund Program awarded resources to 187 classrooms in 88 schools.

In order to assess the effectiveness of the JUMP Math program, 24 classrooms consisting of the non-blended grade 3 and grade 6 classes were selected for testing. The classrooms were located in British Columbia (11), Ontario (4), Quebec (2) and Alberta (1). Teachers were asked to administer the Math Computation subtest of the Wide Range Achievement Test, Fourth Edition (WRAT-4)[4] to their students in October 2011 and again in May 2012. Teachers were sent two alternate forms of the WRAT-4, designated the green form and the blue form, containing different questions but considered equally difficult. Detailed instructions on how to administer the test and return envelopes were provided to each teacher. In the fall, teachers were asked to administer the blue form to half of their students and the green form to the remaining half. For the spring testing, tests forms were pre-labelled with students' names to ensure that they received the alternate coloured form. Completed tests were sent back to JUMP Math and scored by a qualified teacher and the researcher. Standard scores were determined for each student in the spring and fall by looking up their raw test score in a conversion table that corresponds to the student's grade, test form (blue versus green), and time of testing (fall versus spring).

Results:

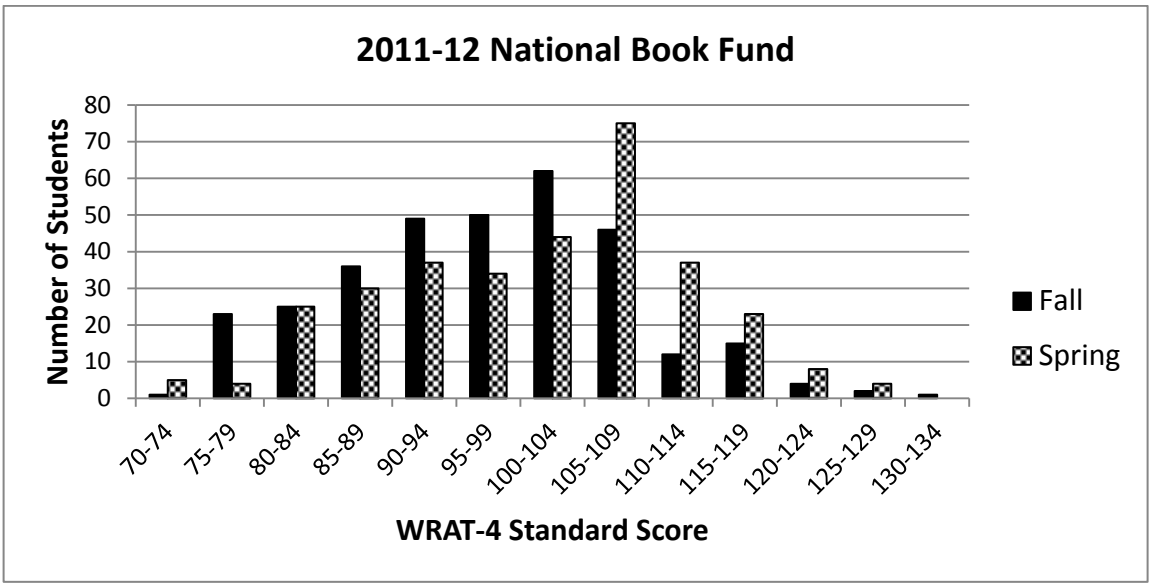
Teachers from 18 of the 24 classrooms selected for testing administered the WRAT-4 in both the fall of spring of the 2011-12 school year. Three hundred and sixty-nine students completed the WRAT-4 test in the fall (275 grade 3, 94 grade 6) and 349 students completed the test in the spring (259 grade 3, 90 grade 6). Standard scores were determined for the 326 students in 18 classes (14 grade 3 and 4 grade 6) who completed the tests in both the fall and spring of the 2011-12 school year. Combining all students, the mean standard score in the spring (mean (standard deviation (SD)) = 100.9 (11.8)) was significantly higher than the mean standard score in the fall (96.8 (10.9), t (paired) = -7.96, $p < 0.001$). The corresponding percentile rank of students (relative to the WRAT-4 standardization sample) increased from the 42nd percentile in the fall to the 53rd percentile in the spring. Using the

[4] Wilkinson G. & Robertson G. (2006). *Wide range achievement test (4th ed.).* Lutz, FL: Psychological Assessment Resources, Inc.

published standard deviation for the WRAT-4 (SD = 15), this increase in standard score corresponds to an effect size of 0.27 ((100.9 – 96.8)/15).

The frequency distribution of standard scores obtained in the fall and spring is shown below in Figure 1. This graph includes only those students (N=326) who wrote the test in *both* the fall and spring of the 2011-12 school year. The graph illustrates that the distribution of scores obtained in the spring is shifted to the right (towards higher scores) when compared to the distribution obtained in the fall. The interval containing the mode (most frequent score) increased from 100-104 in the fall to 105-109 in the spring. Although both distributions are skewed to the right (or "positively" skewed) as indicated by the longer right-hand tail, the distribution in the spring is more bell-shaped or symmetrical. The number of students scoring 'above average' or higher (a standard score of 110 or higher) increased by 112% in the spring (72 students) compared to the fall (34 students). The number of students scoring 'below average' or lower (a standard score of 89 or lower) decreased by 25% in the spring (64 students) compared to the fall (85 students).

Figure 1:



In the fall, 22 of the grade 3 students received a raw score below 5; in these cases an oral test must be administered in order to determine a standard score. Because teachers were not asked to administer oral tests to their students, standard scores could not be assessed for these students and they were excluded from the data analysis. The majority of students with raw scores less than 5 in the fall had raw scores greater than 5 in the spring (N = 17). We were therefore able to determine a standard score in the spring for 17 of the 22 students that were excluded in the fall. Scores for the other 5 students remained too low to determine whether they had made gains. In order to obtain a conservative estimate of the change in math achievement for the 17 students we calculated their maximum possible standard score in the fall by assuming they had achieved a perfect score on the

oral test (15/15). The average maximum possible standard score in the fall for these low-scoring students was 72.3 (SD = 3.1) whereas their average standard score in the spring was 84.2 (SD = 8.1). Thus, this group of students increased their rank relative to the WRAT-4 standardization sample from a maximum of the 3rd percentile in the fall to the 14th percentile in the spring.

We hypothesized that increases in student math achievement would strongly depend on classroom-level variables such as the teacher's fidelity to the JUMP Math program. Figure 2 shows the mean standard score (error bars denote the standard error of the mean (SEM)) in the fall and spring for each of the grade 3 classrooms; classrooms are ordered on the graph from the lowest (left) to highest mean standard score in the fall. The corresponding change in percentile rank for each of the grade 3 classrooms is shown in Figure 3. The change in mean standard score and percentile rank for each of the grade 6 classes is similarly shown in Figure 4 and Figure 5. These graphs comparing the data for each classroom demonstrate that there is considerable variability across classrooms with regards to growth in student achievement, particularly for the grade 3 classes. We suspect that the variability across classrooms reflects several factors including teacher fidelity to the JUMP Math program, previous experience with the JUMP Math program, and other variables related to student learning and teacher effectiveness. It will be the challenge of future studies to determine the classroom-level factors that predict student increases in math achievement.

Figure 2:

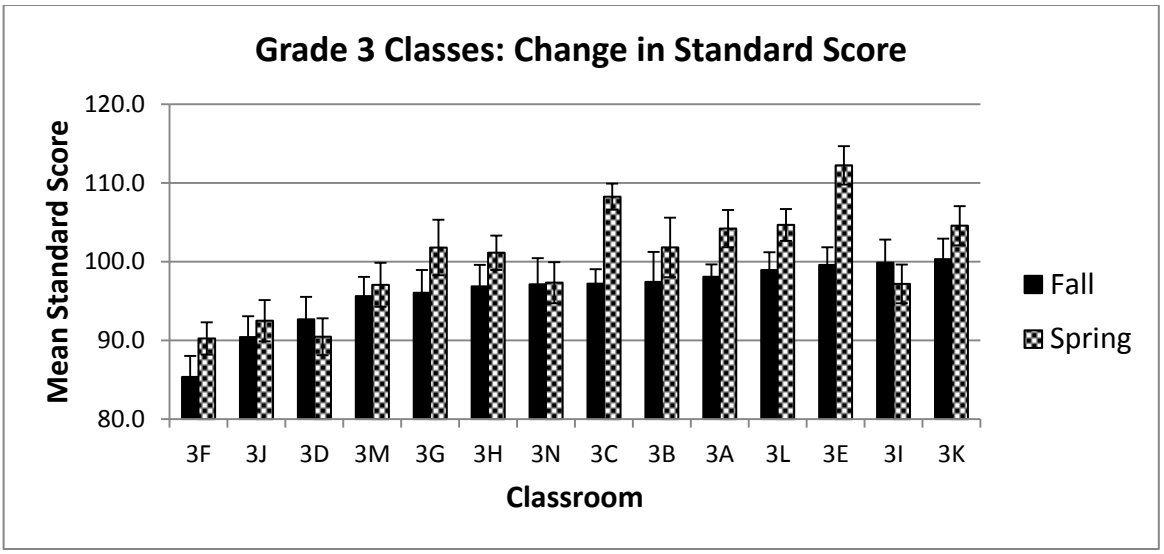


Figure 3:

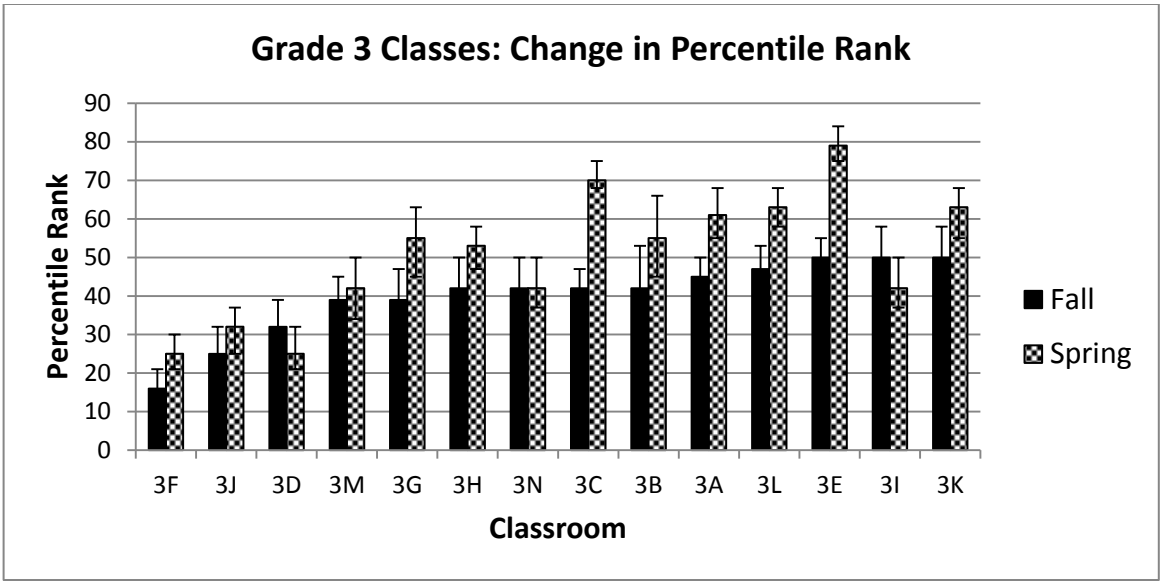


Figure 4:

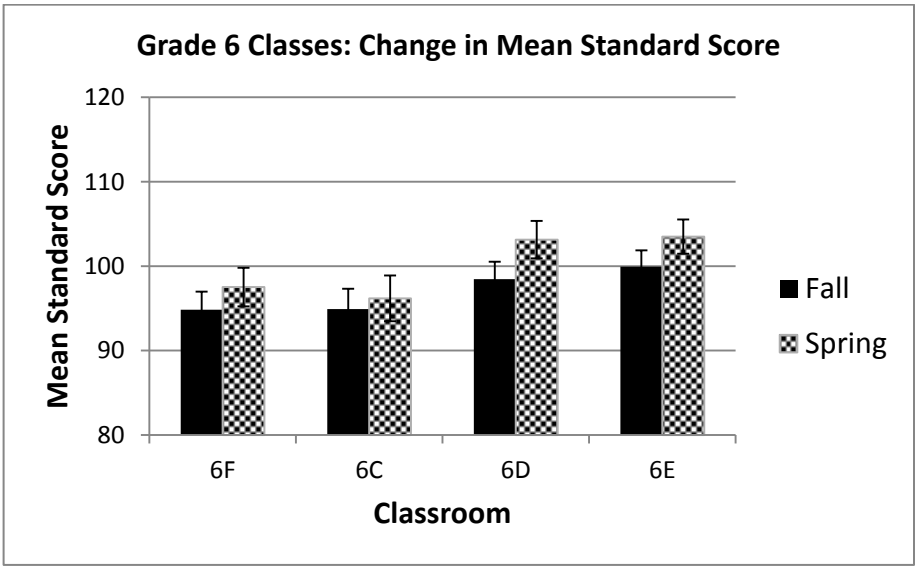
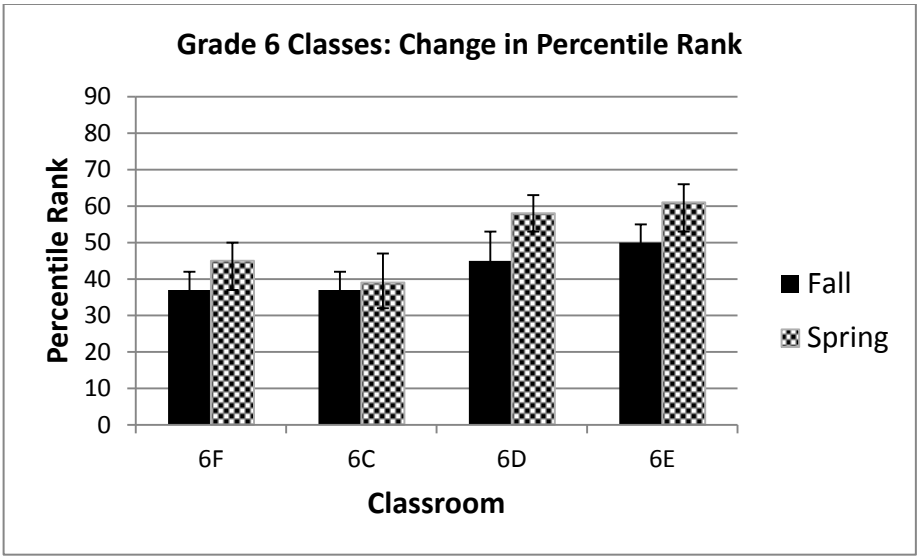


Figure 5:

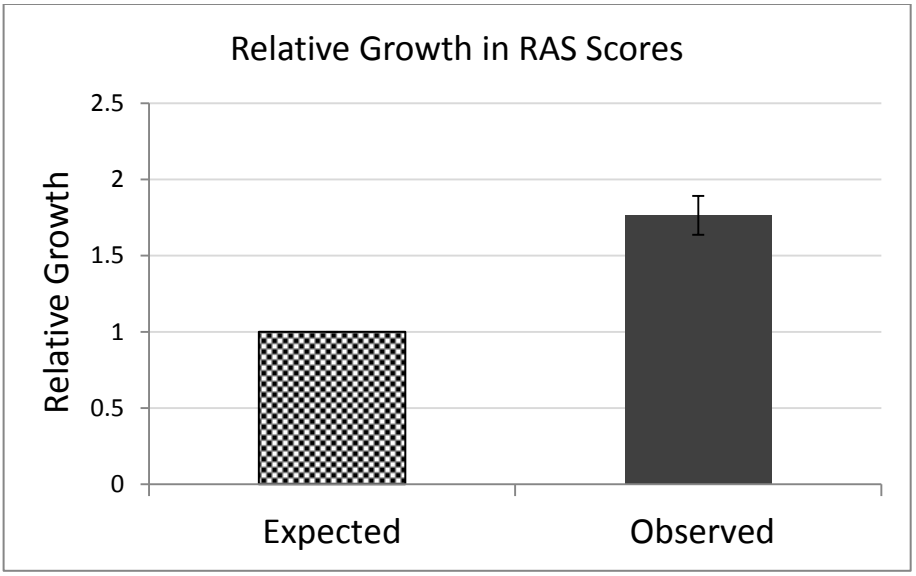


Each student's raw score on the WRAT-4 math test was also converted to a Rasch Ability Scaled (RAS) score using conversion tables for the blue and green forms of the WRAT-4. A student's RAS score will increase over time as their math achievement (raw score) increases; the RAS score is therefore well suited to measuring growth in student achievement from one time to another.[5] In contrast, standard scores will remain constant over time if the student grows at the same rate as the standardization sample. We defined *observed growth* as the difference between a student's RAS score in the spring and their RAS score in the fall (observed growth = RAS score in spring – RAS score in fall). *Expected growth* was also calculated for each student by subtracting their RAS score in the fall from their expected RAS score in the spring (expected growth = expected RAS score in spring – RAS score in fall). The expected RAS score in the spring was determined for each student by calculating the raw score in the spring that would result in the same standard score the student had obtained in the fall. We defined each student's *relative growth* in math achievement (relative to the WRAT-4 standardization sample) as the ratio of observed growth to expected growth (relative growth = observed growth/expected growth). Thus, a student with a relative growth score of 1 grew at the same rate as students from the WRAT-4 standardization sample with the same fall standard score. Growth in student math achievement was 1.5 (grade 3) and 2.5 (grade 6) times the growth of students in the WRAT-4 standardization sample. Across both grades, math achievement of students in the 2011-12 NBF grew at 1.8 times the rate of the WRAT-4 standardization sample (Figure 6).[6]

[5]RAS scores are derived such that the mean RAS score for the grade 5 normative group is equal to 500.

[6]In the original version of this report, dated September 11, 2013, we calculated expected growth by grade using the average achievement level for that grade, whereas in this update we calculate expected growth individually for each student. We believe the latter approach yields a more accurate estimate of student growth.

Figure 6:



In order to assess whether a student's grade, gender, initial math achievement, and a contextual classroom variable (mean SS.fall for student's class) could be used to predict their math achievement in the spring, a hierarchical linear mixed model was fit to the data using the lme (linear mixed effects) function in R, an open-source statistical package[7]. In contrast to linear models that include only fixed effects (e.g. the traditional ANOVA), mixed models (that include both fixed and random effects) are able to account for the effect of clustering of students in classrooms (i.e. students from the same classroom tend to be more alike than students from different classrooms). A student's grade (3 versus 6), initial math achievement (SS.fall) and the average SS.fall for their classroom (cvar(SS.fall, class) were significant predictors of their math achievement in the spring (SS.spring, see Table I). Gender, however, was not a significant predictor variable. The complete statistical analysis in R is provided in Appendix I.

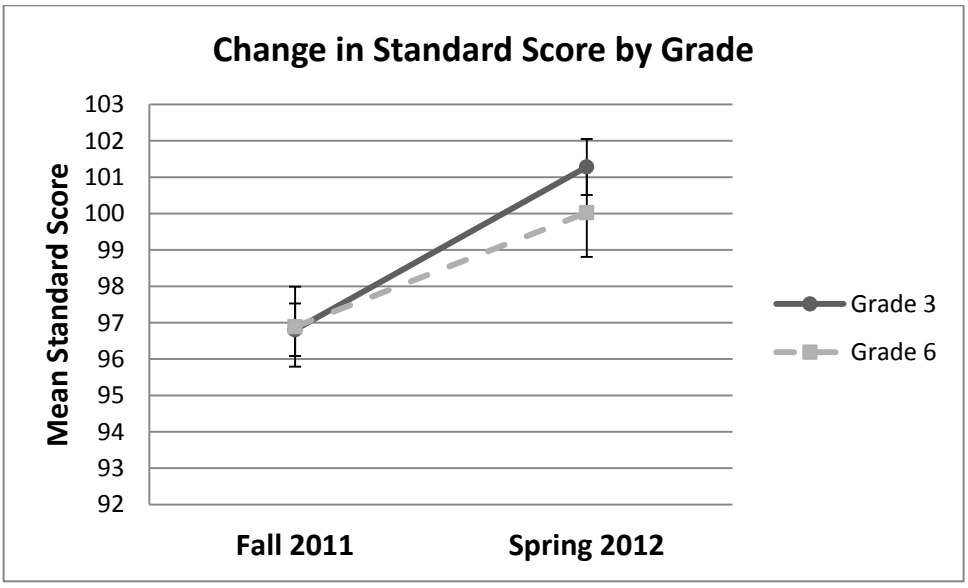
[7] R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Table I:

Predictor	Coefficient	Standard Error	Degrees of Freedom	t-value	p-value
SS.Fall	0.60	0.049	306	12.28	0.000
Grade	-26.55	9.52	15	-2.79	0.014
Mean Class SS.Fall (cvar(SS.fall, class)	0.62	0.27	15	2.26	0.039
Interaction SS.Fall: Grade	0.26	0.096	306	2.73	0.0068

The negative coefficient for the predictor "Grade" (see Table I) and the significant interaction between SS.fall and Grade ($p < 0.01$) reflects the fact that grade 6 students showed less growth than grade 3 students although their initial scores in the fall were similar. This finding is illustrated in Fig. 7 (see below) which shows the mean standard score in the fall and spring for grade 3 versus grade 6 students (+/- sem). The different slopes of the two lines indicate an interaction between SS.fall and Grade.

Figure 7:



Discussion:

Standard scores provide a convenient measure for comparing student achievement: a student with a standard score of 100 has achieved a score equal to the mean score of the sample of students used to standardize the test. Their percentile ranking is thus 50% because half of the students in the standardization sample scored above 100 and half scored below 100. Students composing the standardization sample in the case of the WRAT-4 were tested in both the fall and spring of the school year. Thus, a student that demonstrates the same growth rate as the standardization sample will achieve the same standard score in the fall and spring of the school year. Students participating in JUMP Math's 2011-12 National Book Fund Program showed significant increases in mean standard score in the spring (100.9) compared to the fall (96.8). The corresponding percentile rank of Book Fund students increased from the 42nd percentile in the fall to the 53rd percentile in the spring. The number of students scoring 'above average' or higher (a standard score of 110 or higher) increased by 112% in the spring (72 students) compared to the fall (34 students). The number of students scoring 'below average' or lower (a standard score of 89 or lower) decreased by 25% in the spring (64 students) compared to the fall (85 students). Finally, growth of 2011-12 National Book Fund students in math achievement was 1.8 times that of the WRAT-4 standardization sample.

Whereas the standard scores for the WRAT-4 standardization sample have a normal distribution (i.e. a symmetric, bell-shaped curve), the standard scores obtained in the present study have a skewed distribution, particularly for scores obtained in the fall. This is illustrated in Figure 1 where the distribution of scores in the fall is positively skewed (i.e. the right tail of the distribution is longer than the left tail). This right-skewed distribution could be due to the selection process for the National Book Fund Program. The classrooms that were selected for the program were (mostly) from high-need communities where math achievement was below national standards. In the spring, however, the distribution of scores shifted towards higher standard scores and more closely resembles a normal distribution.

The results of this study are consistent with the findings of a study led by researchers from the Hospital for Sick Children in Toronto and the Ontario Institute for Studies in Education, at the University of Toronto[8]. The study by Solomon et al. was a randomized control trial (RCT) in which classrooms were randomly assigned to either the treatment group (JUMP Math) or control group (incumbent math program). The RCT is considered the gold-standard in research design and permits causal inferences to be made regarding the effect of the treatment on the variable(s) being measured in the study. Using a more extensive battery of tests, Solomon et al. found that students in the JUMP

[8] Solomon, T., Martinussen, R., Dupuis, A., Gervan, S., Chaban, P., Tannock, R., Ferguson, B. (2011) Investigation of a Cognitive Science Based Approach to Mathematics Instruction, peer-reviewed data presented at the Society for Research in Child Development Biennial Meeting, Montreal, March 31 - April 2, 2011.

Math program progressed in their math achievement at approximately twice the rate of students in the control group.

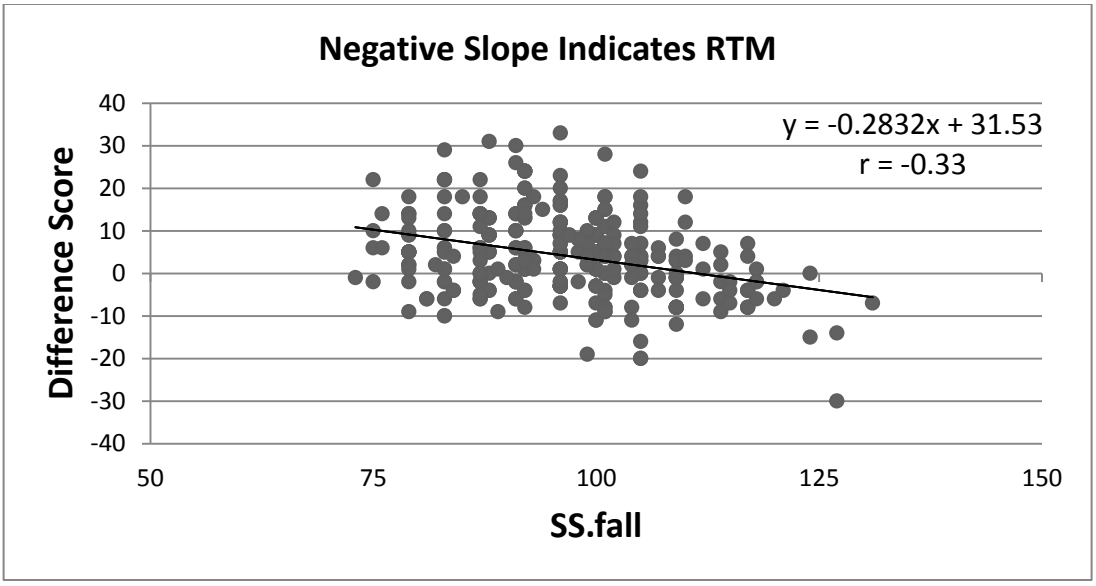
In contrast to the RCT design used by Solomon et al., the current study is an example of a single-group, pre- and post-test research design. This design is also referred to as "pre-experimental" because subjects have not been randomly assigned to treatment and control groups as in a true experimental design. The lack of randomized control and treatment groups in this study limits our ability to make causal inferences due to possible confounding factors. There are four well-recognized confounding factors unique to pre-experimental research studies: history, maturation, test effects, and regression-to-the-mean[9]. We have reviewed the potential impact of each of these factors and conclude that it is unlikely they can account for all of the gains in math achievement that were obtained in this study. In particular, our use of a standardized test with two alternate forms eliminates any possible practice effect that may occur when students complete the same test in the fall and spring of the same school year.

One potential confounding factor that requires careful consideration in studies employing a pre-experimental design is regression-to-the-mean (RTM). RTM is a statistical phenomenon whereby a distribution of measurements (e.g. test scores) narrows with repeated observations[10]. The effect is due to the greater measurement error in the tails of the distribution. Students with very low test scores will be more likely to have higher scores on a subsequent test. Similarly, students with very high test scores will be more likely to have lower scores on a subsequent test. The effects of RTM can lead education researchers to erroneously conclude that their treatment had a greater effect for low-achieving students. In order to determine whether RTM was evident in this data set, the difference score (SS.diff = SS.spring – SS.fall) for each student was plotted against their standard score in the fall (SS.fall). If RTM was present, students with a low score in the fall (low SS.fall) would have a higher difference score (SS.diff) than students with a high score in the fall (high SS.fall). Thus, we would expect SS.diff to be negatively correlated with SS.fall (i.e. a regression line through the points would have a negative slope). The scatter plot and regression line in Figure 8 below indicates that RTM was present in these data: the slope of the regression line is negative (-0.28).

[9] Emma Marsden & Carole J. Torgerson (2012): Single group, pre- and post-test research designs: Some methodological concerns, Oxford Review of Education, 38:5, 583-616.

[10] Adrian G Barnett, Jolieke C van der Pols & Annette J Dobson (2005): Regression to the mean: what it is and how to deal with it, International Journal of Epidemiology, 34:215–220 .

Figure 8:



RTM cannot account for all of the changes in standard score observed in this study. The histogram in Figure 1 shows that the shift to the right in the distribution of standard scores was most pronounced for standard scores in the centre of the distribution (between standard scores of 90 and 120). If RTM was entirely responsible for the increases in standard score, we would expect that the largest shifts would be observed for the lowest test scores. In addition, RTM alone would not produce a change in the mean standard score, but would decrease the spread of standard scores (i.e. decrease the standard deviation of the standard scores). However, the mean standard score in the spring was significantly higher than the mean standard score in the fall, whereas the standard deviation did not decrease. We therefore conclude that other factors must have contributed to the increases in standard score and the effect of RTM was most likely greatest at the tails of the distribution. The impact of RTM can be summarized as an overestimation of gains for low-scoring students and an underestimation of gains for high-scoring students.

Future Directions:

One of the main challenges for future research will be to delineate the classroom-level variables that predict gains in student math achievement. The magnitude of the increase in standard score had a wide range across classrooms, especially for the grade 3 classrooms. In some classrooms, students showed impressive gains whereas other classrooms showed little or no improvement in standard score. We hypothesise that earlier completion of JUMP Math professional development (before beginning use of the program) will increase fidelity to the JUMP Math program and result in enhanced and more consistent student achievement in mathematics across all classes. In the present study, less than half of the teachers had completed JUMP Math professional development by mid-October 2011 (only 8 of 18 teachers). Four teachers finished the 2011-12 school year without completing any JUMP Math professional development. The remaining 6 teachers received their professional development mid-way through the school year. The program requirements for the National Book Fund Program were altered in 2012-13 such that teachers are now required to complete a JUMP Math professional development program prior to the start the school year. We anticipate that this change will improve teacher fidelity to the program and further enhance gains in student math achievement.

Appendix I:

```
#Linear Mixed Model Analysis of BF 2011-12 Data in R
>#  Data file = BF
>
>
> head(BF)
    student gender   school    class   grade ver.fall ss.fall ver.spring
1 student20   Male School 4 Class 3A Grade 3    Green      88       Blue
2 student11   Male School 4 Class 3A Grade 3     Blue      91      Green
3  student6 Female School 4 Class 3A Grade 3     Blue      NA      Green
4  student2 Female School 4 Class 3A Grade 3    Green      83       Blue
5  student8 Female School 4 Class 3A Grade 3     Blue      96      Green
6 student22 Female School 4 Class 3A Grade 3    Green      88       Blue
  ss.spring
1        84
2        85
3        89
4        93
5        93
6        97
> fit0 <- lm(ss.spring ~ ss.fall, data = BF) #start with basic linear model
comparing student scores in spring (y) with student scores in fall (x)
> summary(fit0) #ss.spring = 31.5 + 0.7(ss.fall); Model accounts for 44% of
variance ss.spring

Call:
lm(formula = ss.spring ~ ss.fall, data = BF)

Residuals:
     Min       1Q   Median       3Q      Max
-25.5585  -6.2059  -0.4718   4.9400  28.6611

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 31.52987    4.36678    7.22 3.72e-12 ***
ss.fall      0.71676    0.04482   15.99  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.798 on 324 degrees of freedom
  (44 observations deleted due to missingness)
Multiple R-squared: 0.4412,    Adjusted R-squared: 0.4394
F-statistic: 255.8 on 1 and 324 DF,  p-value: < 2.2e-16

> anova(fit0) #ss.fall is a significant predictor of ss.spring
Analysis of Variance Table

Response: ss.spring
           Df Sum Sq Mean Sq F value    Pr(>F)
ss.fall     1  19800 19799.9  255.78 < 2.2e-16 ***
Residuals 324  25081    77.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> fit1 <- lm( ss.spring ~ ss.fall *gender, data = BF) #Does gender of student
influence ss.spring?
> summary (fit1)

Call:
lm(formula = ss.spring ~ ss.fall * gender, data = BF)

Residuals:
     Min       1Q   Median       3Q      Max
-25.7318  -6.1420  -0.6205   5.0449  28.8580
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        31.6136892  6.6355642   4.764 2.88e-06 ***
ss.fall             0.7174653  0.0682947  10.505  < 2e-16 ***
genderMale         -0.2921926  8.8450680  -0.033    0.974
ss.fall:genderMale -0.0005847  0.0907968  -0.006    0.995
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.834 on 321 degrees of freedom
  (45 observations deleted due to missingness)
Multiple R-squared: 0.4417,    Adjusted R-squared: 0.4365
F-statistic: 84.64 on 3 and 321 DF,  p-value: < 2.2e-16

> anova (fit0, fit1) #does not run because some students missing data for gender
Error in anova.lmlist(object, ...) :
  models were not all fitted to the same size of dataset
> BFMOD <- BF[!is.na(BF$gender),] #delete rows in dataset where gender = na
> fit0.mod <- lm(ss.spring ~ ss.fall, data = BFMOD) #fit basic linear model again
with modified dataset
> summary(fit0.mod)

Call:
lm(formula = ss.spring ~ ss.fall, data = BFMOD)

Residuals:
    Min      1Q  Median      3Q     Max
-25.547  -6.192  -0.457   4.959  28.675

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 31.50580    4.37187   7.206 4.08e-12 ***
ss.fall      0.71686    0.04487  15.978  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.808 on 323 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared: 0.4414,    Adjusted R-squared: 0.4397
F-statistic: 255.3 on 1 and 323 DF,  p-value: < 2.2e-16

> anova (fit0.mod,fit1) #adding gender (Level 1 variable) does not significantly
improve model
Analysis of Variance Table

Model 1: ss.spring ~ ss.fall
Model 2: ss.spring ~ ss.fall * gender
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    323 25059
2    321 25049  2    9.8642 0.0632 0.9388
>
> fit2 <- lm( ss.spring ~ ss.fall *grade, data = BF) #does grade (3 vs 6) of
student influence ss.spring?
> summary (fit2) #model now accounts for 45% of variance in ss.spring

Call:
lm(formula = ss.spring ~ ss.fall * grade, data = BF)

Residuals:
     Min       1Q   Median       3Q      Max
-24.2101  -5.6690  -0.5104   5.4770  28.2517

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          37.38260    5.01091   7.460  8.1e-13 ***
ss.fall               0.66006    0.05143  12.835  < 2e-16 ***
gradeGrade 6        -23.30096   10.00884  -2.328   0.0205 *
```

```
ss.fall:gradeGrade 6   0.22694    0.10272   2.209   0.0279 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.74 on 322 degrees of freedom
  (44 observations deleted due to missingness)
Multiple R-squared: 0.452,    Adjusted R-squared: 0.4469
F-statistic: 88.53 on 3 and 322 DF,  p-value: < 2.2e-16

> anova (fit0,fit2) #adding grade (Level 2 variable) significantly improves model
(p<0.05)
Analysis of Variance Table

Model 1: ss.spring ~ ss.fall
Model 2: ss.spring ~ ss.fall * grade
  Res.Df   RSS Df Sum of Sq     F  Pr(>F)
1    324 25081
2    322 24595  2    485.93 3.181 0.04285 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> fit3 <- lm( ss.spring ~ ss.fall *grade*gender, data = BF) #Try adding gender
back to model.  Maybe interacts with grade?
> summary (fit3) #model still acounts for 45% of variance in ss.spring (same as
before)

Call:
lm(formula = ss.spring ~ ss.fall * grade * gender, data = BF)

Residuals:
     Min       1Q   Median       3Q      Max
-25.0247  -5.8333  -0.4996   5.5004  28.2531

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     34.78770    7.62288   4.564  7.2e-06 ***
ss.fall                          0.68691    0.07869   8.730  < 2e-16 ***
gradeGrade 6                   -13.18002   15.28541  -0.862    0.389
genderMale                       4.53508   10.17438   0.446    0.656
ss.fall:gradeGrade 6             0.12794    0.15666   0.817    0.415
ss.fall:genderMale              -0.04707    0.10449  -0.450    0.653
gradeGrade 6:genderMale         -17.59547   20.32345  -0.866    0.387
ss.fall:gradeGrade 6:genderMale   0.16964    0.20863   0.813    0.417
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.79 on 317 degrees of freedom
  (45 observations deleted due to missingness)
Multiple R-squared: 0.454,    Adjusted R-squared: 0.442
F-statistic: 37.66 on 7 and 317 DF,  p-value: < 2.2e-16

> anova(fit2,fit3) #doesn't run because there are some students with gender = na
Error in anova.lmlist(object, ...) :
  models were not all fitted to the same size of dataset
> fit2.mod <- lm( ss.spring ~ ss.fall *grade, data = BFMOD) #use dataset without
missing gender
> summary (fit2.mod)

Call:
lm(formula = ss.spring ~ ss.fall * grade, data = BFMOD)

Residuals:
     Min       1Q   Median       3Q      Max
-24.1956  -5.6717  -0.5728   5.5514  28.2699

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)           37.35298    5.01722   7.445    9e-13 ***
ss.fall                0.66018    0.05149  12.822   <2e-16 ***
gradeGrade 6         -23.27134   10.02090  -2.322   0.0208 *
ss.fall:gradeGrade 6   0.22682    0.10284   2.206   0.0281 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.75 on 321 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared: 0.4522,    Adjusted R-squared: 0.4471
F-statistic: 88.33 on 3 and 321 DF,  p-value: < 2.2e-16

> fit3.mod <- lm( ss.spring ~ ss.fall *grade*gender, data = BFMOD)
> summary (fit3.mod)

Call:
lm(formula = ss.spring ~ ss.fall * grade * gender, data = BFMOD)

Residuals:
     Min       1Q   Median       3Q      Max
-25.0247  -5.8333  -0.4996   5.5004  28.2531

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    34.78770    7.62288   4.564  7.2e-06 ***
ss.fall                         0.68691    0.07869   8.730  < 2e-16 ***
gradeGrade 6                  -13.18002   15.28541  -0.862    0.389
genderMale                      4.53508   10.17438   0.446    0.656
ss.fall:gradeGrade 6            0.12794    0.15666   0.817    0.415
ss.fall:genderMale             -0.04707    0.10449  -0.450    0.653
gradeGrade 6:genderMale       -17.59547   20.32345  -0.866    0.387
ss.fall:gradeGrade 6:genderMale   0.16964    0.20863   0.813    0.417
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.79 on 317 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared: 0.454,     Adjusted R-squared: 0.442
F-statistic: 37.66 on 7 and 317 DF,  p-value: < 2.2e-16

> anova(fit2.mod,fit3.mod) #gender does not significantly improve model
Analysis of Variance Table

Model 1: ss.spring ~ ss.fall * grade
Model 2: ss.spring ~ ss.fall * grade * gender
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    321 24577
2    317 24494  4    82.137 0.2657 0.8999
> #Do not need to add gender to model.  Can go back to fit2
>
> fit4 <- lm( ss.spring ~ ss.fall *grade + cvar(ss.fall,class), data = BF)
> #fit4 adds Level 2 (derived) contextual variable to model (mean ss.fall for each
class)
> summary (fit4) #model now accounts for 48% of variance in ss.spring

Call:
lm(formula = ss.spring ~ ss.fall * grade + cvar(ss.fall, class),
    data = BF)

Residuals:
     Min       1Q   Median       3Q      Max
-23.7910  -5.7787   0.0244   5.2639  26.2433

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -29.1668    16.0279  -1.820   0.0697 .
ss.fall                0.6018     0.0518  11.616  < 2e-16 ***
```

```
gradeGrade 6          -25.2413     9.7505  -2.589   0.0101 *
cvar(ss.fall, class)    0.7484     0.1717   4.359 1.76e-05 ***
ss.fall:gradeGrade 6    0.2473     0.1001   2.471   0.0140 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.505 on 321 degrees of freedom
  (44 observations deleted due to missingness)
Multiple R-squared: 0.4826,    Adjusted R-squared: 0.4762
F-statistic: 74.86 on 4 and 321 DF,  p-value: < 2.2e-16

> anova (fit2,fit4) #contextual variable significantly improves model
Analysis of Variance Table

Model 1: ss.spring ~ ss.fall * grade
Model 2: ss.spring ~ ss.fall * grade + cvar(ss.fall, class)
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    322 24595
2    321 23220  1    1374.3 18.998 1.764e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> fit5 <- lm( ss.spring ~ ss.fall *grade*cvar(ss.fall,class), data = BF)
> #fit5 adds interactions between contextual variable and ss.fall and grade
> summary (fit5) #still 48% of variance (no improvement)

Call:
lm(formula = ss.spring ~ ss.fall * grade * cvar(ss.fall, class),
    data = BF)

Residuals:
     Min       1Q   Median       3Q      Max
-24.1000  -5.7199   0.1224   5.1562  26.5517

Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 3.318e+01  1.625e+02   0.204    0.838
ss.fall                                    -1.542e-01  1.767e+00  -0.087    0.930
gradeGrade 6                               -2.041e+00  3.788e+02  -0.005    0.996
cvar(ss.fall, class)                        1.091e-01  1.682e+00   0.065    0.948
ss.fall:gradeGrade 6                        3.568e-01  3.916e+00   0.091    0.927
ss.fall:cvar(ss.fall, class)                7.756e-03  1.825e-02   0.425    0.671
gradeGrade 6:cvar(ss.fall, class)          -2.625e-01  3.941e+00  -0.067    0.947
ss.fall:gradeGrade 6:cvar(ss.fall, class) -9.047e-04  4.068e-02  -0.022    0.982

Residual standard error: 8.535 on 318 degrees of freedom
  (44 observations deleted due to missingness)
Multiple R-squared: 0.4839,    Adjusted R-squared: 0.4725
F-statistic: 42.59 on 7 and 318 DF,  p-value: < 2.2e-16

> anova (fit4,fit5) #interactions with contextual variable do not significantly
improve model
Analysis of Variance Table

Model 1: ss.spring ~ ss.fall * grade + cvar(ss.fall, class)
Model 2: ss.spring ~ ss.fall * grade * cvar(ss.fall, class)
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    321 23220
2    318 23164  3    56.066 0.2566 0.8566
>
> fit6 <- lme(ss.spring ~ ss.fall *grade+cvar(ss.fall,class), BF,
+             random = ~ 1 |class,
+             na.action = "na.exclude")
> #fit 6 adds random effects to model (accounts for students clustered in
classrooms).  This is now a mixed model (both fixed and random effects).
> summary(fit6)
Linear mixed-effects model fit by REML
```

```
 Data: BF
       AIC      BIC    logLik
  2315.183 2341.583 -1150.592

Random effects:
 Formula: ~1 | class
        (Intercept) Residual
StdDev:    3.453057 7.949255

Fixed effects: ss.spring ~ ss.fall * grade + cvar(ss.fall, class)
                            Value Std.Error  DF   t-value p-value
(Intercept)           -17.011064 26.059570 306 -0.652776  0.5144
ss.fall                 0.600623  0.048918 306 12.278118  0.0000
gradeGrade 6          -26.551247  9.520132  15 -2.788958  0.0138
cvar(ss.fall, class)    0.619544  0.273805  15  2.262720  0.0389
ss.fall:gradeGrade 6    0.260529  0.095538 306  2.726961  0.0068
 Correlation:
                      (Intr) ss.fll grdGr6 c(.,c)
ss.fall               -0.017
gradeGrade 6          -0.019  0.494
cvar(ss.fall, class) -0.983 -0.162 -0.074
ss.fall:gradeGrade 6  0.034 -0.508 -0.973  0.057

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.64026580 -0.62291327 -0.03423169  0.61248189  3.10129620

Number of Observations: 326
Number of Groups: 18
> getVarCov(fit6) #Random effects account for 12% of variance in ss.spring.
Variance due to Level 2 (class).
Random effects variance covariance matrix
            (Intercept)
(Intercept)      11.924
  Standard Deviations: 3.4531
> anova(fit6, fit4) #Now know you must list lme fit BEFORE lm fit!
     Model df      AIC      BIC    logLik   Test  L.Ratio p-value
fit6     1  7 2315.183 2341.583 -1150.592
fit4     2  6 2334.485 2357.114 -1161.243 1 vs 2 21.30181  <.0001
> #fit 6 (with random intercept) signifcantly improves model (compared to fit4
without random intercept)
>
> fit7 <- lme(ss.spring ~ ss.fall *grade + cvar(ss.fall,class), BF, random = ~ 1 +
ss.fall| class,
+             na.action = "na.exclude")
> #fit 7 adds random slopes to model
> summary (fit7)
Linear mixed-effects model fit by REML
 Data: BF
       AIC      BIC    logLik
  2317.911 2351.854 -1149.955

Random effects:
 Formula: ~1 + ss.fall | class
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev     Corr
(Intercept) 9.95224950 (Intr)
ss.fall     0.06953756 -0.986
Residual    7.91704322

Fixed effects: ss.spring ~ ss.fall * grade + cvar(ss.fall, class)
                            Value Std.Error  DF   t-value p-value
(Intercept)           -19.043761 27.206539 306 -0.699970  0.4845
ss.fall                 0.592860  0.052471 306 11.298917  0.0000
gradeGrade 6          -27.630928 10.960819  15 -2.520882  0.0235
cvar(ss.fall, class)    0.648387  0.286953  15  2.259555  0.0392
ss.fall:gradeGrade 6    0.271546  0.103885 306  2.613919  0.0094
```

```
 Correlation:
                     (Intr) ss.fll grdGr6 c(.,c)
ss.fall               0.007
gradeGrade 6         -0.001  0.490
cvar(ss.fall, class) -0.980 -0.201 -0.097
ss.fall:gradeGrade 6  0.007 -0.503 -0.981  0.090

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.6566125 -0.6148083 -0.0199539  0.6494606  3.1607491

Number of Observations: 326
Number of Groups: 18
> getVarCov(fit7)
Random effects variance covariance matrix
            (Intercept)    ss.fall
(Intercept)    99.04700 -0.6825800
ss.fall        -0.68258  0.0048355
  Standard Deviations: 9.9522 0.069538
> anova(fit6, fit7) #fit 7 does not significantly improve model
     Model df      AIC      BIC    logLik   Test  L.Ratio p-value
fit6     1  7 2315.183 2341.583 -1150.592
fit7     2  9 2317.911 2351.854 -1149.955 1 vs 2 1.272369  0.5293
>
> #Therefore, we will use fit6 as the best model for these data
```