Increased Math Achievement in Elementary Students Participating in JUMP Math's 2012-13 National Book Fund Program.

Beverley Murray, Ph.D.

February 3, 2015

Executive Summary

JUMP Math characterizes its approach to math instruction as *guided discovery*, a combination of direct instruction, discovery learning, and varied practice.[1] 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.[2] 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 grade 4 to 7 students who participated in JUMP Math's 2012-13 National Book Fund (NBF) program. A total of 286 students in fourteen classrooms completed the math computation subtest of the *Wide Range Achievement Test, Fourth Edition* (WRAT-4) in October 2012 and May 2013. Average student growth in math achievement was 2.8 times that of the WRAT-4 standardization sample, and mean standard scores in the spring ($M = 94.5$, $SD = 12.2$) were significantly higher than mean standard scores in the fall ($M = 90.7$, $SD = 12.7$), paired $t(285) = -7.3$, $p < 0.001$. The corresponding percentile rank of students increased from the 27th percentile in the fall to the 37th percentile in the spring. The number of students scoring 'above average' or higher increased by 70% in the spring (34 students) compared to the fall (20 students). The number of students scoring 'below average' decreased by 29% in the spring (105 students) compared to the fall (135 students). A policy change implemented in the 2012-13 NBF required teachers to complete JUMP Math professional development prior to receiving JUMP Math resources. This may account for the fact that students in the 2012-13 NBF had higher growth rates than students in the 2011-12 NBF (2.8 versus 1.8 times the growth rate of the WRAT-4 standardization sample). 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] http://www.jumpmath.org/

[2] 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 2012-13 school year, JUMP Math's National Book Fund Program awarded resources to 3,600 students in 140 classrooms across 5 Canadian provinces (AB, BC, MB, ON, and QC).

In order to assess the growth in math achievement for students participating in the NBF program, 16 non-blended classrooms spanning grades 4 to 7 were selected for testing. Teachers were asked to administer the math computation subtest of the *Wide Range Achievement Test, Fourth Edition* (WRAT-4)[3] to their students in October 2012 and again in May 2013. Teachers were sent two alternate forms of the WRAT-4, designated the green form and the blue form, consisting of 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 14 of the 16 classrooms selected for testing administered the WRAT-4 in both the fall and spring of the 2012-13 school year. Standard scores were determined for the 286 students who completed the tests in both the fall and spring of the 2012-13 school year. The mean standard score in the spring spring (M[4] = 94.5, SD[5] = 12.2) was significantly higher than the mean standard score in the fall (M = 90.7, SD = 12.7), t(285) = -7.3 , $p < 0.001$.). We would expect students in the 2012-13 NBF to have the same standard score in the fall and spring if their math achievement had increased at the same rate as the WRAT-4 standardization sample. The fact that their mean standard score was significantly higher in the spring indicates that their math achievement grew at a higher rate than the

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

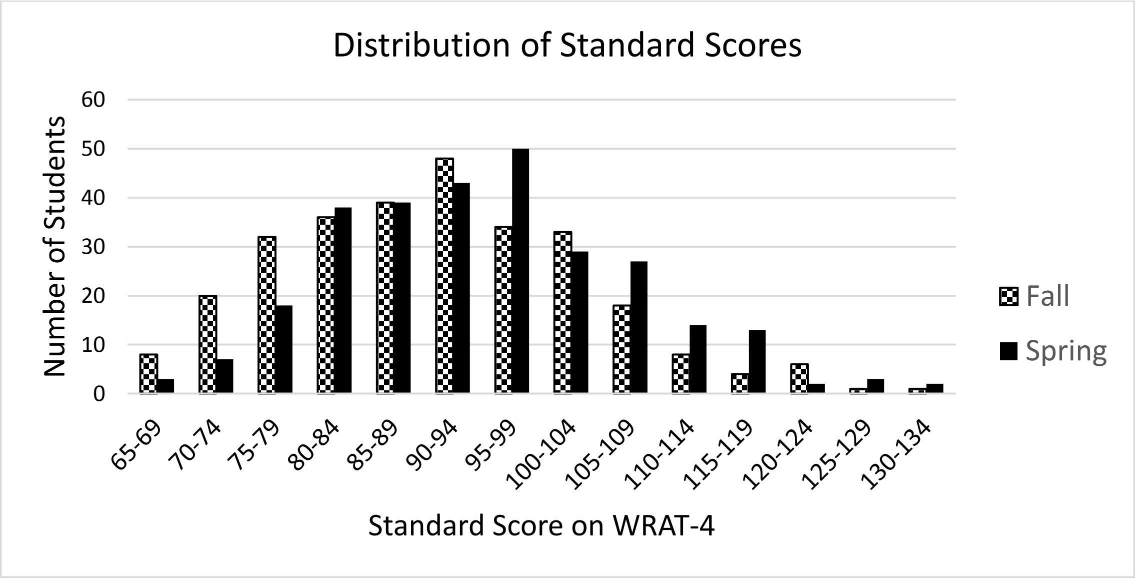
[4] M= mean

[5] SD = standard deviation

WRAT-4 standardization sample. The corresponding percentile rank of students (relative to the WRAT-4 standardization sample) increased from the 27th percentile in the fall to the 37th percentile in the spring. Using the published standard deviation for the WRAT-4 (SD = 15), this increase in standard score corresponds to an effect size of 0.25 ((94.5 – 90.7)/15).

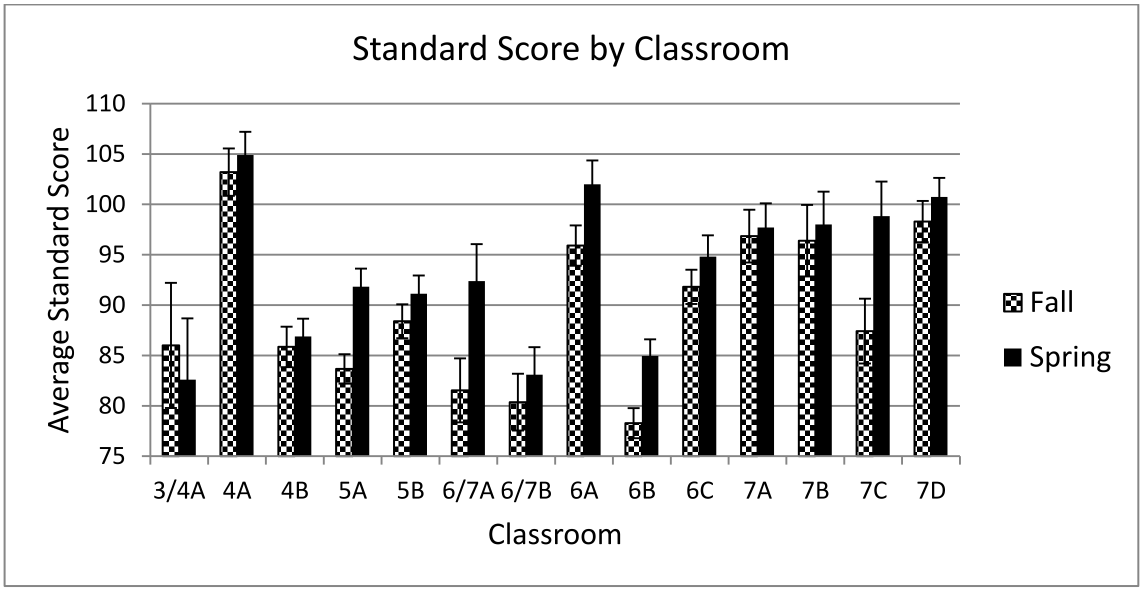
The frequency distributions of standard scores obtained in the fall and spring are shown below in Figure 1. The distributions include only those students (N=286) who completed either a blue or green test in the fall and then completed the alternate coloured test in the spring (students who completed the same test in both the fall and spring were excluded from the analysis). The graph illustrates that the distribution of scores obtained in the spring is shifted to the right (towards higher scores) compared to the distribution obtained in the fall. The median score increased from 90 in the fall to 93 in the spring. The number of students scoring 'above average' (equivalent to a standard score of 110 or higher) increased by 70% in the spring (34 students) compared to the fall (20 students) whereas the number of students scoring 'below average' (equivalent to a standard score of 89 or lower) decreased by 29% in the spring (135 students) compared to the fall (105 students).

Figure 1:



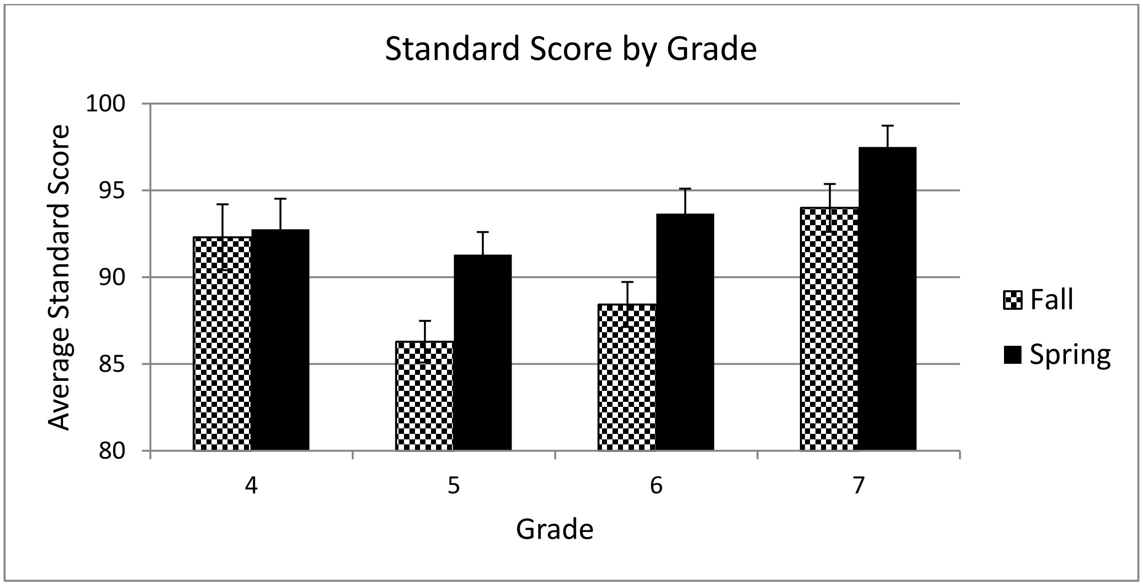
Mean standard scores in the fall and spring for each of the fourteen classrooms are shown in Figure 2 (error bars for all graphs denote the standard error of the mean (SEM)); classrooms are ordered on the graph according to grade level. Although non-blended (i.e. single-grade classrooms) were selected for testing, in three cases the classroom had been re-designated as blended by the time of testing. In one case, the classroom included grade 3 students (classroom 3/4A) who were excluded from the analysis. Figure 2 illustrates the variability across classrooms with respect to both the initial level of student achievement and the change in student achievement over the school year. We hypothesize that average increases in student math achievement depend in part on classroom-level variables such as the teacher's fidelity to the JUMP Math program.

Figure 2:



The average standard score was calculated for each grade in both the fall and spring (Figure 3). Increases in standard score ranged from 6% (grades 5 and 6) to 0.09% (grade 4); grade 7 classes had an average increase in standard score of 4%.

Figure 3:



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. 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. The average expected and observed RASS difference scores were calculated for each grade (Figure 4).

We defined each student's *relative growth* in math achievement (relative to the WRAT-4 standardization sample) 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. On average, math achievement of students in the 2012-13 NBF grew at 2.8 times the rate of the WRAT-4 standardization sample (Figure 5).

Figure 4:

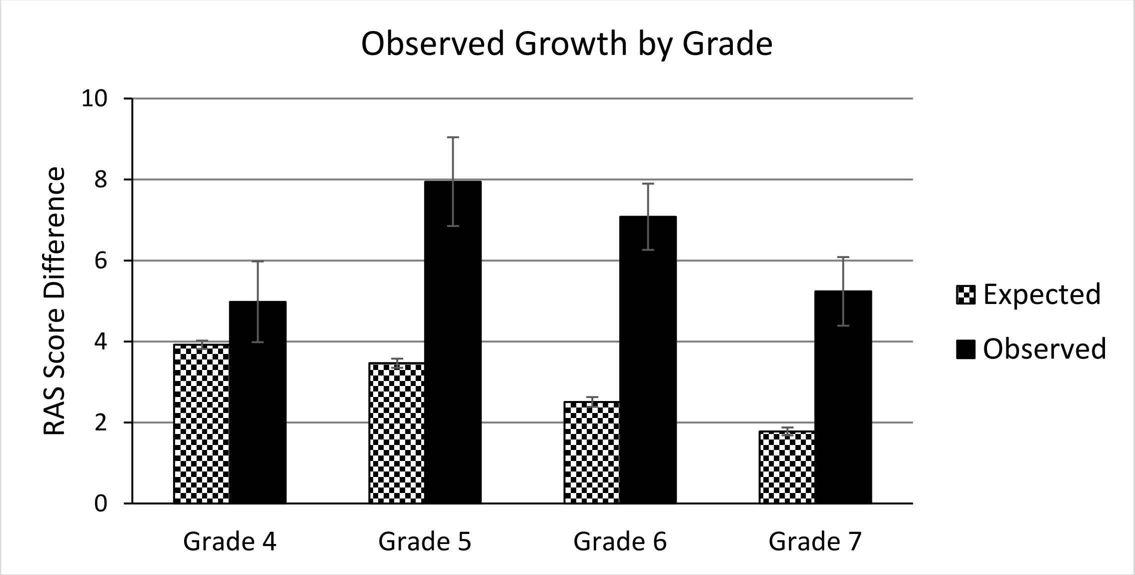
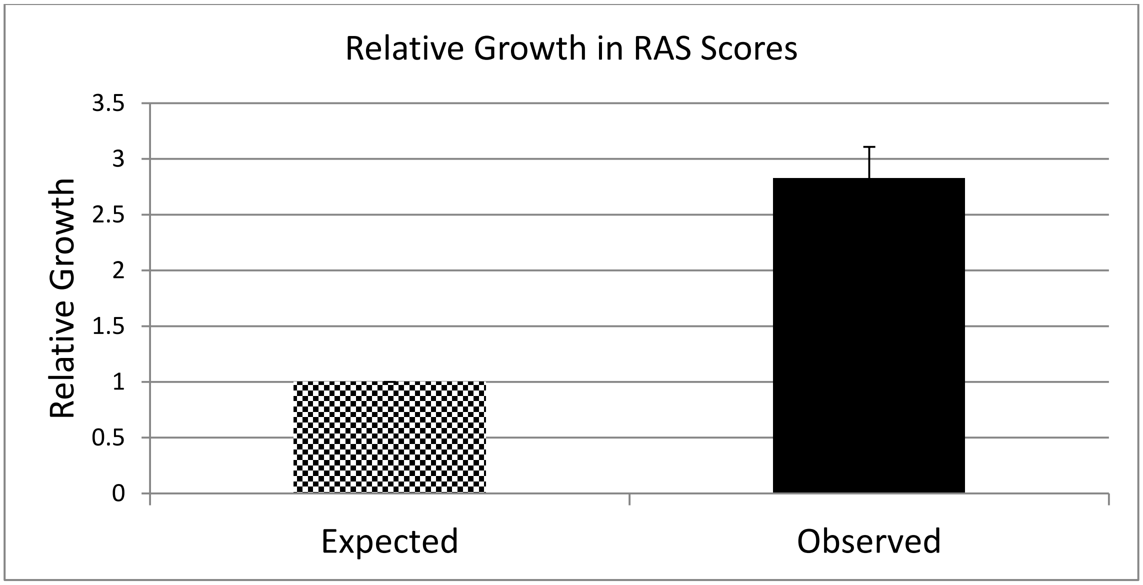


Figure 5:



In order to assess whether a student's initial math achievement, grade, gender, and a classroom-level variable (mean standard score for each class) could be used to predict student growth on the math computation subtest of the WRAT-4, a step-wise linear regression was performed on the data using the lm (linear model) function in R, an open-source statistical package[6]. Initial math achievement (SS.fall) and gender were significant predictors of student growth whereas grade level and the mean classroom standard score were not significant predictors of student growth (see Appendix I.A). A hierarchical linear mixed model was subsequently fit to the data using the lme (linear mixed effects) function in R; SS.fall and gender were included as fixed variables and classroom was included as a random variable (see Appendix I.B). Mixed models that include both fixed and random effects are able to account for the effect of clustering of students in classrooms. Students from the same classroom tend to be more alike than students from different classrooms as reflected by the intraclass correlation (ICC). The ICC of 0.10 for these data was calculated by dividing the component of the variance due to the random variable 'classroom' by the total variance (see Appendix I.C).

[6] 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: Linear Mixed Effects Model

Predictor	Coefficient	Standard Error	Degrees of Freedom	t-value	p-value
Intercept	28.85	3.45	267	8.36	0.000
SS.Fall	-0.26	0.04	267	-7.00	0.000
Gender	1.95	0.84	267	2.33	0.02

The negative coefficient for the predictor "SS.Fall" (see Table I) reflects the fact that students with low standard scores in the fall tended to show more growth than students with high standard scores in the fall. This negative correlation is demonstrated in Figure 6; each point on the graph represents a student's growth on the WRAT-4 versus their standard score in the fall. The regression line through these points has a slope of -0.27 and a correlation coefficient (R^2) of 0.19. The negative slope of the regression line is evidence of a statistical phenomenon known as regression-to-the-mean (RTM) and is unlikely to reflect any selective effect of the JUMP program on low-achieving students. The implications of RTM will be discussed below (see page 11).

Gender was also a significant predictor of growth (p = 0.02; Table I). The average RAS difference score for male students was 2.22 points higher than the average RAS difference score for female students (male: M=7.44, SD=8.56; female: M=5.22, SD=7.34). The average RAS score for male and female students in the fall and spring is shown in Figure 7. Female students started the school year with a slightly higher average RAS score compared to male students; male students ended the school year with a slightly higher average RAS score compared to female students.

Figure 6:

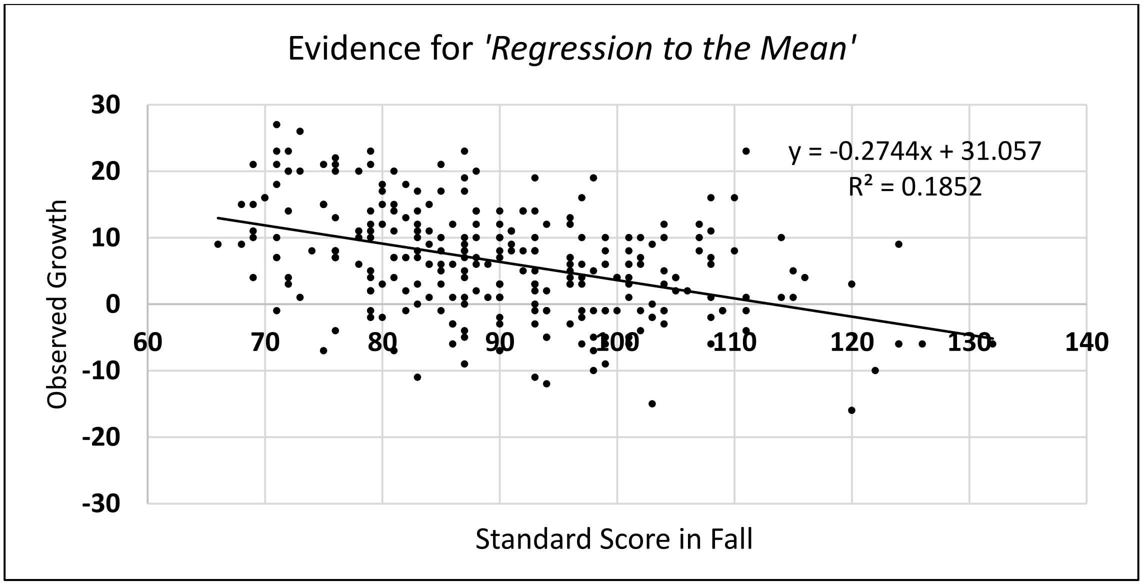
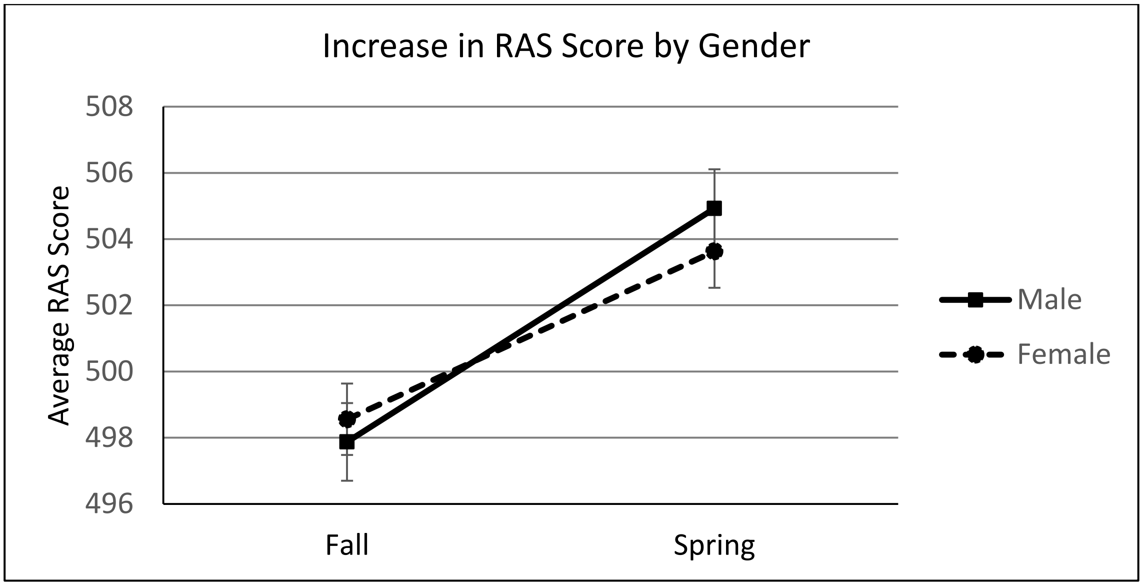


Figure 7:



Discussion

Standard scores are a useful measure for comparing student achievement on a standardized test: 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 corresponding percentile rank is 50%; half of the students in the standardization sample scored above 100 and half scored below 100. Students who composed the WRAT-4 standardization sample 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 2012-13 National Book Fund Program showed significant increases in mean standard score in the spring (M = 94.5, SD = 12.2) compared to the fall (M = 90.7, SD = 12.7). The corresponding percentile rank of NBF students increased from the 27th percentile in the fall to the 37th percentile in the spring. The number of students scoring 'above average' (equivalent to a standard score of 110 or higher) increased by 70% in the spring (34 students) compared to the fall (20 students) whereas the number of students scoring 'below average' (equivalent to a standard score of 89 or lower) decreased by 29% in the spring (135 students) compared to the fall (105 students). On average, students participating in the 2012-13 National Book Fund grew at 2.8 times the rate of the WRAT-4 standardization sample.

Whereas the distribution of standard scores for the WRAT-4 standardization sample has a normal distribution (i.e. symmetric and bell-shaped curve) centered on a standard score of 100, the distribution of standard scores obtained in the present study is positively skewed, particularly in the fall (Figure 1). A positive skew occurs when the right tail of the distribution is longer than the left tail. Our skewed distribution could be due to the selection process for the National Book Fund Program. Classrooms that were selected for the program were (mostly) from high-need communities where math achievement was below national standards. In the spring the distribution of scores shifted towards higher standard scores and was less skewed.

JUMP Math has used the WRAT-4 to assess student math achievement since 2011-12[7]. A comparison of the test results for the past 2 years is shown in Table II. We observed an increase in student growth between the 2011-12 and 2012-13 NBF (1.8 vs 2.8, respectively). This increase could be due to policy changes that were implemented in 2012-13, the most significant of which was requiring all participating teachers to complete a JUMP Math professional development session prior to the start of the school year. In contrast, less than half of the teachers selected for classroom testing in the 2011-12 NBF had completed JUMP Math professional development by mid-October 2011 (only 8 of 18 teachers). In addition, changes were made in the process for assessing NBF applications which may have improved our ability to identify high-needs classrooms. This may account for the increase in the percentage of low-scoring students participating in the 2012-13 NBF. Either or both of these policy changes could have impacted growth in student math achievement. Another difference between the two studies was the finding this year that gender was a statistically significant predictor of student growth; gender was not statistically significant in the 2011-12 NBF.

[7] Murray, B. (2013, September 18). *Increased math achievement in grade 3 and 6 students participating in JUMP Math's 2011-12 National Book Fund Program.* Retrieved from http://www.jumpmath.org/cms/sites/default/files/Student%20Achievement%20From%202011-12%20JUMP%20Math%20Book%20Fund%20%282013%29.pdf

Table II: NBF Student Test Results

	2012-13 NBF	2011-12 NBF
# of students tested in both fall & spring	286	326
Grades tested	4 to 7	3 and 6
SS fall vs SS spring	90.8 vs 94.6	96.8 vs 100.9
Percentile rank fall vs spring	27th vs 37th	42nd vs 53rd
Average student growth relative to WRAT-4	2.8	1.8
% of students scoring 'above average' in fall vs spring	7% vs 12%	10% vs 22%
% of students scoring 'below average' in fall vs spring	47% vs 36%	26% vs 20%

The results presented here 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]. Solomon et al. (2011) employed 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[9]. Using a more extensive battery of tests, Solomon et al. found that students in the JUMP 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[10]. We have reviewed the potential impact of each of these factors and

[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.

[9] Murnane, R.J. & Willett, J.B. (2011). *Methods matter: improving causal inference in educational and social science research*. New York, NY: Oxford University Press.

[10] 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.

conclude that it is unlikely they can account for the statistically significant gains in math achievement obtained for both the 2011-12 NBF and 2012-13 NBF. 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[11]. 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, observed growth for each student (RAS score spring – RAS score.fall) was plotted against their standard score in the fall (Figure 6). If RTM was present, students with a low score in the fall would tend to have a higher score in the spring (and therefore greater growth) whereas students with a high score in the fall would tend to have a lower score in the spring (and therefore lower growth). Thus, we would expect growth to be negatively correlated with standard score in the fall (i.e. a regression line through the points would have a negative slope). The scatter plot and regression line in Figure 6 indicate that RTM was present in these data: the slope of the regression line is negative (-0.27).

RTM cannot account for all of the changes in standard score observed in this study as RTM alone would not produce an overall change in the mean standard score that was obtained in this study. In addition, the histogram in Figure 1 shows that the shift to the right in the distribution of standard scores was evident for standard scores in the centre of the distribution (between standard scores of 95 to 99). If RTM was entirely responsible for the increases in standard score, we would expect that the shifts in standard score would be observed for the lowest and highest test scores. 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. 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.

Teacher Fidelity

The program requirements for the NBF were altered in 2012-13 such that teachers were required to complete a JUMP Math professional development program prior to the start of the school year. We anticipated that this change would improve teacher fidelity to the program and enhance gains in

[11] 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 .

student math achievement compared to the 2011-12 NBF. Student growth did improve in the 2012-13 NBF compared to the 2011-12 NBF, however it is not clear whether the improvement is due to increased fidelity to the JUMP Math program or other differences between the two samples of teachers and/or students. NBF teachers provide self-reported estimates of their fidelity to the JUMP Math program in a teacher feedback survey administered in the spring. Ten of the 14 teachers who administered the 2012-13 student assessments also completed the teacher feedback survey. Teachers were asked to rate how frequently they used the lesson plans and student books on a 7-point scale (1 = never; 7 = every class). On average, teachers reported using the lesson plans for approximately 50% of classes (M=4.4, SD=1.3) and the student books for slightly more than 80% of classes (M=6.0, SD=1.8). Ten of the 18 teachers who administered the 2011-12 student assessments also completed a teacher feedback survey and reported similar usage of the program (lesson plans: M=4.8, SD=1.1; student books: M=5.9, SD=1.5). Thus, we have no evidence that the frequency with which teachers used JUMP Math increased in the 2012-13 NBF (although it is still possible). Fidelity to the JUMP Math program includes not only the frequency of use but also how effectively the teacher delivers the program. Thus, teachers may not have increased the frequency of their use of the program but may have more effectively delivered the program due to the increased professional development provided in the 2012-13 NBF.

Appendix I

A:

```
fit <- lm(growth ~ ss.fall + gender + grade + class.avg.fall, data = BF)
> step <- stepAIC(fit, direction="both")
Start:  AIC=1131.9
growth ~ ss.fall + gender + grade + class.avg.fall

                 Df Sum of Sq   RSS    AIC
- grade           3     68.96 14769 1127.2
- class.avg.fall  1      2.49 14703 1130.0
<none>                        14700 1131.9
- gender          1    275.58 14976 1135.2
- ss.fall         1   2031.89 16732 1166.5

Step:  AIC=1127.22
growth ~ ss.fall + gender + class.avg.fall

                 Df Sum of Sq   RSS    AIC
- class.avg.fall  1      0.90 14770 1125.2
<none>                        14769 1127.2
- gender          1    274.23 15043 1130.4
+ grade           3     68.96 14700 1131.9
- ss.fall         1   2044.24 16813 1161.9

Step:  AIC=1125.24
growth ~ ss.fall + gender

                 Df Sum of Sq   RSS    AIC
<none>                        14770 1125.2
+ class.avg.fall  1      0.90 14769 1127.2
- gender          1    273.77 15044 1128.4
+ grade           3     67.37 14703 1130.0
- ss.fall         1   3012.91 17783 1175.8
> step$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
growth ~ ss.fall + gender + grade + class.avg.fall

Final Model:
growth ~ ss.fall + gender


              Step Df   Deviance Resid. Df Resid. Dev      AIC
1                                      276   14700.21 1131.898
2          - grade  3 68.9570192       279   14769.16 1127.222
3 - class.avg.fall  1  0.8993103       280   14770.06 1125.240
```

B.

```
> fit2 <-lme(growth ~ ss.fall + gender, BF, random = ~ 1|class, na.action = "na.ex
clude")

> summary(fit2)
Linear mixed-effects model fit by REML
 Data: BF
       AIC      BIC    logLik
```

```
  1927.036 1945.21 -958.5182

Random effects:
 Formula: ~1 | class
        (Intercept) Residual
StdDev:    2.331228 6.970438

Fixed effects: growth ~ ss.fall + gender
                Value Std.Error  DF   t-value p-value
(Intercept) 28.853715  3.453417 267  8.355121  0.0000
ss.fall     -0.257727  0.036796 267 -7.004199  0.0000
genderM      1.954270  0.837109 267  2.334548  0.0203
 Correlation:
        (Intr) ss.fll
ss.fall -0.968
genderM -0.183  0.069

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.54709364 -0.61286899 -0.01510766  0.63063140  2.72434283

Number of Observations: 283
Number of Groups: 14
```

C.

```
> VarCorr(fit2)
class = pdLogChol(1)
            Variance  StdDev
(Intercept)  5.434626 2.331228
Residual    48.587003 6.970438
> 5.434626/(5.434626 + 48.587003) #/ICC calculation
[1] 0.1006009
```