Increased Math Achievement in Elementary Students Participating in JUMP Math's 2013-14 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 students who participated in JUMP's 2013-14 National Book Fund (NBF) program. A total of 241 students in twelve classrooms completed the math computation subtest of the *Wide Range Achievement Test, Fourth Edition* (WRAT-4) in October 2013 and May 2014. Average student growth in math achievement was 2.5 times that of the WRAT-4 standardization sample, and mean standard scores in the spring (M = 95.3, SD = 11.0) were significantly higher than mean standard scores in the fall (M = 89.6, SD = 11.6), paired t(240) = 9.7, $p < 0.001$. The corresponding percentile rank of students increased from the 25th percentile in the fall to the 37th percentile in the spring. The number of students scoring 'above average' or higher increased by 92% in the spring (25 students) compared to the fall (13 students). The number of students scoring 'below average' decreased by 35% in the spring (85 students) compared to the fall (130 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] 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 Group, 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 2013-14 school year, JUMP Math's National Book Fund Program awarded resources to 3,380 students in 133 classrooms across 8 Canadian provinces (AB, BC, MB, NB, NS, ON, QC, and SK) and one Canadian territory (YT).

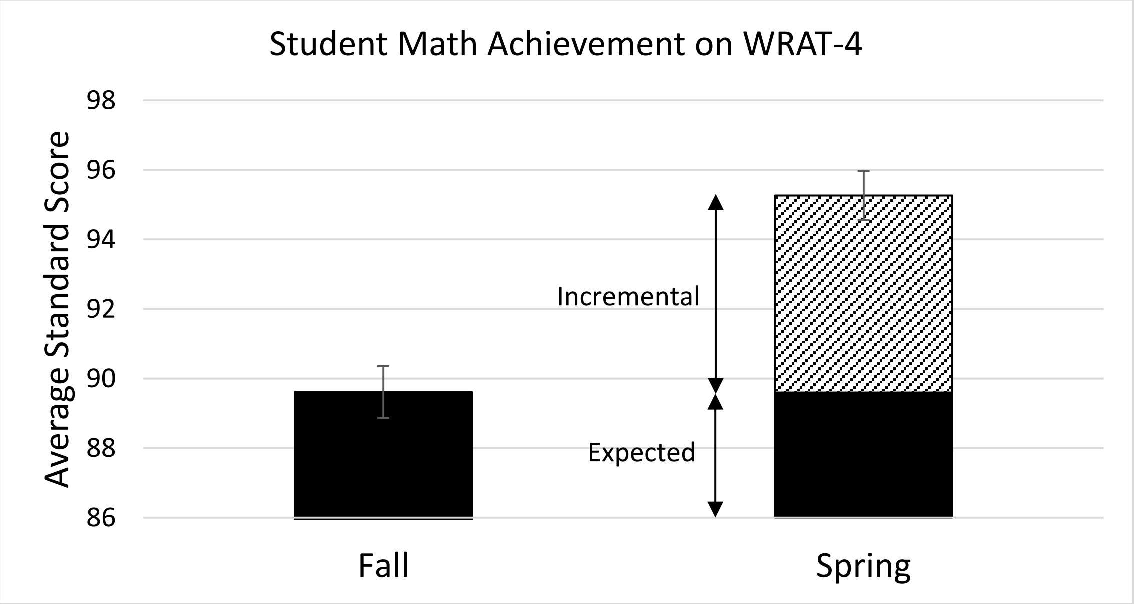
In order to assess the growth in math achievement for students participating in the NBF program, all of the non-blended grade 4 classrooms were selected for testing. Teachers were asked to administer the math computation subtest of the *Wide Range Achievement Test, Fourth Edition* (WRAT-4)[3] to their students in October 2013 and again in May 2014. 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).

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

Results

Teachers from 12 of the 14 classrooms selected for testing administered the WRAT-4 in both the fall and spring of the 2013-14 school year. Standard scores were determined for the 241 students who completed the tests in both the fall and spring of the 2013-14 school year. The mean standard score in the spring (M[4] = 95.3, SD[5] = 11.0) was significantly higher than the mean standard score in the fall (M = 89.6, SD = 11.6), paired $t(240) = -9.7$, $p < 0.001$ (see Figure 1). We would expect the students in the 2013-14 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 WRAT-4 standardization sample. The corresponding percentile rank of students (relative to the WRAT-4 standardization sample) increased from the 25th 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 0.38 ((95.3 – 89.6)/15).

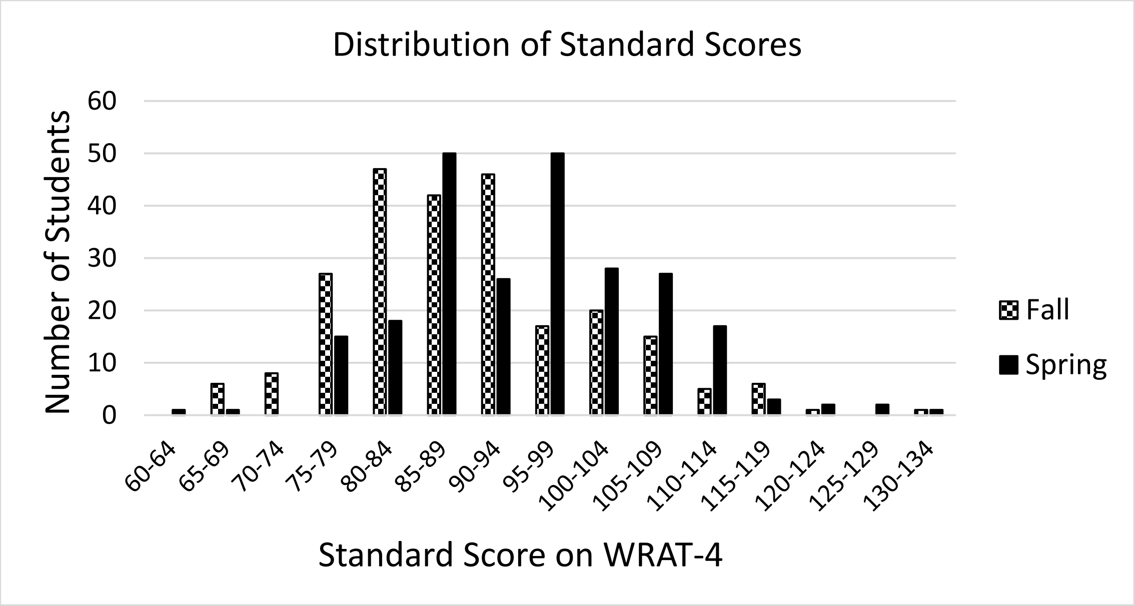
Figure 1:



[4] M= mean
[5] SD = standard deviation

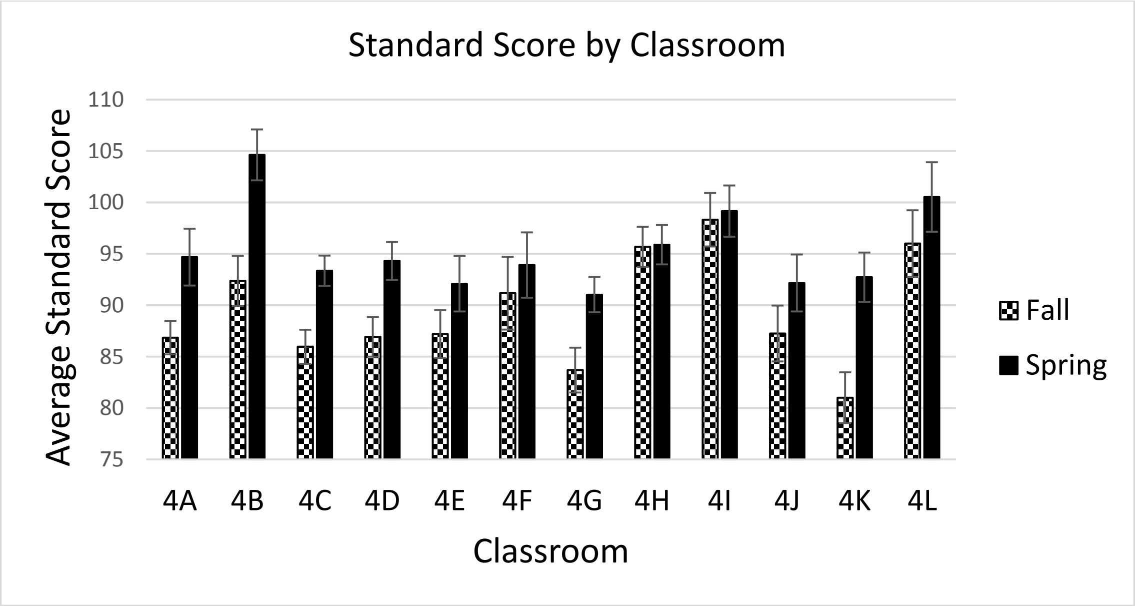
The frequency distributions of standard scores obtained in the fall and spring are shown below in Figure 2. The distributions include only those students (N=241) who completed either a blue or green test in the fall and then completed the alternate coloured test in the spring (students who completed 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 87 in the fall to 95 in the spring. The distribution of standard scores in the spring is bimodal; there are two clear peaks separated by a trough. The number of students scoring 'above average' or higher increased by 92% in the spring (25 students) compared to the fall (13 students). The number of students scoring 'below average' decreased by 35% in the spring (85 students) compared to the fall (130 students).

Figure 2:



Mean standard scores in the fall and spring for each of the twelve classrooms are shown in Figure 3 (error bars for all graphs denote the standard error of the mean (SEM)). Mean standard scores in the fall ranged from 81.0 to 98.3 whereas mean standard scores in the spring ranged from 91.0 to 104.6. The percent increase in standard score ranged from 0.2% (classroom 4H) to 14.5% (classroom 4K).

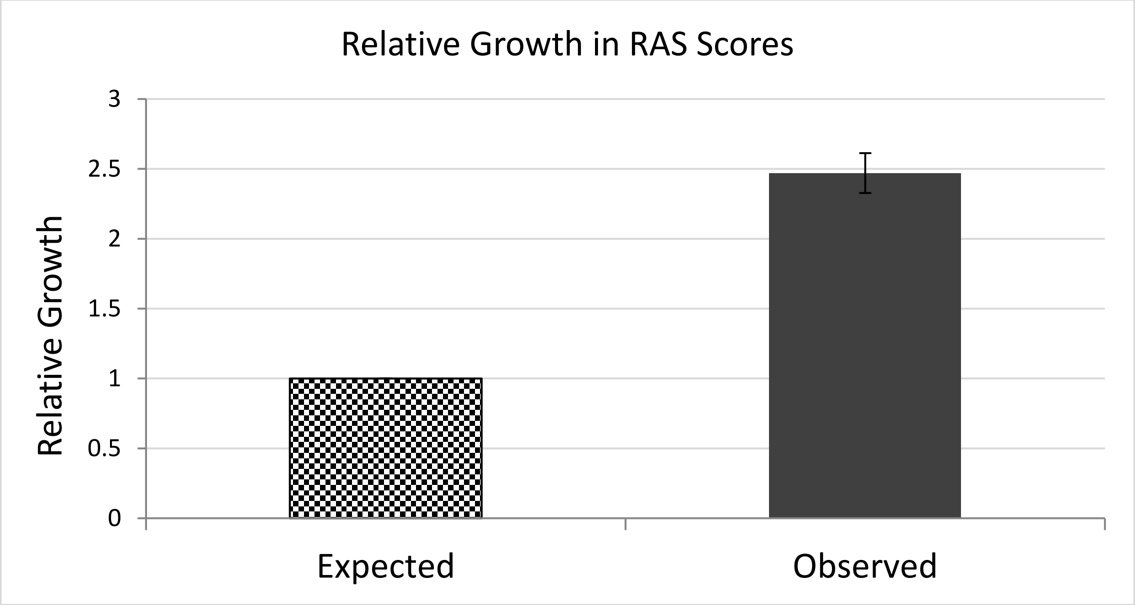
Figure 3:



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. 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. On average, math achievement of students in the 2013-14 NBF grew at 2.5 times the rate of the WRAT-4 standardization sample (Figure 4).

Figure 4:



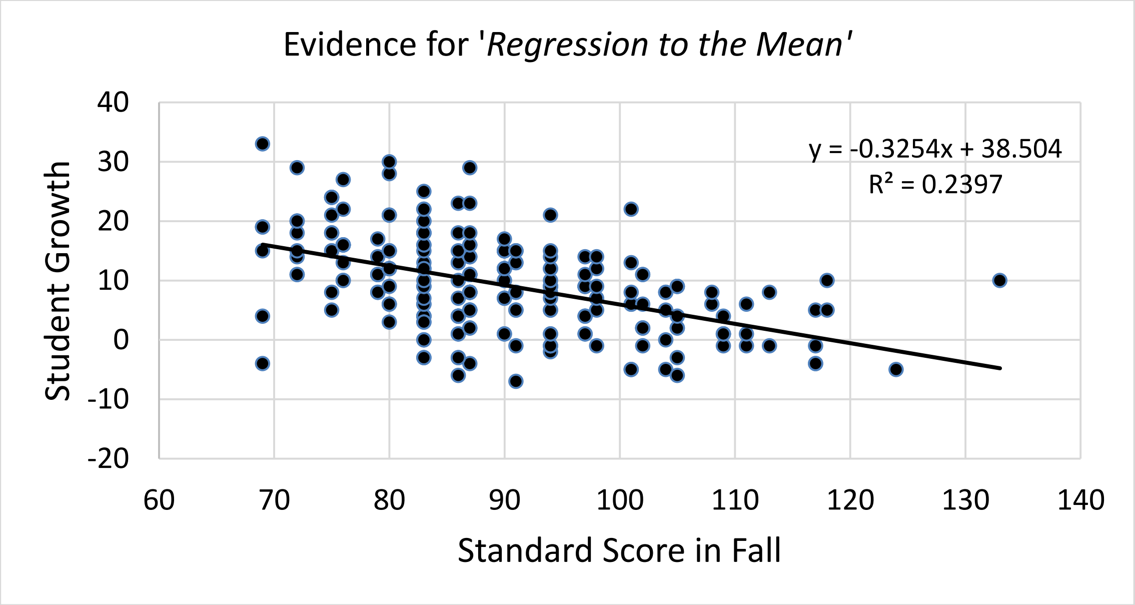
A regression analysis was performed to determine which variables were significant predictors of student achievement in the spring. We started with a simple linear model in which standard score in the spring (ss.spring) was regressed against centred standard score in the fall (ss.fall.c) and average classroom standard score in the fall (avg.ss.fall, see Appendix I.A). Standard scores in the fall were 'centred' by subtracting the overall average standard score in the fall, thus making the value of the intercept more meaningful. We did not include gender in the model because few students listed their gender on the test form. Standard score in the fall was a significant predictor of standard score in the spring ($p < 0.001$) whereas average classroom standard score in the fall was not a significant predictor ($p > 0.1$). A linear mixed model with ss.fall as a fixed variable and classroom as a random variable was also used; mixed models (that include both fixed and random effects) are able to account for the effect of clustering of students in classrooms (Appendix I.B). Coefficients for the linear mixed model are shown in Table I. Students in the same classroom tend to be more alike than students from different classrooms, as indicated by a non-zero intraclass correlation (ICC). An ICC of 0.07 for these data was calculated by dividing the component of the variance due to the random variable 'classroom' by the total variance (Appendix I.C).

Table I: Linear Mixed Effects Model

Predictor	Coefficient	Standard Error	Degrees of Freedom	t-value	p-value
Intercept	95.44	0.82	228	116.96	0.00
ss.fall (centred)	0.66	0.05	228	14.25	0.00

Figure 5 (see below) illustrates the relationship between student growth on the WRAT-4 and standard score in the fall. The regression line through these points has a slope of -0.33 and a correlation coefficient (R^2) of 0.24. 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 are discussed below (see page 10).

Figure 5:



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. The 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 who 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 2013-14 National Book Fund Program showed significant increases in mean standard score in the spring (M = 95.3, SD = 11.0) compared to the fall (M = 89.6, SD = 11.6). The corresponding percentile rank of NBF students increased from the 25th 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 92% in the spring (25 students) compared to the fall (13 students) whereas the number of students scoring 'below average' (equivalent to a standard score of 89 or lower) decreased by 35% in the spring (85 students) compared to the fall (130 students). On average, students participating in the 2013-14 National Book Fund grew at 2.5 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. a symmetric, 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. 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. We also observe a more prominent bi-modal distribution in the spring; there are two peaks in the distribution of standard scores in the spring separated by a trough. A bi-modal distribution could indicate that we had two distinct populations in our sample. In the fall, the distribution of standard scores for the two populations may have overlapped so that separate peaks are not as obvious. If one population of students experienced more growth in math achievement over the school year then this could account for the presence of two separate peaks in the spring.

JUMP Math has used the WRAT-4 to assess student math achievement since 2011-12[6]. A comparison of the test results for the past 3 years is shown in Table II. We observed a step increase in student growth between the 2011-12 and 2012-13 NBF (1.8 vs 2.8, respectively) that remained high in 2013-

[6] 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

14 (2.5). The step 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.

Table II: NBF Student Test Results

	2013-14 NBF	2012-13 NBF	2011-12 NBF
# of students tested in both fall & spring	241	286	326
Grades tested	4	4 to 7	3 and 6
SS fall vs SS spring	89.6 vs 95.3	90.8 vs 94.6	96.8 vs 100.9
Percentile rank fall vs spring	25th vs 37th	27th vs 37th	42nd vs 53rd
Average student growth relative to WRAT-4	2.5	2.8	1.8
% of students scoring 'above average' in fall vs spring	5% vs 10%	7% vs 12%	10% vs 22%
% of students scoring 'below average' in fall vs spring	54% vs 35%	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[7]. 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

[7] 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.

in the study[8]. 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[9]. We have reviewed the potential impact of each of these factors and conclude that it is unlikely they can account for the statistically significant gains in math achievement obtained for all of the 3 years of the NBF in which student testing has been completed and analyzed. 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 observed growth for each student (RAS score in spring – RAS score in fall) was plotted against their standard score in the fall (Figure 5). If RTM was present, students with a low standard score in the fall would tend to have a higher score in the spring (and therefore greater growth) whereas students with a high standard 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 5 indicate that RTM was present in these data: the slope of the regression line is negative (-0.33).

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

[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 .

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 2 shows that the shift to the right in the distribution of standard scores was evident for standard scores in the centre of the distribution (most dramatically 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.

Appendix I: Statistical Analysis in R[11]

A. Simple Linear Model

```
> model1 <- lm(ss.spring ~ ss.fall.c + class.avg.fall, data = BF.2013.14) #simple
linear model
> summary(model1)

Call:
lm(formula = ss.spring ~ ss.fall.c + class.avg.fall, data = BF.2013.14)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.6829  -5.9028  -0.1923   4.6593  26.0851 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    108.71478   10.05640  10.811   <2e-16 ***
ss.fall.c        0.67665    0.05026  13.462   <2e-16 ***
class.avg.fall  -0.15009    0.11207  -1.339    0.182    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

Residual standard error: 8.052 on 238 degrees of freedom
Multiple R-squared:  0.4674,	Adjusted R-squared:  0.4629 
F-statistic: 104.4 on 2 and 238 DF,  p-value: < 2.2e-16

> model2 <- lm(ss.spring ~ ss.fall.c*class.avg.fall, data = BF.2013.14) #check for
interactions
> summary(model2)

Call:
lm(formula = ss.spring ~ ss.fall.c * class.avg.fall, data = BF.2013.14)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.0028  -5.7008  -0.2364   5.0817  26.1867 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)              109.051816  10.072043  10.827   <2e-16 ***
ss.fall.c                  0.046215   0.775985   0.060    0.953    
class.avg.fall            -0.155929   0.112384  -1.387    0.167    
ss.fall.c:class.avg.fall   0.006957   0.008545   0.814    0.416    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.058 on 237 degrees of freedom
Multiple R-squared:  0.4689,	Adjusted R-squared:  0.4622 
F-statistic: 69.74 on 3 and 237 DF,  p-value: < 2.2e-16

> #no interactions
```

B. Linear Mixed Effects Model

```
> model3 <- lme(ss.spring ~ ss.fall.c, data = BF.2013.14, random = ~ 1|class)
> #add random effects to model
> summary(model3)
Linear mixed-effects model fit by REML
```

[11] 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/.

```
 Data: BF.2013.14
       AIC      BIC    logLik
  1694.85 1708.756 -843.4249

Random effects:
 Formula: ~1 | class
        (Intercept) Residual
StdDev:    2.179724 7.814654

Fixed effects: ss.spring ~ ss.fall.c
               Value Std.Error  DF  t-value p-value
(Intercept) 95.43609 0.8159871 228 116.9578       0
ss.fall.c    0.66479 0.0466461 228  14.2518       0
 Correlation:
          (Intr)
ss.fall.c 0.006

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-2.259874800 -0.725746333  0.003476434  0.591423333  3.282284301

Number of Observations: 241
Number of Groups: 12

> model4 <- lme(ss.spring ~ ss.fall.c, data = BF.2013.14, random = ~ 1+ss.fall.c|c
lass)
> #add random slopes to model
> summary(model4)
Linear mixed-effects model fit by REML
 Data: BF.2013.14
        AIC     BIC    logLik
  1698.292 1719.15 -843.1458

Random effects:
 Formula: ~1 + ss.fall.c | class
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev     Corr
(Intercept) 2.10529786 (Intr)
ss.fall.c   0.09083618 0.36
Residual    7.76871487

Fixed effects: ss.spring ~ ss.fall.c
               Value Std.Error  DF   t-value p-value
(Intercept) 95.41796 0.8059294 228 118.39493       0
ss.fall.c    0.67146 0.0547277 228  12.26902       0
 Correlation:
          (Intr)
ss.fall.c 0.163

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.20140149 -0.69154789 -0.03313782  0.61013530  3.29559662

Number of Observations: 241
Number of Groups: 12

> anova(model3,model4) #compare model 3 and model 4
       Model df      AIC      BIC    logLik   Test   L.Ratio p-value
model3     1  4 1694.850 1708.756 -843.4249
model4     2  6 1698.292 1719.150 -843.1458 1 vs 2 0.5581514  0.7565
> #model 4 does not significantly improve upon model 3
```

C. Calculation of Intraclass Correlation (ICC)

```
> VarCorr(model3)
class = pdLogChol(1)
            Variance  StdDev
(Intercept)  4.751195 2.179724
Residual    61.068810 7.814654
> 4.751195/(4.751195+61.068810) #calculation of ICC
[1] 0.07218466
```