Jessica Guise

National Gradewide Effect Size Trends

Summary: Based on seven years of standardized testing data collected from 37 states, plus the District of Columbia, in the United States, various comparisons of effect sizes for ST Math users can be explored-specifically, 156 sub-studies are analyzed. Looking at different testing calendar years, length of school experience with ST Math, and the math metric used to calculate effect size, no consistent trend emerged. One exception is comparing the treatment (ST Math) group based on high or low percent completion of the program. The higher percent progress (greater than 60% by April 15) demonstrated consistently higher effect sizes- with an average benefit of 0.25 effect size points for z-score of percent proficient (Z_prof), and an average benefit of 0.14 effect size points for z-score of scale score (Z_ss). Both of these findings are statistically significant. Thus, MIND will reference two standard effect sizes for Z_prof: 0.31 for the group completing at least 40% progress by April 15 and 0.48 for the group completing at least 60% progress by April 15.

Background: Since these analyses use grades as the unit of analysis, and states publish grade-mean state standardized test scores, the data for student math outcomes is collected from each state education agency's research files (retrieved from state websites). The treatment students use ST Math student accounts served by MIND. Student ST Math usage data is aggregated to grade-level means by MIND.

The Treatment grades pool originated with all schools and grades using ST Math in the United States (with sufficient state data). Grades are filtered out based on the appropriate school year of use and the length of time using the program (i.e. 2 Year 1819).

Because the analysis uses grade-mean data, such as z-score of grade-mean scale scores or z-score grade-mean proficiency level percentages, it is necessary that the program also be a grade-wide treatment, with the great majority of students in each grade receiving treatment. Otherwise, the grade-means reported by the state of 100% *tested* students would not be valid measures of a smaller fraction of *treatment* students. MIND's site implementation requirement is that an entire grade, including all teachers and all classes within that grade, use the ST Math program. We validate how closely this is the case for each individual treatment grade by comparing the number of ST Math student accounts at a grade level to the reported enrollment at that grade level. We discard from the Treatment pool any grade with a ratio of ST Math student accounts to reported grade enrollment lower than 85%.

Furthermore, the outcomes measure is a summative year-end test, i.e. that state's standardized math assessment (for example, CAASPP in CA). The math assessment thus covers all of the math standards for that entire grade level. Meanwhile, the ST Math

program curriculum (arranged into Learning Objectives) is also aligned to each state's math standards. To infer that the ST Math content is having a valid effect on student outcomes on the summative assessment, we discard any grade with grade-mean of ST Math Progress for its students lower than 40% by April 15 (Prog2A).

Progress is a percentage, and is defined as Levels completed by the student, divided by the total number of Levels in the grade-level curriculum. Note that student achievement of at least 40% progress in ST Math is accomplished primarily by teacher assignment of computer session time to students. With sufficient time on task, students make progress. The program helps them self-pace through providing real-time informative feedback for each puzzle.

In order to accumulate enough grades for a single analysis, there must be a sufficient number of schools using ST Math with fidelity. This is often the case at a district-level, but it can also be accomplished by looking across an entire state. When conducting a national study across different states, each state's standardized test score must be normalized for valid comparisons-calculating a new z-score is the statistical method used in these analyses. Two different z-score metrics are used: z-score of math proficiency (Z_prof)- i.e. percentage of students in the top two proficiency levels and z-score of mean scale score (Z_ss). These calculations are done by state, by year, by grade. This is also valuable when comparing the same state across different exams.

Previous studies demonstrate the positive correlation between more ST Math progress and higher gains on state assessments when compared to matched controls. The below chart summarizes findings from a 1718 study aggregated across 31 states with 1-5 year(s) of ST Math usage.

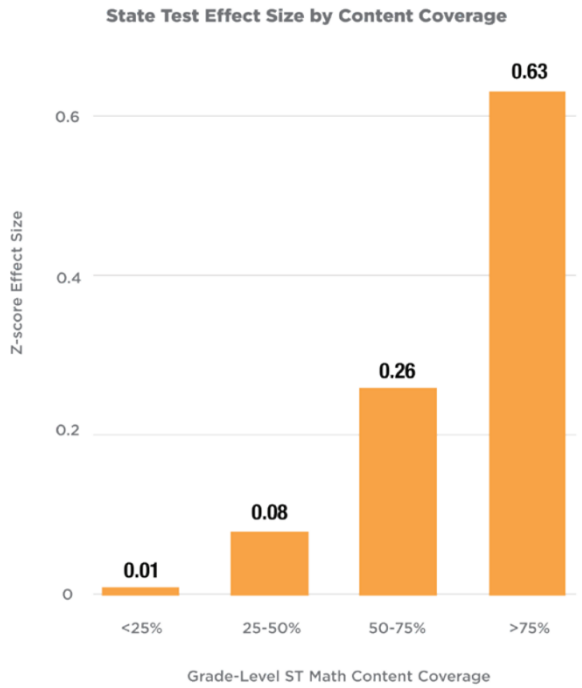


Figure 1 https://info.mindresearch.org/hubfs/Collateral/Infographics/Infographic_ST_Math_TheoryofChange.pdf

MIND believes in going beyond the single “gold standard” study to test an edtech product. While a randomized control trial (RCT) is certainly valuable, it is often time consuming, expensive, and hard to achieve sufficient power for significance; thus, it is difficult to stay up to date on testing a product by RCT, alone. An alternative to an RCT is a quasi-experimental analysis. This allows us to analyze multiple cohorts of schools year after year, yielding repeatable, statistically significant results.

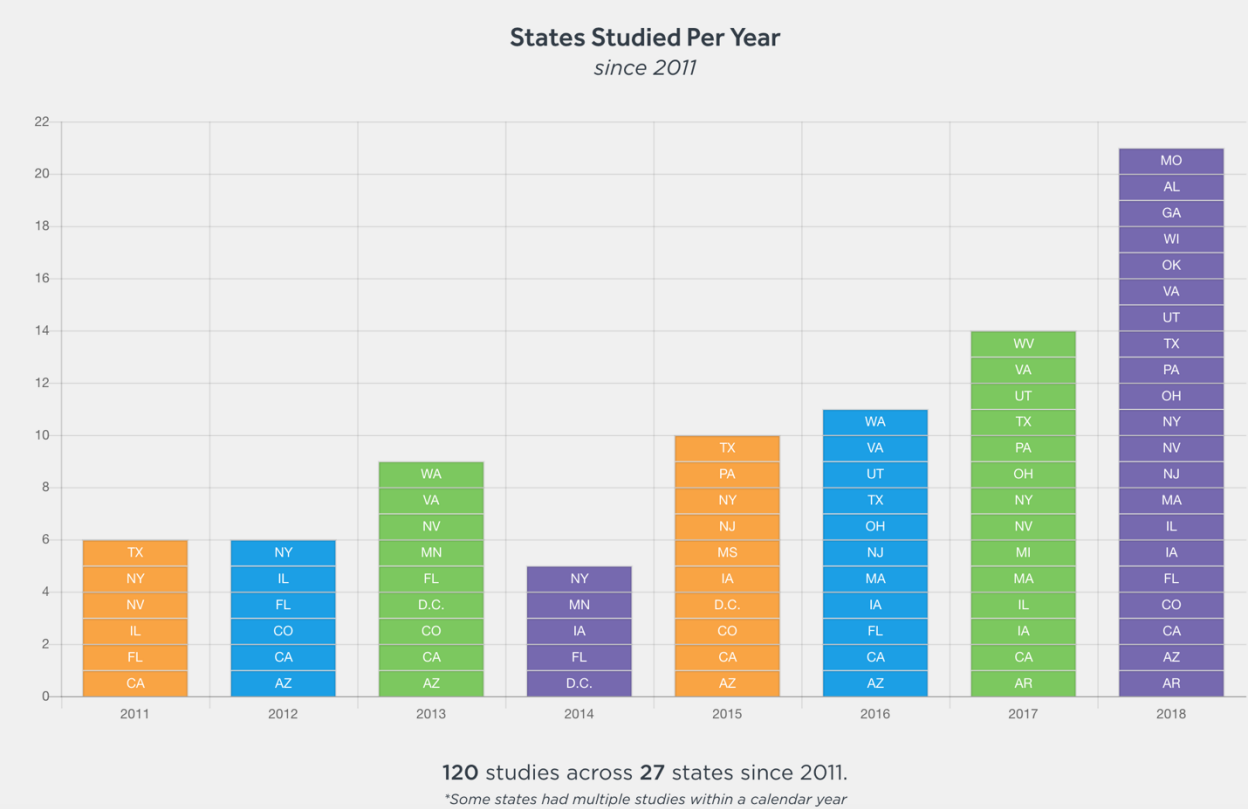


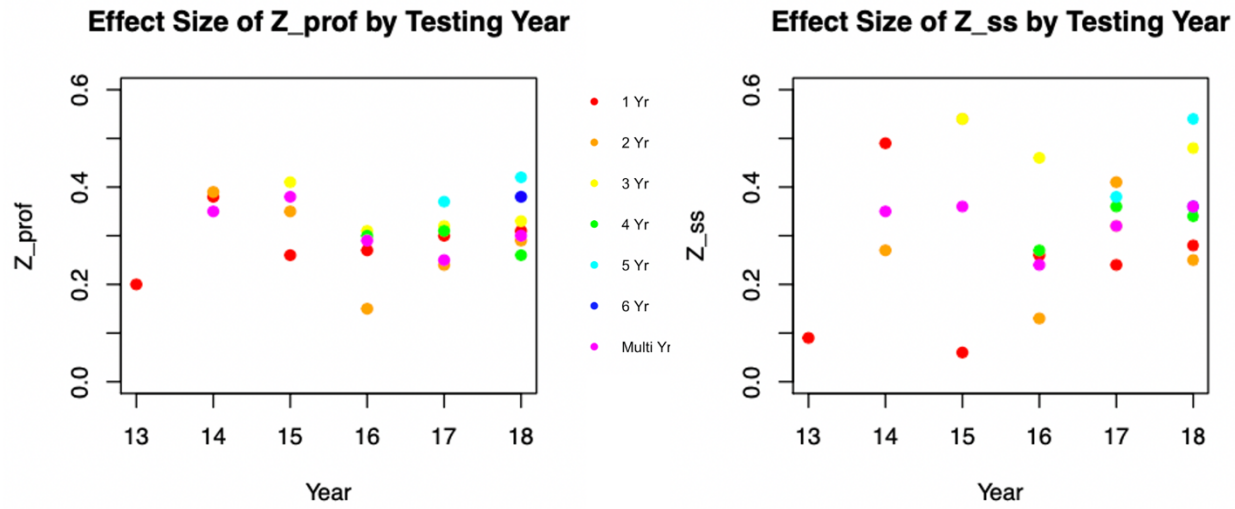
Figure 2 https://www.stmath.com/impact

MIND also believes in accountability and transparency. An independent, third party research group, WestEd, validated MIND’s methodology and published their own nationwide study using 1516 state test results. WestEd matched ST math users to a control group via propensity score matching. Then, outcome measures were analyzed using multiple linear regression and hierarchical linear modeling. This differs from MIND’s use of matching by a Monte Carlo method and use of t-tests on determining statistical significant of difference of means between the two groups. Both methods yield similar, positive results in favor of the ST Math group.

Dataset: Multiple data sources are utilized to conduct this research. MIND’s national grade-aggregated dataset with all years of ST Math usage data is merged with each state’s grade-aggregated data by relevant years-before beginning ST Math use and the final year of ST Math use. Finally, demographic data is obtained from MDR data for matching purposes.

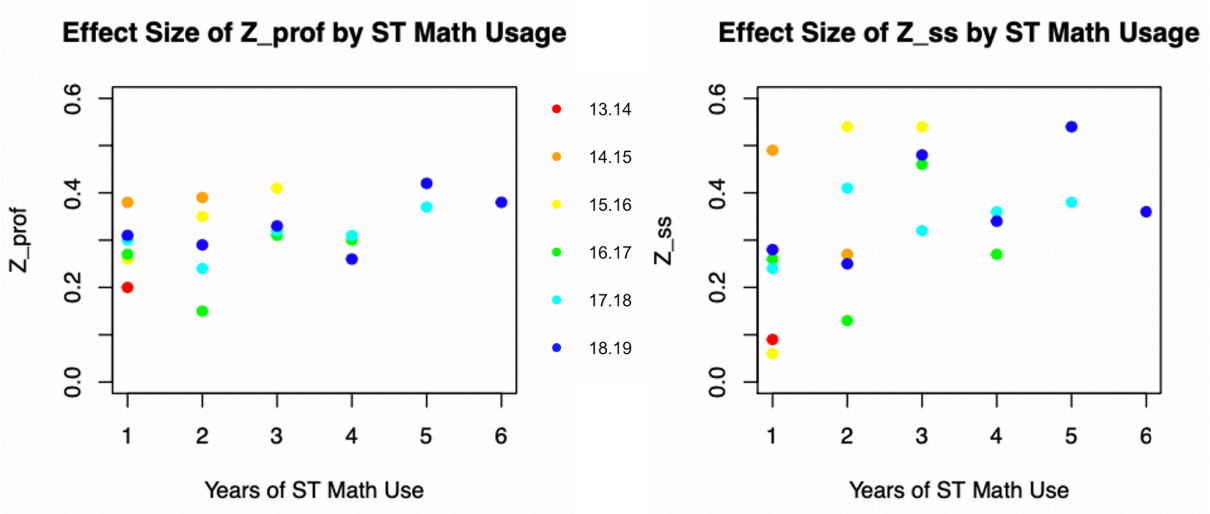
Research Questions: This report aims to answer three questions that arose from the national usage trend data: 1. Is there any difference in effect size based on the testing school calendar year? 2. Does effect size increase as the school's number of years using ST Math increases? 3. Is there a difference in effect size between the lower progress (40-60% by April 15) and higher progress (60+% progress by April 15) subgroups?

Question 1: Based on the currently available data, effect size does not vary by school calendar year. Despite a few outliers, the data stays relatively stable.



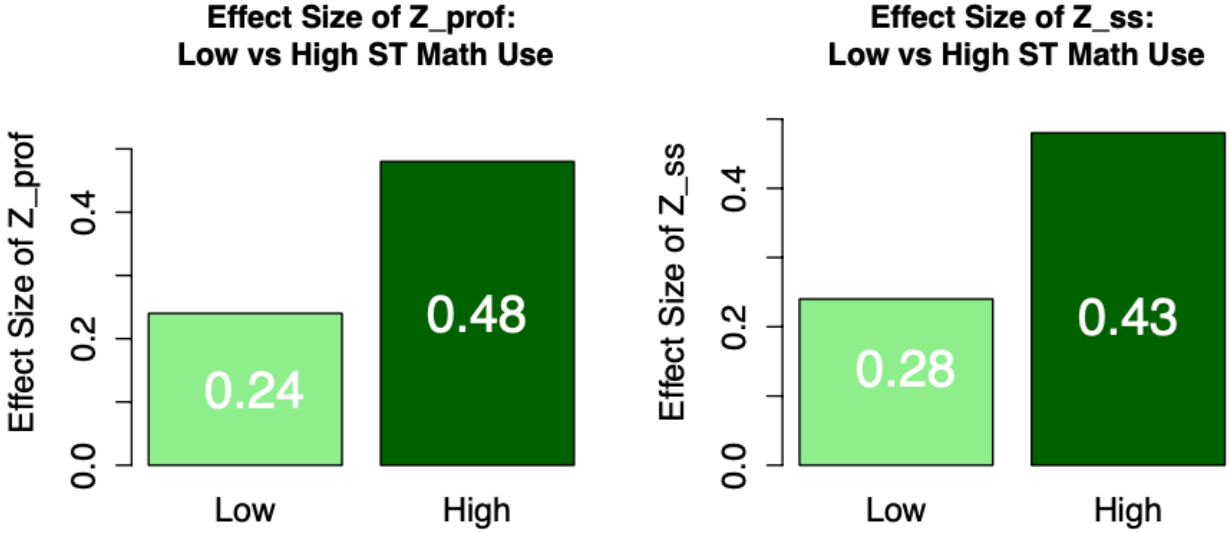
The above plots illustrate the variability in effect sizes. While a few differences between individual school years proved to be statistically significant, taken altogether, this is not enough to notice a trend in the data by school-year (Index a).

Question 2: Similarly to the first question, effect sizes vary by years of experience without a clear pattern. Overall, the effect sizes are bounded between 0.06 and 0.54 (for all grades, not split by progress). However, the low effect sizes (below 0.1) are only seen in 1 Year and 2 Year studies. The minimum effect size for either Z_prof or Z_ss goes up to 0.26 for 3-6 year studies. Despite this anecdotal difference, when grouping the effect sizes together in two groups: 1-2 years of ST Math use versus 3+ years of ST Math use, the difference was not statistically significant.



The above plots show the distribution of effect sizes by years of ST Math usage. Although a few differences proved to be statistically significant, taken together, there is not enough evidence to support a trend (Index b).

Question 3: As mentioned in MIND's Theory of Change, the more time on ST Math, the more progress made through the program, and subsequently, the more math gains noticed on standardized tests. Particularly, in this report, the treatment set is divided into a low dose (Prog2A between 40% and 60%) and a high dose (Prog2A>60%).



When aggregating across all school years and years of using ST Math, the above graphs demonstrate the higher effect sizes for the high dose group. The high dose group outperformed the low group with a statistically significant difference of 0.24 for effect size of Z_prof, in addition to a statistically significant difference of 0.15 for effect size of Z_ss (Index c).

Limitations: The availability of sufficient state data is one limitation. Some state data was not found easily on state websites and data requests did not yield a response. Some states have years of missing data due to transitioning of exams or issues with administration. Finally, some schools/grades in particular have missing data due to only a small number of students being tested and the need to protect privacy of student data.

Further, the appearance of a correlation between an increase in math gains among those using ST Math for three or more years isn't necessarily due to more years of experience- rather, these grades might be more engaged with or have higher usage of ST Math (or Math in general).

Finally, the repeated use of the same cohort of schools in multiple sub-studies (i.e. studying the same district's schools after 1 year, 2 years, 3 years, etc. of use) within this analysis could be seen as a redundancy or even cherry-picking. However, the rationale behind this choice was considered in light of the purposes of this study. Overlap of analytic samples' schools was inevitable based on the comparisons between years using the program, testing year, math attribute, and percent completion of the program.

Future Research: This work can be expanded on as more schools continue to use ST Math with fidelity and as more state data becomes available. In addition, this work establishes a 6 year-long baseline of ST Math Gen 5 effectiveness from which to compare future MIND Research Institute math program generations.

Index a:

Testing Year	Years of Use	Progress Group	Num Grades (Z_p)	Avg Prog (Z_p)	Effect Size (Z_p)	Num Grades (Z_ss)	Avg Prog (Z_ss)	Effect Size (Z_ss)
1314	1	40-100	175	54.58	0.2	154	55.08	0.09
1314	1	40-60	125	48.79	0.13	107	48.86	0.07
1314	1	60-100	50	69.05	0.4	47	69.24	0.16
1415	1	40-100	133	57.17	0.38	80	62.12	0.49
1415	1	40-60	93	50.05	0.39	44	52.2	0.57
1415	1	60-100	40	73.73	0.34	36	74.24	0.39
1415	2	40-100	333	55.86	0.39	292	55.92	0.27
1415	2	40-60	223	49.08	0.25	198	49.21	0.22
1415	2	60-100	110	69.62	0.71	94	70.05	0.36
1415	Multi	40-100	473	56.19	0.35	372	57.25	0.35
1415	Multi	40-60	321	49.31	0.35	242	49.76	0.35
1415	Multi	60-100	152	70.73	0.35	130	71.21	0.38
1516	1	40-100	153	55.9	0.26	121	56.37	0.06
1516	1	40-60	101	49.94	0.21	82	50.47	-0.01
1516	1	60-100	52	67.47	0.35	39	68.76	0.22
1516	2	40-100	216	56.76	0.35	93	61.68	0.54
1516	2	40-60	135	49.57	0.18	39	51.55	0.51
1516	2	60-100	81	68.73	0.72	54	68.99	0.58
1516	3	40-100	379	56.34	0.41	93	61.68	0.54
1516	3	40-60	250	49.44	0.33	39	51.55	0.51
1516	3	60-100	129	69.72	0.57	54	68.99	0.58
1516	Multi	40-100	748	56.37	0.38	557	57.42	0.36
1516	Multi	40-60	486	49.58	0.34	343	50	0.26
1516	Multi	60-100	262	68.97	0.44	214	69.31	0.51
1617	1	40-100	285	54.86	0.27	213	55.13	0.26
1617	1	40-60	199	49.24	0.19	148	49.37	0.19
1617	1	60-100	86	67.84	0.5	65	68.26	0.4
1617	2	40-100	274	56.02	0.16	209	56.89	0.13
1617	2	40-60	178	49.12	0.03	129	49.07	0.14
1617	2	60-100	96	68.83	0.35	80	69.5	0.12
1617	3	40-100	226	56.92	0.31	94	64.01	0.46
1617	3	40-60	141	49.6	0.16	33	53.84	0.3
1617	3	60-100	85	69.06	0.63	61	69.51	0.62
1617	4	40-100	381	60.32	0.3	351	61.16	0.27
1617	4	40-60	197	50.25	0.16	171	50.53	0.1
1617	4	60-100	184	71.1	0.46	180	71.23	0.45
1617	Multi	40-100	1166	57.32	0.29	867	58.95	0.24
1617	Multi	40-60	715	49.56	0.16	481	50.01	0.12
1617	Multi	60-100	451	69.61	0.48	282	69.61	0.47
1718	1	40-100	108	51.33	0.3	87	51.45	0.24
1718	1	40-60	91	49.1	0.3	75	49.63	0.23
1718	1	60-100	17	63.28	0.32	12	63.04	0.09
1718	2	40-100	235	52.73	0.24	163	52.53	0.41
1718	2	40-60	184	49.07	0.17	130	48.91	0.4
1718	2	60-100	51	65.95	0.47	33	66.77	0.41
1718	3	40-100	201	53.2	0.32	131	54.25	0.32
1718	3	40-60	155	49.03	0.27	98	49.68	0.26
1718	3	60-100	46	67.24	0.48	33	67.83	0.46
1718	4	40-100	180	58.23	0.31	88	61.72	0.36
1718	4	40-60	103	50.57	0.26	40	53.12	0.3
1718	4	60-100	77	68.48	0.41	48	68.88	0.44
1718	5	40-100	319	58.16	0.37	306	58.2	0.38
1718	5	40-60	195	49.97	0.29	187	49.97	0.28
1718	5	60-100	124	71.04	0.51	119	71.12	0.55
1718	Multi	40-100	1028	55.4	0.25	774	56	0.32
1718	Multi	40-60	713	49.56	0.21	528	49.83	0.29
1718	Multi	60-100	315	68.62	0.34	246	69.61	0.39
1819	1	40-100	247	51.9	0.31	187	52.14	0.28
1819	1	40-60	200	47.93	0.28	151	48.05	0.25
1819	1	60-100	47	68.79	0.46	36	69.32	0.41
1819	2	40-100	263	53.57	0.29	178	51.13	0.25
1819	2	40-60	199	48.63	0.28	137	48.34	0.23
1819	2	60-100	64	68.92	0.35	41	69.13	0.3
1819	3	40-100	230	53.9	0.33	148	55.01	0.48
1819	3	40-60	172	49.23	0.29	106	49.51	0.46
1819	3	60-100	58	67.73	0.48	42	68.87	0.58
1819	4	40-100	204	54.91	0.26	155	55.25	0.34
1819	4	40-60	144	49	0.19	107	48.73	0.25
1819	4	60-100	60	69.1	0.47	48	69.77	0.61
1819	5	40-100	149	57.59	0.42	95	59.14	0.54
1819	5	40-60	93	49.68	0.26	55	50.06	0.42
1819	5	60-100	56	70.74	0.84	40	71.62	0.71
1819	6	40-100	212	58.58	0.38	185	58.05	0.36
1819	6	40-60	124	50.31	0.22	110	50.41	0.32
1819	6	60-100	88	70.23	0.6	75	69.25	0.43
1819	Multi	40-100	1305	54.79	0.3	948	55.14	0.36
1819	Multi	40-60	932	48.98	0.24	666	49.01	0.32
1819	Multi	60-100	373	69.33	0.48	282	69.61	0.47

Index b: Absolute Effect Sizes by ST Math Dose (Percent Progress)

	Z_prof Effect Size	Z_ss Effect Size
P2A>=40	0.31	0.33
P2A>=60	0.48	0.43

Index c: Comparison by ST Math Dose (Percent Progress)

	Estimate	P.Value	Int.Low	Int.High
Z_prof Effect sizes: 40-60% P2A vs 60-100% P2A	-0.25	0.00	-0.31	-0.18
Z_ss Effect sizes: 40-60% P2A vs 60-100% P2A	-0.14	0.00	-0.22	-0.06

Index d: Comparison of School Calendar Years

	Estimate	P.Value	Int.Low	Int.High
18.19 vs 17.18	0.03	0.34	-0.03	0.09
18.19 vs 16.17	0.06	0.12	-0.02	0.15
18.19 vs 15.16	-0.02	0.58	-0.12	0.07
18.19 vs 14.15	-0.05	0.09	-0.10	0.01
17.18 vs 16.17	0.03	0.36	-0.05	0.12
17.18 vs 15.16	-0.05	0.23	-0.15	0.04
17.18 vs 14.15	-0.08	0.01	-0.13	-0.02
16.17 vs 15.16	-0.09	0.09	-0.19	0.02
16.17 vs 14.15	-0.11	0.02	-0.19	-0.03
15.16 vs 14.15	-0.02	0.54	-0.12	0.07

Table 1 Differences in Effect Sizes of Z_prof Across School Years

	Estimate	P.Value	Int.Low	Int.High
18.19 vs 17.18	0.03	0.47	-0.07	0.14
18.19 vs 16.17	0.10	0.17	-0.05	0.25
18.19 vs 15.16	-0.00	0.99	-0.34	0.34
18.19 vs 14.15	0.00	0.97	-0.22	0.22
17.18 vs 16.17	0.07	0.30	-0.08	0.21
17.18 vs 15.16	-0.04	0.77	-0.39	0.31
17.18 vs 14.15	-0.03	0.68	-0.27	0.21
16.17 vs 15.16	-0.10	0.45	-0.44	0.23
16.17 vs 14.15	-0.10	0.30	-0.32	0.12
15.16 vs 14.15	0.01	0.97	-0.34	0.35

Table 2 Differences in Effect Sizes of Z_ss Across School Years

Index e: Comparison of ST Math Usage (Seniority)

	Estimate	P.Value	Int.Low	Int.High
1 Yr vs 2 Yr	0.00	0.96	-0.11	0.12
1 Yr vs 3 Yr	-0.06	0.14	-0.13	0.02
1 Yr vs 4 Yr	-0.00	0.91	-0.07	0.06
1 Yr vs 5 Yr	-0.11	0.05	-0.22	-0.00
2 Yr vs 3 Yr	-0.06	0.27	-0.18	0.06
2 Yr vs 4 Yr	-0.01	0.90	-0.12	0.11
2 Yr vs 5 Yr	-0.11	0.07	-0.24	0.02
3 Yr vs 4 Yr	0.05	0.12	-0.02	0.12
3 Yr vs 5 Yr	-0.05	0.23	-0.17	0.06
4 Yr vs 5 Yr	-0.11	0.08	-0.25	0.04

Table 3 Differences in Effect Sizes of Z_prof by ST Math Usage

	Estimate	P.Value	Int.Low	Int.High
1 Yr vs 2 Yr	-0.08	0.40	-0.30	0.13
1 Yr vs 3 Yr	-0.21	0.03	-0.39	-0.03
1 Yr vs 4 Yr	-0.09	0.25	-0.25	0.08
1 Yr vs 5 Yr	-0.22	0.14	-0.59	0.15
2 Yr vs 3 Yr	-0.13	0.17	-0.33	0.07
2 Yr vs 4 Yr	-0.00	0.97	-0.20	0.19
2 Yr vs 5 Yr	-0.14	0.29	-0.50	0.22
3 Yr vs 4 Yr	0.13	0.07	-0.02	0.27
3 Yr vs 5 Yr	-0.01	0.93	-0.48	0.46
4 Yr vs 5 Yr	-0.14	0.32	-0.83	0.55

Table 4 Differences in Effect Sizes of Z_ss by ST Math Usage