edmentum™

A Quasi-Experiment of the Effectiveness of Study Island Technology on Elementary Student Achievement

Jeff McLeod, PhD

Edmentum, Inc.
Bloomington, Minnesota

Individualized learning, simply stated, is the strategy of teaching across the full range of students in a classroom. Computer technology to a very great extent makes this possible by automating the assignment of lessons and practice to learners based on their readiness to learn. Computer assisted instruction, as one form of individualized learning, provides supplementary tutoring and practice testing to reinforce student learning. Automated platforms continuously engage, monitor, and assess student progress. Mastery based programs provide lessons and brief quizzes with feedback about the correct and incorrect answers, and can be repeated until mastery is achieved. This individualized mastery approach, introduced by Carroll (1963, 1989), and Bloom (1973) have a strong evidentiary basis.

Evidence of Effect of Computer Assisted Practice Testing

Guskey and Gates (1986) were first to publish a major research synthesis showing the mastery testing approach has positive effects on learning even without technology. Kulik, Kulik, and Bangert-Drowns (1990) published a widely cited meta-analysis of interventions from both paper-pencil and computer assisted environments. They found a robust, positive benefit to individualized mastery-learning for a wide range of outcomes including increasing end-of-course standardized test scores. The authors found the average effect size of frequent mastery testing was .52, which is considered moderately strong.

Kulik et al confirmed their earlier findings in a subsequent meta-analysis (Kulik, Kulik, and Bangert-Drowns, 1991) where they focused exclusively on studies using frequent practice testing. They confirmed that practice testing produces moderate gains in achievement. The positive benefits of practice with frequent testing is greater for students who are struggling the most.

The evidence-base for mastery learning continues to grow as seen in further meta-analytic reviews by Kulik, Kulik, and Bangert-Drowns (1991), Cheung and Slavin (2012), and Cheung and Slavin (2013). The influence of the practice-testing model has begun to spread outside American schools, having been found to be effective in international contexts (e.g., Jain, 2015). These newer studies, which focus exclusively on computer-based interventions, confirm the finding that computer-assisted instruction has a small to moderate effect in the subject of Math, and a small but statistically significant effect for the subject of Reading.

Mechanism of the Treatment Effect

Empirical evidence in favor of a learning model should rest on a plausible theory of learning. Previous researchers suggested that the mechanism of effect derives from stimulating practice and review, opportunity for feedback, and positive influence on study time. More recently, a plausible learning theory has developed among cognitive educational psychologists. These studies report that practice-test items that are featured in computer assisted instruction give learners *retrieval practice*. Karpicke and Roediger (2008) published an article in the journal Science entitled *The Critical Importance of Retrieval for Learning*, in which they provide evidence that retrieval practice is *more effective* than common-sense homework tasks like reviewing

material. The authors indicated it is more effective from a cognitive perspective to *practice producing correct answers to questions about the lessons*. Experimental research by Roediger (2014), Roediger and Karpicke (2006), and McDermott, Agarwal, D'Antonio, and Roediger & McDaniel (2014) has confirmed the "testing effect" and the benefits of mastery learning in computer mediated environments.

The achievement gains have been documented both on teacher tests, classroom summative tests, and end-of-course standardized tests. These authors found, like the authors cited above, that learning through mastery testing was particularly effective for lower performing students.

Study Island is one form of online computer-assisted instruction that provides the retrieval practice that researchers suggest is potentially effective for learning. This online learning platform offers a deep store of practice items, motivational tools like blue- ribbons, gamified interfaces, and interactive lessons. Like other high quality learning platforms, the exercises and test items have been aligned to national standards. Study Island is understood to make subject matter information more accessible to learners, and thereby promote growth by preparing the student for further learning.

Research Question

Does usage of the Study Island educational platform contribute to academic growth in Math and Literacy/Language Arts for children in Elementary schools, grades 3-6?

Method

Participants. A sample of eight schools was drawn from a national study of school districts in a norm-setting study for Edmentum's Exact Path vertical scale. Study Island is an Edmentum product as well, so it was possible to identify a treatment group of 4 schools in the norm sample who also had active licenses to use Study Island in the classrooms during the 2016-17 school year, and had data indicating the platform was being used. A comparison group of 4 schools was selected based on nearest neighbor matches to the four experimental schools based on pre-test scores. All four comparison schools were verified not to have had access to the Study Island platform.

The use of multiple schools in each of the treatment and comparison groups mitigates the confounding of treatment effects with school.

Limited demographic information was attached to student records at the school level. Table 1 below compares the treatment and control groups on these district level variables. The profiles suggest that the treatment group, on average, had higher Latino/Latina student populations, a higher proportion of English Language Learners (ELL), a higher student/teacher ratio, and fewer Caucasian students.

Table 1.

Demographics of Groups

Group		ELL	IEP	Stu/Teacher Ratio	Latino	Caucasian
Control	Mean	.01	.16	13.71	.05	.92
	N	1055	1055	1055	1055	1055
	SD	.01	.03	2.16	.02	.08
Treatment	Mean	.06	.11	15.89	.08	.89
	N	582	582	582	582	582
	SD	.05	.03	.51	.12	.12
Total	Mean	.03	.14	14.49	.06	.91
	N	1637	1637	1637	1637	1637
	SD	.04	.04	2.05	.08	.10

Note. ELL = proportion of ELL in the school district, IEP is proportion individual education plan students in the district, stu/teacher ratio is the student teacher ratio, Latino = proportion of Latino students in the district, and Caucasian = proportion of Caucasian students in the district.

Procedure. As mentioned, this study was coordinated within a larger study in which Edmentum's Math, Reading, and Language Arts achievement tests were being administered to collect growth norms for a technical report. This study used a national, representative sample of fourteen school districts was recruited from the states of Florida, California, Pennsylvania, Wisconsin, Idaho, New Jersey, Arizona, Minnesota, and Michigan.

Schools in the participating districts were given access to Edmentum's Exact Path achievement test at no cost. The assessments were to be administered during three testing windows across the 2016-17 school year (fall, winter, and spring).

There were four schools in the sample that had access to the Study Island learning platform. All schools were confirmed to be active users by consulting operational usage data. A group of four non-Study Island schools was selected from the remaining schools in the beta sample. The only condition for selection was that the pre-test mean scores for the control group should be comparable to the pre-test scores of the treatment group.

Treatment

Study Island. Study Island is an e-learning program, a form of computer assisted instruction, which presents online assessments, instruction, and test preparation – i.e., practice testing – using curriculum and test questions that are strongly aligned to nationwide and state standards including Common Core College and Career Readiness standards. Students can be assigned to practice on grade appropriate skills. Progress can be checked using brief quizzes where mastery is defined as 80% items correct.

The product gives Blue Ribbons when milestones are met in standards mastery. There is not a single implementation model for Study Island. Some children simply do not work to attain the blue ribbons. The decision to compete for blue ribbons is a motivational variable which can be viewed as the personal engagement of the child.

Time. The design is an intent-to-treat paradigm, in which the treatment was allowed to be administered in the most natural ways in the classroom without these cases being eliminated. However, a treatment window was carefully enforced in which a total of 12 weeks of computer assisted instruction and testing had elapsed between pre-test and post-test measures.

Measures

Exact Path Achievement Scale. The achievement scales for ELA, Math, and Reading are sophisticated vertical scales delivered my means of computer adaptive logic. They are constructed using modern IRT methods, which means that all items are calibrated so that each has an objective difficulty level. This makes it possible for the adaptive test to deliver equivalent forms of the test at each administration.

Because the scale is a vertical scale, it is equated across multiple grades. In this case, it is equated across grades K-6. The implication of a vertical scale is that scores from one grade to the next are interpretable on the same scale. Furthermore, scores at the beginning and end of an educational intervention are directly interpretable in terms of growth. The post-test score minus the pre-test score is a direct measure of increase in ability.

Edmentum's Exact Path scales have measured reliability between .80 and .95, very strong construct validity, and concurrent validity as indicated by sensitivity of .95 and specificity of .90 - .95. (Edmentum Exact Path Provisional Technical Manual, 2016).

Because the items are not repeated at pre-test and post-test in an adaptive test, the threat of pre-testing on internal validity is minimized. Nor do the tests sensitize students to the dependent variable because an academic situation is one in which testing is expected, and the content of tests is highly recognizable as being part of the usual school curriculum. Nothing is being revealed or tipped off to the learner.

Propensity Score

Rosenbaum and Rubin (1983) developed a method to assist in removing confounds from treatment conditions. This study used logistic regression to create a propensity score. The dependent variable was binary (control group = 0, treatment group = 1). Predictor variables were Ethnic category (Latino percent), student teacher ratio, and a dummy variable for Midwest Census region. All coefficients were statistically significant at $p < .0001$. Correlations of the propensity score with the grouping variable were $r = .74$ (Math), $r = .55$ (ELA), and $r = .72$ (Reading).

Research Design

A quasi-experimental outcome study was performed using a fully crossed 2 x 4 factorial design. The first independent variable was Study Island usage with two levels (treatment vs. no-treatment control). The second independent variable was learner's grade with four levels (3, 4, 5, and 6). The dependent variable was a post-test score from the Exact Path vertical growth scale. Pre-test scores were taken by students in the fall. These scores were analyzed to determine whether there was baseline equivalence. Analysis indicated a need for a propensity-score matching (Rosenbaum and Rubin, 1983) to control for baseline differences between treatment and control.

Covariates. As discussed, the study involved a pre-test for each subject (Language Arts, Math, and Reading). This is normal practice for classrooms that use formative testing as an educational strategy. All schools administered the pre-tests at the beginning of the school year, and a follow-up test later after 12 weeks of instruction.

The researchers used the achievement pre-test as a covariate in the analysis of variance design along with the propensity score to statistically control for pre-existing differences and achieve some degree of pseudo-random assignment. This has the effect of ruling out the rival hypothesis to explain systematic differences in achievement at the time of sample selection.

Dependent Variables. There are three dependent variables, a Math test, a Reading test, and a Language Arts test. Each test has its own pre-test. The researchers consider this to be three separate statistical analyses.

Statistical Analysis

Analysis. A 2 x 4 univariate linear model was fit with each of the dependent variables separately (Math, Reading, and Language Arts achievement scales). The two main effects (treatment, and grade) and the interaction were tested with the propensity score entered into the general linear model as a covariate.

Statistical Power. The post-hoc statistical power for a 2-way analysis of covariance was performed using G*Power software. Since there are 3 dependent variables, a conservative alpha level of .01 is selected. The statistical model thus specified has power over .99 to detect a medium effect size.

Effect Size. The effect size partial eta-squared will be reported, and will be converted to Cohen's d for comparison with effects that have been obtained in previous studies of the effect of computer assisted instruction on achievement outcomes.

Results

Baseline Equivalence

The four schools in the treatment condition were matched as closely as possible to 4 control schools based on the pre-test scores. To measure baseline equivalence, differences between pre-test scores on the three dependent variables were calculated and expressed in control-group standard deviation units. Mean ELA pre-test scores differed by .17 standard deviations, mean Reading pre-test scores differed by .13 standard deviations, and mean Math pre-test scores differed by .10 standard deviations. These are small but non-trivial deviances. Therefore, propensity scores were computed, as described in the method section, to match the conditions with greater precision. Propensity scores correct for systematic but decisive factors that might disproportionately influence the probability of a student being included in the treatment group and thus bias group comparisons.

Method Check for Treatment Implementation

Table 2 shows that the treatment was in fact implemented during the 12-week instruction window. The treatment group had an average of 25 sessions, or 2 sessions per week on the Study Island platform. The average minutes per week is just over 10 minutes. This is lower than the amount of time that Cheung and Slavin (2012, 2013) found to be the optimum dose of computer assisted instruction, which these researchers found to be just under 30 minutes per week.

Students in the treatment group were administered quizzes averaging around 13 items per week. The learners achieved on average 60% accuracy in their learning tasks. This is very important because it indicates that children were intentionally engaged in practice. This is further indicated

by the fact that the average number of motivational rewards or “blue ribbons” earned by children was 3.61 over the 12- week instruction period.

Overall, it appears that this intent-to-treat design resulted in a valid implementation of the treatment, with the cautionary note that the actual time on task was lower than is optimal for computer assisted instruction.

Table 2.

Study Island Usage Averages for Treatment Group

Usage Variable	M	N	SD
Sessions	25.60	1809	29.83
Minutes of Usage	121.06	1809	126.65
Total Questions	158.44	1809	193.03
Percent Correct	60.50	1809	18.85
Blue Ribbons	3.61	1809	4.93

Univariate Tests of Treatment Effect

Math

Test number one was for the effect of treatment, grade, and treatment x grade interaction with the propensity score as a covariate. The dependent variable is the post-test math achievement score after 12 weeks of instruction.

The propensity score was significant (F (1, 2157) = 14.32, $p < .01$; $\eta^2 = .007$), as was the pre-test score ((F (1, 2157) = 3964.30, $p < .01$; $\eta^2 = .65$), treatment (F (1, 2157) = 32.82, $p < .01$; $\eta^2 = .015$), and grade (F (1, 2157) = 8.72, $p < .01$, $\eta^2 = .012$). The treatment x grade interaction term was marginally significant (F (3, 2157) = 2.18, $p = .08$, $\eta^2 = .003$).

The Cohen’s d effect size equivalent for the treatment effect $\eta^2 = .015$, using methods described by Cohen (1988) is $d = .25$, which is considered a small but statistically significant effect.

Table 3.
Adjusted Post-test means of treatment versus control for Math

			95% Confidence Interval	
Group	Mean	Std. Error	Lower Bound	Upper Bound
Control	1002.827[b]	2.120	998.670	1006.985
Treatment	1023.301[b]	2.277	1018.837	1027.765

a. Subject = Math

b. Covariates appearing in the model are evaluated at the following values: propensity = .4710, pre-test = 981.6104.

The confidence interval for pre- post-test gain for the treatment group is 95% C.I = [13.47, 27.48]

Table 4.
Adjusted Post-test within grade means of treatment versus control for Math

				95% Confidence Interval	
Group	Grade	Mean	Std. Error	Lower Bound	Upper Bound
Control	3	988.919[b]	4.256	980.573	997.265
	4	1006.194[b]	3.530	999.271	1013.117
	5	1007.331[b]	2.941	1001.564	1013.098
	6	1008.866[b]	4.558	999.928	1017.804
Treatment	3	1011.560[b]	4.021	1003.674	1019.446
	4	1016.297[b]	3.503	1009.428	1023.165
	5	1031.677[b]	3.776	1024.271	1039.083
	6	1033.670[b]	5.115	1023.638	1043.701

a. Subject = Math

b. Covariates appearing in the model are evaluated at the following values: propensity = .4710, pre-test = 981.6104.

Table 5.
Un-Adjusted Post-test within grade means of treatment versus control for Math

Group	Grade	Mean	Std. Deviation	N
Control	3	889.68	82.38	212
	4	966.08	99.86	269
	5	1054.24	97.02	466
	6	1097.15	97.42	227
	Total	1012.62	119.83	1174
Treatment	3	933.04	71.92	255
	4	999.15	95.90	347
	5	1066.60	95.34	268
	6	1086.89	114.72	123
	Total	1011.24	108.43	993
Total	3	913.35	79.75	467
	4	984.71	98.94	616
	5	1058.75	96.53	734
	6	1093.54	103.78	350
	Total	1011.99	114.72	2167

Reading

Next was the effect of treatment, grade, and treatment x grade interaction with the propensity score as a covariate. The dependent variable is the post-test reading achievement score after 12 weeks of instruction.

The propensity score was significant (F (1, 2228) = 18.25, $p < .01$; partial $\eta^2 = .008$), as was the pre-test covariate (F (1, 2228) = 3111.39, $p < .01$; partial $\eta^2 = .58$), treatment (F (1, 2228) = 47.061, $p < .01$; $\eta^2 = .21$), and grade (F (3, 2228) = 19.16, $p < .01$, $\eta^2 = .025$). Treatment by grade results did not meet statistical significance, indicating that treatment effect – the slope of the growth -- did not differ by grade.

The Cohen's d effect size equivalent for the treatment effect $\eta^2 = .021$, using methods described by Cohen (1988) is $d = .29$. It is considered a small but statistically significant effect.

Table 6.
Adjusted Post-test means of treatment versus control for Reading

			95% Confidence Interval	
Group	Mean	Std. Error	Lower Bound	Upper Bound
Control	1088.961[b]	2.523	1084.014	1093.908
Treatment	1119.565[b]	2.881	1113.915	1125.214

a. Subject = R

b. Covariates appearing in the model are evaluated at the following values: propensity = .4334, pre-test = 1076.1906.

The confidence interval for the difference between pre- and post-test means 95% C.I. = [21.85, 39.35].

Table 7.
Adjusted Post-test within-grade means of treatment versus control for Reading

				95% Confidence Interval	
Group	Grade	Mean	Std. Error	Lower Bound	Upper Bound
Control	3	1071.187[b]	4.709	1061.952	1080.422
	4	1099.380[b]	3.726	1092.072	1106.687
	5	1104.400[b]	3.543	1097.452	1111.347
	6	1080.878[b]	5.217	1070.647	1091.108
Treatment	3	1100.131[b]	4.890	1090.541	1109.722
	4	1123.766[b]	4.389	1115.159	1132.374
	5	1128.564[b]	4.665	1119.416	1137.713
	6	1125.796[b]	6.207	1113.624	1137.969

a. Subject = R

b. Covariates appearing in the model are evaluated at the following values: propensity = .4334, pre-test = 1076.1906.

Table 8.
Un-Adjusted Post-test within-grade means of treatment versus control for Reading

group	Grade	Mean	Std. Deviation	N
Control	3	1024.03	108.67	261
	4	1093.04	103.40	369
	5	1134.57	105.28	410
	6	1156.80	112.22	229
	Total	1103.77	116.28	1269
Treatment	3	1031.02	101.66	245
	4	1102.47	103.92	333
	5	1151.14	104.40	269
	6	1149.09	121.12	122
	Total	1103.79	115.68	969
Total	3	1027.41	105.29	506
	4	1097.51	103.68	702
	5	1141.14	105.17	679
	6	1154.12	115.28	351
	Total	1103.78	116.00	2238

ELA

Finally, there was the effect of treatment, grade, and treatment x grade interaction with the propensity score as a covariate. The dependent variable is the post-test ELA achievement score after 12 weeks of instruction.

The propensity score was not significant, but the pre-test covariate was significant (F (1, 1021) = 1731.07, $p < .01$; partial $\eta^2 = .63$), as well as treatment (F (1, 1021) = 31.92, $\eta^2 = .03$), and grade (F (3, 1021) = 3.43, $p < .01$, $\eta^2 = .01$). The interaction term was marginally significant (F (3, 1021) = 2.58, $p < .052$, $\eta^2 = .008$).

The Cohen's d effect size equivalent for the treatment effect $\eta^2 = .03$, using methods described by Cohen (1988) is $d = .35$.

Table 9.

Adjusted Post-test means of treatment versus control for ELA

group	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Control	1050.898[b]	3.638	1043.758	1058.038
Treatment	1080.930[b]	3.739	1073.592	1088.267

a. Subject = E

b. Covariates appearing in the model are evaluated at the following values: propensity = .3421, pre-test = 1045.6195.

The confidence interval for the difference between pre- and post-test means is 95% C.I. = [19.60, 40.46].

Table 10.

Adjusted Post-test within grade means of treatment versus control for ELA

Group	Grade	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Control	3	1048.914[b]	11.081	1027.169	1070.659
	4	1075.728[b]	8.926	1058.212	1093.244
	5	1035.342[b]	6.241	1023.095	1047.589
	6	1043.608[b]	6.100	1031.637	1055.579
Treatment	3	1072.758[b]	7.279	1058.475	1087.042
	4	1091.021[b]	6.843	1077.593	1104.449
	5	1085.293[b]	6.563	1072.415	1098.171
	6	1074.647[b]	6.242	1062.399	1086.895

a. Subject = E

b. Covariates appearing in the model are evaluated at the following values: propensity = .3421, pre-test = 1045.6195.

Table 11.

Un-Adjusted Post-test within grade means of treatment versus control for ELA

Group	Grade	Mean	Std. Deviation	N
Control	3	898.73	129.37	40
	4	1030.61	103.87	74
	5	1064.98	115.99	198
	6	1101.04	110.43	224
	Total	1062.90	124.50	536
Treatment	3	1003.36	95.08	85
	4	1046.02	102.33	188
	5	1104.26	90.08	101
	6	1099.90	107.69	121
	Total	1063.75	106.60	495
Total	3	969.88	117.43	125
	4	1041.67	102.80	262
	5	1078.25	109.38	299
	6	1100.64	109.32	345
	Total	1063.31	116.19	1031

Attrition Analysis

A potential threat to analysis of experimental data is the presence of differential attrition. While it is not always possible to eliminate threat completely, it is helpful to be cognizant of its presence and degree while interpreting results.

Tables 12 through 14 show the attrition rate for each of the three subjects. Each subject is analyzed separately because teachers occasionally opted not to administer specific tests in the second testing window. This is most likely to occur in the case of ELA, as teachers typically try to reduce testing time by using only the Math and Reading tests. The presumable idea is that Reading and ELA are to some extent redundant in their estimation.

Thus, the ELA test is the only one of the three that seems to have a significant differential in attrition rate. Clearly the control group was more likely to drop out of the second testing window. It

might be suggested that users of Study Island tend to be already committed to the value of testing, and thus the treatment group teachers are less likely to let the learners out of the second testing window.

Another possible explanation is that the control group in general had higher pre-test scores on the achievement test scales. Again, the control group teachers might have been more comfortable letting the ELA test drop in the second window because the learners do not exhibit cause for concern.

Still, it is important to be aware of the differential attrition in the ELA test. A follow up analysis showed that it was indeed the high performing learners who dropped out of the ELA test, thus giving an edge in favor of the treatment group mean.

Table 12.

ELA Attrition

			Attrition		
			NO	YES	Total
	Control	Count	884	171	1055
		% within group	83.8%	16.2%	100.0%
	Treatment	Count	566	16	582
		% within group	97.3%	2.7%	100.0%
Total		Count	1450	187	1637
		% within group	88.6%	11.4%	100.0%

Table 13.

Reading Attrition

		Attrition		
		NO	YES	Total
Control	Count	1403	120	1523
	% within group	92.1%	7.9%	100.0%
Treatment	Count	1051	38	1089
	% within group	96.5%	3.5%	100.0%
Total	Count	2454	158	2612
	% within group	94.0%	6.0%	100.0%

Table 14.

Math Attrition

		Attrition		
		NO	YES	Total
Control	Count	1538	75	1613
	% within group	95.4%	4.6%	100.0%
Treatment	Count	1055	76	1131
	% within group	93.3%	6.7%	100.0%
Total	Count	2593	151	2744
	% within group	94.5%	5.5%	100.0%

Discussion

The study measured the effect of 12 weeks of Study Island usage on achievement tests. The Cohen's *d* effect sizes of .25 - .35 suggest that up to a third of a standard deviation is gained in the treatment period. For purposes of comparison, it might be noted that a recent study by Scammacca, Fall, and Roberts (2015) computed national growth norms for a full academic year for learners in the lower part of the percentile range. They reported typical annual growth for grades 3-6 in the range of .30 to .55 in Math and Reading. This growth benchmark represents the 50th percentile. The authors used Hedges *g* statistic, which is extremely close to Cohen's *d* for large samples, and is therefore comparable for purposes of this discussion.

The finding of about a third of a standard deviation gain in this study, potentially attributable to educational technology, represents *clinically important growth*. It would appear to be comparable to 3/5ths of the *total* growth that occurs in an entire academic year. It must be appreciated that the Cohen's *d* statistics reported in the present study are incremental effects above and beyond a treatment as usual control group, not net growth due to all classroom learning combined which is what Scammacca et. al (2015) reported as a benchmark.

The findings of this study correspond very closely to the effects found for computer assisted instruction with frequent testing. There were small- to medium effects, all statistically significant, for all three subjects. The Cohen's *d* statistics are very much in line with what has been found in the literature discussed in the introduction. It should be observed that the treatment effect, while important, should be understood as the effect of about 10 minutes per week on the platform, which – as noted earlier – is below the recommended dosage. Other researchers such as Cheung and Slavin (2012, 2013) found that 30 minutes per week is the ideal dosage for the full treatment effect. Consequently, one might suggest that the effect sizes reported here are the lower bound of the potential effect size.

Nevertheless, the fact that the effect size magnitudes are closely aligned with previous research gives confidence in the results. Moreover, the present study found that the Study Island learning platform was more effective for younger learners. This is consistent with prior research suggesting that computer based individualized learning works best for those at the lower end of the ability scale.

Similarly, it is noteworthy that the Study Island platform was used in many schools as an intervention for disadvantaged students, and in schools perhaps with less funding – as suggested by the higher student teacher ratios. The gains found in this study are encouraging to schools who serve these populations.

References

Bloom, B. S. (1973). Recent developments in mastery learning. *Educational Psychologist*, *10*, 204-221.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record*, *64*, 723-733.

Carroll, J. B. (1989). The Carroll Model: A 25-Year Retrospective and Prospective View. *Educational Researcher, 18,* 1, 26-31.

Cheung, A. C. K., & Slavin, R. E. (2012). How features of educational technology applications affect student reading outcomes: A meta-analysis. *Educational Research Review, 7,* 3, 198-215.

Cheung, A. C. K., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. *Educational Research Review, 9,* 88-113.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences, 2nd ed., New Jersey: Lawrence Erlbaum Associates, Inc.

Guskey, T. R., & Gates, S. L. (1986). Synthesis of research on the effects of mastery learning in elementary and secondary classrooms. *Educational Leadership*, *43*(8), 73 – 80.

Jain, S. (2015). Improving Student Achievement through Mastery Learning. *Asian Journal of Multidisciplinary Studies*, *3*(4), 113-118.

Karpicke, J. D., & Roediger, H. L. (2008). The Critical Importance of Retrieval for Learning. *Science*, *319*, 5865, 966.

Kulik, C.C. & Kulik, J. (1986-87). Mastery Testing and Student Learning: A Meta-Analysis. *Journal of Educational Technology Systems*, 15, 325-345.

Kulik, C-L. C., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research, 60*(2)*,* 265–299. Retrieved from http://www.jstor.org/stable/1170612.

Kulik, C-L. C., Kulik, J. A., & Bangert-Drowns, R. L. (1991). Effects of frequent classroom testing. *Journal of Educational Research*, 85(2), 89-99.

McDermott, K. B., Agarwal, P. K., D'Antonio, L. D., Roediger III, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, *20*(1), 3–21.

Roediger III, H. L. (2014). How tests make us smarter. *The New York Times*. Pp SR12.

Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255.

Rosenbaum, P. R. & Rubin, D. B. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*, 33–38.

Scammacca, N. K., Fall, A.-M., & Roberts, G. (2015). Benchmarks for Expected Annual Academic Growth for Students in the Bottom Quartile of the Normative Distribution. *Journal of Research on Educational Effectiveness*, *8*, 3, 366-379.