# How big is that?

## Reporting the Effect Size and Cost of ASSISTments in the Maine Homework Efficacy Study

March 2017

SRI Education™
A DIVISION OF SRI INTERNATIONAL

## Authors

Jeremy Roschelle, Robert Murphy, Mingyu Feng, Marianne Bakia

This project report was produced as part of a research collaboration of SRI Education, Worcester Polytechnic Institute, and Center for Research and Evaluation (CRE) at the University of Maine.

This material is based on work supported by the Institute of Education Sciences of the U.S. Department of Education under Grant R305A120125. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Institute of Education Sciences.

## Suggested Citation

Roschelle, J., Murphy, R., Feng, M., & Bakia, M. (2017). *How big is that? Reporting the Effect Size and Cost of ASSISTments in the Maine Homework Efficacy Study.* Menlo Park, CA: SRI International.

# Contents

# Introduction

In a rigorous evaluation of ASSISTments as an online homework support conducted in the state of Maine, SRI International reported that "the intervention significantly increased student scores on an end-of-the-year standardized mathematics assessment as compared with a control group that continued with existing homework practices." (Roschelle, Feng, Murphy & Mason, 2016). Naturally, education stakeholders want to know how big the improvement was.

To answer this type of question, researchers report an effect size as a simple way of quantifying the difference between two groups. We reported an effect size of $g = 0.18$ of a standard deviation ($t(20) = 2.992$, $p = 0.007$) based on a two-level hierarchical linear model (Roschelle et al., 2016). An effect size is calculated by dividing the difference in scores between the two groups by the pooled standard deviation (Hedges, 1981). The underlying idea is that the strength of an effect depends both on the magnitude of the score difference and on how much the scores vary naturally. Consider this analogy to a commute: If it takes exactly 25 minutes to get to work every day, then a reduction to 22 minutes might mean a lot. Yet if the commute time varies between 10 minutes and 60 minutes, a reduction of the average time of 25 minutes to 22 minutes might not feel like much.

Roschelle and colleagues (2016) also reported an improvement index corresponding to the effect size: "Students at the 50th percentile without the intervention would improve to the 58th percentile if they received the ASSISTments treatment." An improvement index is the expected percentile gain for the average student in the control group—the student who scored at the 50th percentile on the outcome measure—if that student had attended a school where the intervention was implemented. Reporting the effect size or an improvement index does not appear to answer educators' questions completely, however. To an educator, the implications of whether such numbers are high or low may not be unclear.

In this technical report, we present alternatives for explaining the effect size, building on the guidance of Lipsey et al. (2012), a leading researcher who developed broad recommendations for effect size reporting. First, we provide additional detail on how we calculated the effect size and highlight the range of values that might be considered valid for this study. Second, we give comparisons with conventional benchmarks, a strategy that Lipsey and colleagues criticized but that still bears reporting. Third, we offer comparisons based on the recommendations of Lipsey et al. The report closes with a discussion of the challenges of interpreting effect sizes. The sidebar at the end of the report provides sample statements that educators may use to describe the study.

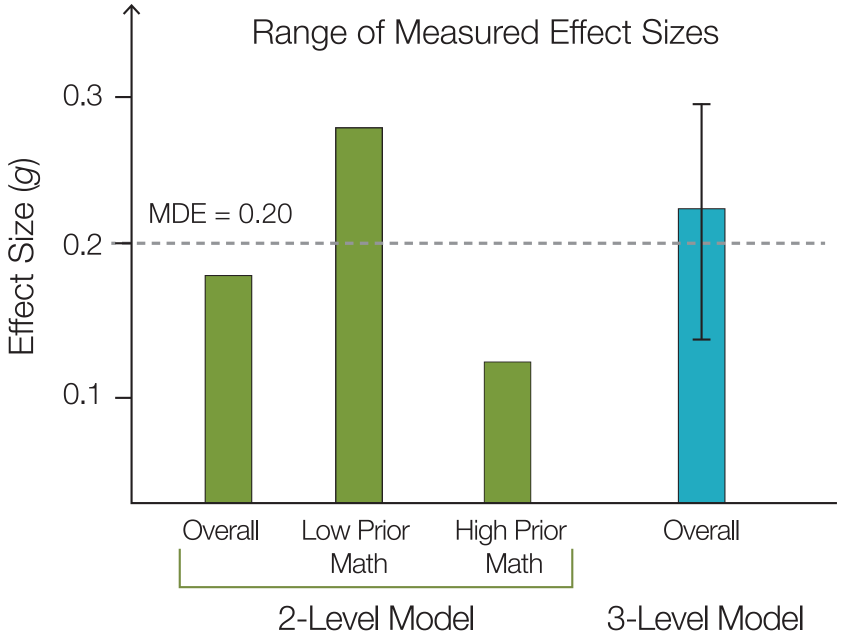
## The Range of Effect Sizes

In our first evaluation report (Roschelle et al., 2016), we not only reported the overall effect size of 0.18, but also considered how the effect size varied for students with lower or higher prior mathematics scores. Using the scores on the state standardized test from a school year before the ASSISTments intervention, we divided the student sample into two groups: (1) a group whose prior math score was at or below the overall median math score and (2) a group whose prior math score was above the median math score. We found that students with a lower prior mathematics score experienced a greater benefit from the ASSISTments intervention: Students with low prior mathematics scores gained 14.35 points on the TerraNova Common Core assessment, the primary outcome measure of the study, compared with 5.84 points for the students with a higher prior mathematics score. This interaction effect was statistically significant ($t(2770) = 2.432$, $p = 0.015$). TerraNova is a standardized test with established technical qualities, and the Common Core version aligns to the state of Maine's adoption of the Common Core State Standards. An IES review of the standardized tests compared assessments that are commonly used in the Mid-Atlantic region and gave TerraNova especially high marks, stating " only one was truly a predictive study and demonstrated strong evidence of predictive validity (TerraNova)." (Brown & Coughlin, 2007, p. iv)

We also calculated effect sizes for each group; the effect sizes were 0.29 for the low prior math score group and 0.12 for the high prior math score group. This suggests that educators who choose to use ASSISTments could see variability in the effects in their schools depending on whether their students have lower or higher prior performance in mathematics. Schools that want to close achievement gaps may be particularly interested in the effect size for students with lower prior math scores, which is notably larger than the effect size for all students.

We also continued with data analysis after publishing for first report (Roschelle et al., 2016). We noticed outliers in the data set (see Appendix A) and moved from the two-level HLM model reported earlier to a three-level model that more accurately reflects the structure of the data. The effect size recalculated using a three-level HLM on a data set that excludes the student outliers was $g = 0.22$, with a 95% confidence interval from 0.15 to 0.30. The confidence interval means that if we had the resources to run the experiment 100 times, we could expect to get a treatment effect size in the range of 0.15 to 0.30 95 times. Figure 1 shows the range of effect sizes we measured. For the remainder of the report, we focus on effect size and confidence interval estimated via the 3-level model, as we consider this to be the most meaningful estimate of the impact of ASSISTments on student achievement.

Figure 1: Range of effect sizes measured for impact on student mathematics achievement in the ASSISTments group as compared to a control group. The experiment was planned for a minimal detectable effect size (MDE) of 0.20. The three effect sizes to the left were reported in (Roschelle et al 2016). The 3-level model to the right (see Appendix A) excludes outliers and also produced a confidence interval around the estimated effect size.



## Comparison with Conventional Benchmarks

The most common way to consider the importance of an effect size is to compare it against conventional benchmarks. For example, ASSISTments arose from the intelligent tutoring systems tradition. In this tradition, researchers have long aimed for a "two sigma effect" (Bloom, 1984) that would be realized by increasing learner outcomes by two standard deviations. Bloom suggested an effect size benchmark of 2.0 based on the belief that providing a student with a human tutor has an effect of this magnitude. A more recent meta-analysis (VanLehn, 2011), however, found an average effect size in studies of human tutors of 0.79. For intelligent tutor systems that intervene when a student makes a wrong mathematical step, the average effect size was 0.75. For systems that intervene when students give wrong answers (but not when students err on individual steps during the process of getting the answer), the average effect size relative to conventional instruction was 0.31. Another caution with these benchmarks, however, is that they typically derive from experiments in which the researcher defined the outcome measure; effects are typically lower in experiments that use an externally validated measures, such as the TerraNova assessment used in our study. Indeed, some rigorous studies of intelligent tutoring systems in mathematics have found no effect on student outcomes (e.g., Dynarski et al, 2007). With these considerations in mind, we powered this experiment to be able to detect an effect of 0.20. The measured effect size of 0.22 was slightly higher than our expectations.

As deployed in the Maine Homework Efficacy Study, ASSISTments was not intended to be compared with a human tutor or to a step-oriented intelligent tutoring system; hence, using the 0.31 benchmark for comparison is reasonable.

Besides considering ASSISTments relative to intelligent tutoring systems, another reasonable benchmark would come from research on formative assessment interventions. ASSISTments is a formative assessment intervention because it emphasizes timely feedback to students and teachers and also supports teachers to make instructional decisions based on the feedback. In one formative assessment study that was similar to our evaluation (but did not use technology), 24 middle and high school math and science teachers developed their practices of formative assessment over the course of a year. At the end of the year, student achievement was measured by externally scored standardized tests. The investigators found an increase in student achievement, compared with students of other teachers in the same schools, of 0.32 standard deviations (Wiliam, Lee, Harrison, & Black, 2004). But that was only one study. A meta-analysis of 19 studies of formative assessment in mathematics found a mean effect size of 0.17, with a 95% confidence interval ranging from 0.14 to 0.20 (Kingston & Nash, 2011).

Another reasonable benchmark comes from a meta-analysis of computer-based interventions in mathematics, which focused specifically on high-quality research studies with features such as random assignment and use of an external standardized test to measure the student learning outcome. Cheung and Slavin (2013) found a mean effect size of 0.09.

Rather than using a particular class of interventions for reference, we could use the universe of all possible interventions. Cohen (1988) suggested a set of conventional benchmarks for educational interventions writ large.  The benchmarks describe “small” (0.2), “medium” (0.5), and “large” (0.7) effects in education. By this set of benchmarks, ASSISTments in our evaluation had a small effect. Small is not understood to be negligible or unimportant. For example, McCartney and Rosenthal (2000) reported that for interventions related to heart attacks,  the best fell well below the 0.2 benchmark for a small effect and yet some of the interventions “correspond to reducing the incidence of heart attacks by about half—an effect of enormous practical significance” (p. 4). If students were to experience an intervention in this range for multiple years of schooling, the compound effect could easily become dramatic.

Lipsey et al. (2012) discouraged the use of conventional benchmarks because the underlying factors in the target study and its comparison group are often not similar. They stated:

> The problem is that the normative distribution used as a basis for comparison must be appropriate for the outcome variables, interventions, and participant samples on which the effect size at issue is based. Cohen’s broad categories of small, medium, and large are clearly not tailored to the effects of intervention studies in education, much less any specific domain of education interventions, outcomes, and samples. Using those categories to characterize effect sizes from education studies, therefore, can be quite misleading. It is rather like characterizing a child’s height as small, medium, or large, not by reference to the distribution of values for children of similar age and gender, but by reference to a distribution for all vertebrate mammals. (p. 4)

In the sections that follow, we consider alternatives to conventional benchmarks.

## Comparison with Expected Progress

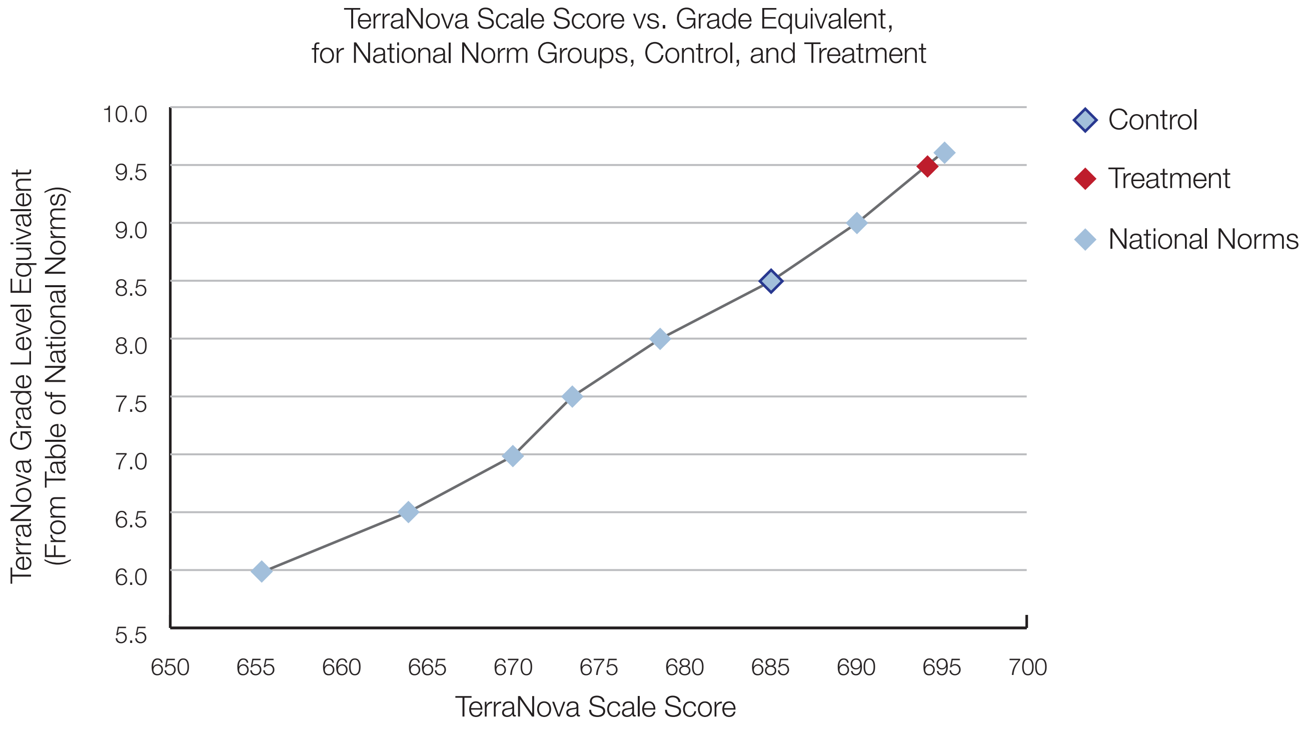
As an alternative, Lipsey et al. (2012) recommended comparing the measured difference between treatment and control groups in a study with expected progress in the same time period and on the same measure. We used the TerraNova Common Core math test as the outcome measure, and the publisher of this test provides extensive backup documentation (CTB/McGraw-Hill LLC, 2011) in addition to the higher reputation the TerraNova has achieved in external validity research (Brown & Coughlin, 2007). The backup document enables us to compare our measured impact on student learning to expected progress.

The TerraNova test is offered for students in grade levels from elementary school through high school, and the scores on each grade-level-appropriate assessment are translated to a uniform scale of 0 to 1,000. Because the scale covers the variation expected across 12 years of school, the differences in test scores from year to year can be small. The publisher provides a table of national norms that translate between particular scores and grade-level equivalents (CTB/McGraw-Hill LLC, 2011). Using this table, we found that TerraNova scores generally increase by 11.66 points each year from sixth grade to ninth grade (our study involved seventh-grade students). If we presume that the average student in our sample progress at the national rate of 11.66 points each year, then the difference in the mean scores of the treatment group (8.49 points, see Appendix A) over the course of 7th grade seems to be an important gain.

Figure 2 is a plot of the national norm data from the TerraNova publisher, showing the expected TerraNova scores and the grade level for students who achieve those scores. At the end of seventh grade, control students in our study had a mean score of 685 and treatment students had a mean score of 694 (using the 3-level HLM and excluding students with a score of 487). Using Table 49 on page 100 of (CTB/McGraw-Hill LLC, 2011), this places the students at grade equivalents of 8.5 and 9.5 respectively, a difference of 1 grade-equivalent (note that both groups are performing above grade-level expectations relative to the nationally normed sample).

We caution that this does not necessarily mean that a student in the treatment group learned everything he or she would learn in an additional full year of school or that the student could skip eighth grade. An interpretation of this finding is “the students in the Maine control group performed at the level we would expect from students in the national sample who are half way through 8th grade and are taking the 7th grade test, and students in the Maine treatment group performed at the level we would expect from students in the national sample who are halfway through 9th grade and are taking the 7th grade test.” The bottom line is that a 7th grade student who achieves at a 9th grade-equivalent on the TerraNova does not mean that student has mastered all the 8th grade math content in Maine. Nonetheless, Figure 2 does show that the difference in TerraNova scores between the treatment and control groups is substantial.

Figure 2: This plot of mean scale scores of the treatment and control groups relative to the TerraNova publisher's national norms for scale scores and grade-level equivalents shows how the intervention effect compares with expected progress.



Another measure of expected progress Lipsey et al. (2012) discussed is the effect size associated with 1 year of school by subject area. In math, the effect size of attending a school in seventh grade is reported to be 0.32. Against this number, our range of reported effect sizes also seems to be notable. For example, the 0.22 effect size corresponds to approximately two-thirds of a year of expected progress—and in the case of this experiment, that is an *additional* two-thirds of a year of expected progress in the ASSISTments group compared to the control group.

## Comparison with Policy-Relevant Performance Gaps

Another point of reference to assess the relative size of an effect is the comparison with student performance gaps on policy-relevant indicators like eligibility for free or reduced-price lunch (FRPL) (an indicator of poverty) and Individualized Education Program (IEP) status (an indicator of special education services for students). In this study, we found that students who are eligible for FRPL or have an IEP had lower scores than other students. For example, in the control group, students with FRPL status scored 4 points lower and students with IEP status scored 9 points lower than their peers. Thus, the 8.49-point difference in scores between the treatment and control groups is meaningful given the size of the performance gaps for these important subgroups. Indeed, our finding that low-performing students benefited more from ASSISTments is particularly relevant because it is a gap-closing finding. FRPL and IEP status correlates with lower prior mathematics scores, and the gain of 14.35 for treatment students with lower prior scores relative to similar students in the control group may make this intervention particularly relevant to schools that wish to close achievement gaps. (In forthcoming work, we expect to investigate and report on whether the ASSISTments intervention specifically closed gaps for students with IEPs and eligible for FRPL.)

## Comparison on the Basis of Cost

A third comparison Lipsey et al (2012) recommended is cost. Thus far, our team has conducted pilot work to estimate the resources associated with the implementation of ASSISTments. In a future study, our team wants to collect additional data to estimate per student costs to compare the relative cost-effectiveness of ASSISTments to conventional approaches to homework. We present the pilot work below.

To make good decisions about implementing particular instructional interventions, educators need to understand the likely costs. Cost is a practical consideration that can dramatically shape how knowledge about effective practices is translated into action in districts. Reliance on effectiveness alone may encourage adoption of interventions that are too expensive to sustain with fidelity (Bakia, Caspary, Wang, Dieterle, & Lee, 2011; Harris, 2009; Hollands et al., 2015). Analytic approaches that examine costs are relevant in the context of rising prices in education and decreasing educational budgets (Bowen 2013; Hollands et al., 2014). In addition to questions of impact and efficiency, policy-makers and administrations require information related to affordability in order to address basic questions like: will a new program or approach increase costs, and if so, by how much? While cost-analyses of educational interventions are not yet common, studies of cost and cost-effectiveness are readily available in the health and human services sector.

It is important to note that, from an economics perspective, "cost" is a term used to represent something conceptually different than "price." Cost refers to the value of resources, no matter who pays. Price, on the other hand, refers more specifically to the money paid for a particular resource. For example, the "price" to a district for an unfunded mandate would be zero, but the time teachers might invest in integrating new resources has value. Their time could have been spent on some other productive endeavor and thus has a cost. These costs can then be matched with associated estimates of impact in order to create cost-effectiveness ratios. These ratios can be used by decision-makers to evaluate the relative value of one alternative to another.

An emerging standard in educational cost analysis is the "ingredients approach," a straightforward method to systematically identify required resources associated with adoption of a program or intervention, regardless of the source of funding (Levin & McEwan, 2001). The ingredients approach includes collecting detailed information regarding the components of an intervention and its alternatives in order to understand the type and amount of resources required to achieve the desired impact. Since various educational alternatives often require common resources (like classroom space, technical infrastructure, or teacher time), analysts have focused more on a program's direct costs and changes in cost rather including estimates of the value of classroom space, technical infrastrastructure and overhead rates (Hollands & Bakir, 2015). With this approach, the costs associated with schooling generally, such as classrooms with adequate furniture and supplies, are not included in cost estimates. However, our pilot work is not yet to the stage that warrants computation of a cost effectiveness ratio, so we focus below only on the resources associated with the cost of implementing ASSISTments.

So, what resources supported the use of ASSISTments during this study?

**Hardware.** The ASSISTments system is web-based. In Maine, no specialized hardware or software was needed to implement ASSISTments beyond the computers already provided to students by the state. In a different location, it might be necessary to estimate a cost to provide hardware and support to students if this is not already available.

**Teacher's Time.** School adoption of ASSISTments for this study also required administrative planning and training for teachers. Direct district resources supporting the implementation of ASSISTments included preparation for the use of ASSISTments:

- modest annual technology support for account set up of about 30 minutes per class (1.5 hours per teacher in Maine)
- about an hour of instructional time for teachers to introduce ASSISTments to students

Teachers also participated in about three days per year of professional development consisting of 2.5 days of in-person training and three hours via webinars. The ASSISTments team prepared and conducted the training sessions.

Available data from site visits suggest that the average amount of class preparation time teachers required did not change with the adoption of ASSISTments. Whereas teachers may save time grading student homework problems when using ASSISTments, they tended to use this time to review data reports generated by ASSISTments.

**Coaching.** ASSISTments also provided coaching and feedback for teachers during the course of the school year. A math coach traveled to participating schools. The coach visited each teacher on average two times in the course of a school year.  During the visit, the coach supported the teacher using ASSISTments during one class period time and, if depending on the teacher's availability, spent additional time with the teacher to answer his/her questions and review use of ASSISTments. Each session is estimated at 1.5 hours of teacher time per teacher, in addition to the coaches' time preparing to visit schools and conducting coaching with teachers in schools.

Future studies will examine the intensity of coaching and professional development required to sustain the intervention over time. Although we evaluated the program with two years of professional development, the program developer expects that a satisfactory implementation might be achieved with only one year of professional development and expects that the intervention could be sustained beyond two years without any additional professional development. Thus, teachers may need less professional development and coaching, so schools may realize lower costs.

We understand that some readers may wish to compare observed cost for ASSISTments with a benchmark range of costs associated with alternative math interventions and conventional approaches. However, as mentioned above, technically sound cost analyses for educational interventions generally and technology-supported interventions specifically are relatively sparse in the literature, making benchmarking difficult. Cost data on educational interventions are not readily available in the literature, but additional examples of studies and related resources are available at http://cbcse.org/publications/. If district or school leaders desire cost-effectiveness comparisons, we would advocate for each district or school making its own comparisons relative to the cost of the specific products it is considering, based on the ingredients list above.

## Discussion

Clearly, simply reporting an effect size does not satisfy educator's desires for a simple answer to the question "how big is that?" or "is the effect important?" We reported an average effect size of 0.22 with a 95% confidence interval of 0.15 to 0.30. We found an even larger effect size (0.29) for students with low prior mathematics achievement. We considered appropriate conventional benchmarks and also went beyond those by putting this effect size in the context of expected progress and in the context of achievement gaps found among low-income students and students with identified special education needs.  We also estimated the cost of the ASSISTments intervention. Each of these comparisons yields information that we hope is helpful to administrators and educators who are considering whether to use ASSISTments or a similar intervention.

In making sense of effect sizes, readers should be aware of the importance of comparing the designs of the studies being compared. The range of expected effect sizes can vary greatly by the scale and rigor of the studies included in the analysis. Generally, reported effect sizes tend to be higher in studies that involve smaller populations and less rigorous designs. Reported effect sizes are also higher in studies that use an investigator-designed assessment. This evaluation was relative large (43 schools and 85 teachers), followed a rigorous design, and used a nationally normed standardized test. It could be unfair to compare it with benchmarks derived from studies with just a few schools or using quasi-experimental and less rigorous designs.

Some further considerations for interpretation were raised in our journal article (Roschelle et al., 2016) and deserve repeating. This study was conducted in Maine, and its population is different from other areas of the United States; it is more rural and less racially diverse, for instance. Maine gives all seventh-grade students a laptop computer to take home, and the effects might be different in a location with less access or less equitable access to technology. Also, our study duration was fairly long: Teachers had a first school year to improve their practice, and then effects on students were measured in a second school year; effects might vary in implementations that are shorter or have different amounts or quality of teacher professional development and coaching.

## Conclusion

In this report, we expanded our reporting of the effect size of ASSISTments in the Maine Homework Efficacy Study.  Building on arguments made by others in the research literature, we believe that comparing the ASSISTments effect with conventional aspirations like the two sigma effect or Cohen's small, medium, and large categories may underplay the practical value of our research findings to educators. Comparing the results instead with expected progress or policy-relevant performance gaps is more appropriate. Such comparisons highlight the impact of ASSISTments and the potential value of working to improve mathematics homework practices using online tools.

Our overall recommendation to schools would be to consider which of the ways of reporting effect size best suits their local situation. The gap-closing effect noted and the comparison based on policy-relevant indicators may be most relevant for some schools. For others, the change in grade-level equivalents on TerraNova may hit home most meaningfully. Other schools may have the opportunity to compare ASSISTments with other invention choices on the basis of cost or effect size and may decide on the relevance of this study to their decisions with cost as a factor. In any event, schools should also consider how their setting differs from the setting of this study in terms of student population, access to technology, or availability of time and resources for teacher learning.

Reasonable summary statements that an educator could use are listed in the sidebar below along with important cautions. Overall, the state of the art in interpreting education research is such that the best practice for educational decision makers is not to focus overly on the magnitude of a single number, single best comparison, or the results from a single study. By considering the more complete set of interpretative guidelines suggested here and the sample summary statements in the sidebar, educators may come to their own most accurate and relevant understanding of "how big is that?" and their own understanding of the importance of the impact on student achievement measured in this study.

## References

Bakia, M., Caspary, K., Wang, H., Dieterle, E., & Lee, A. (2011). *Estimating the effects of online learning for secondary school students: State and district case studies.* Menlo Park, CA: SRI International.

Bloom, B. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4–16.

Bowen, W. (2013). *Higher education in the digital age.* Princeton, NJ: Princeton University Press.

Brown, R. S., & E. Coughlin. (2007). *The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region* (Issues & Answers Report, REL 2007–No. 017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory MidAtlantic.

Cheung, A. C., & 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.). Hillsdale, NJ: Lawrence Earlbaum.

CTB/McGraw-Hill LLC. (2011). *Norms book, spring*. (TerraNova test norms).

Dynarski, M., R. Agodini, Heaviside, S., Novak, T., Carey, N., Campuzano, L., Means, B., Murphy, R., Penuel, W., Javitz, H., Emery, D. & Sussex, W. (2007). *Effectiveness of reading and mathematics software products: Findings from the first student cohort.* Washington, DC. U.S. Department of Education, Institute of Educational Sciences.

Hedges, L. V. (1981). Distribution theory for Glass' estimator of effect size and related estimators. *Journal of Educational Statistics. 6*(2), 107–128. doi:10.3102/10769986006002107

Hollands, F.M. & Bakir, I. (2015). *Efficiency of automated detectors of learner engagement and affect compared with traditional observation methods.* Center for Benefit-Cost Studies of Education, Teachers College, Columbia University.

Hollands, F.M., Brooks, A., Belfield, C., Levin, H.M., Chang, H., Shand, R., Pan, Y.I., & Hanisch-Cerda, B. (2014). Cost-Effectiveness Analysis in Practice: Interventions to Improve High School Completion. *Educational Evaluation and Policy Analysis, vol. 36,* 3: pp. 307-326.

Hollands, F. M., Hanisch-Cerda, B., Levin, H. M., Belfield, C. R., Menon, A., Shand, R., Pan, Y., Bakir, I., & Cheng, H. (2015). *CostOut — The CBCSE Cost Tool Kit.* Center for Benefit-Cost Studies of Education, Teachers College, Columbia University.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28–37.

Harris, D. (2009). Toward policy-relevant benchmarks for interpreting effect sizes: Combining effects with costs. *Educational Evaluation and Policy Analysis. 31*(1), 3- 29.

Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis: Methods and applications* (2nd ed.). Thousand Oaks, CA: SAGE.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms* (Report No. 2013-3000). Washington, DC: National Center for Special Education Research.

McCartney, K., & Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. *Child Development, 71,* 173–180.

Roschelle, J., Feng, M., Murphy, R., & Mason, C. (2016). Online mathematics homework increases student achievement. *AERA Open, 2*(4). doi: 10.1177/2332858416673968

VanLehn, K. (2011): The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221.

Wiliam, D., Lee, C., Harrison, C., & Black, P. J. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education: Principles, Policy & Practice, 11*(1), 49–65.

## Summary Statement About This Study

The Maine Homework Efficacy Study compared test scores for seventh-grade students in schools that used ASSISTments for homework assignment, completion, and review with test scores for students in schools that continued with their existing homework practices. This comparison occurred after teachers had a full school year in which to learn to use ASSISTments with teacher training and coaching. The cost per teacher of the ASSISTments intervention consisted mostly of costs associated with the teacher training and coaching.

Students in schools that used ASSISTments learned more. The mean effect size associated with the use of ASSISTments as compared to a business as usual control group  of +0.22 standard deviations was

- Greater than the 0.09 effect size found across rigorous studies of computer-based interventions in mathematics
- Slightly greater than the 0.17 effect size found in rigorous studies of formative assessment in mathematics
- Slightly greater than the 0.20 effect size we planned the experiment to be able to detect
- About 2/3 of the 0.32 effect size expected for a full additional year of classroom instruction, suggesting the amount of additional learning in the ASSISTments group is important.

Students who had lower mathematics scores before seventh grade benefited more from ASSISTments than students who had higher mathematics scores before seventh grade, a gap-closing effect.

The score gain for students in schools that used ASSISTments also compared favorably with policy-relevant performance gaps such as the gap related to students' eligibility for free or reduced-price lunch or the gap related to student IEP status.

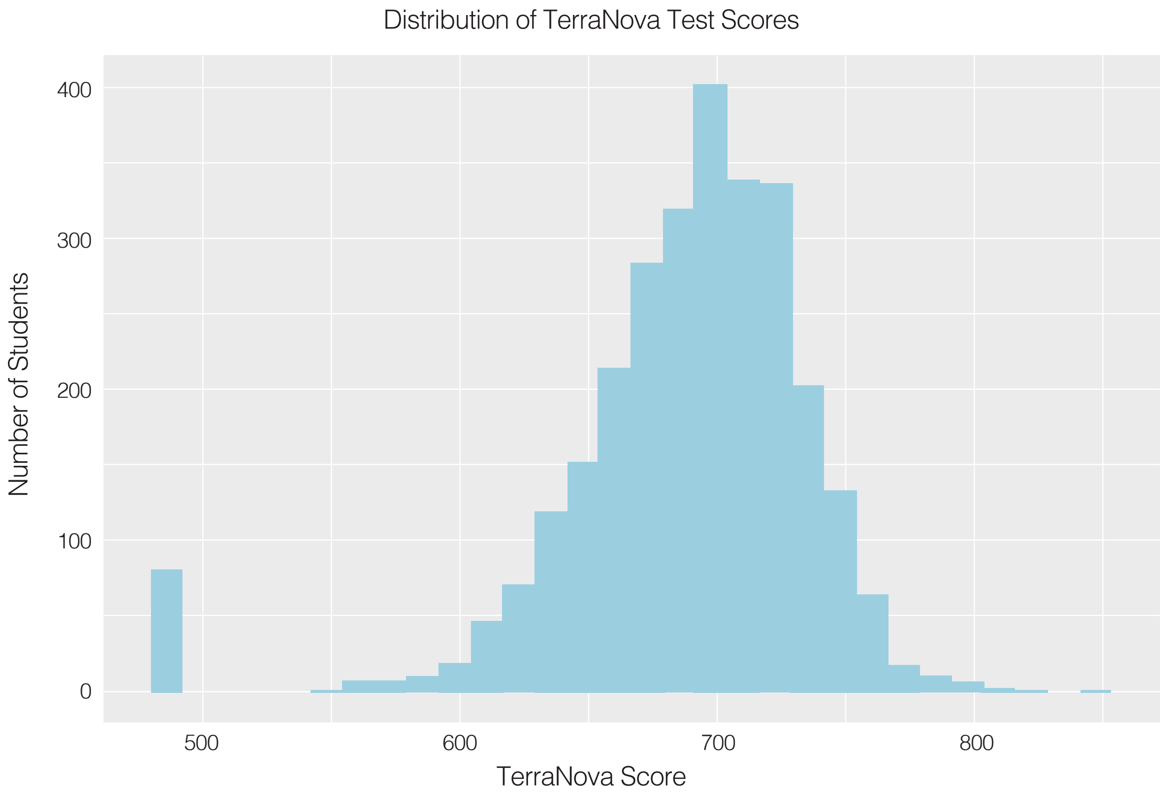
In interpreting the effect size measured in the Maine Homework Efficacy Study, educators should be careful to consider how their school setting may be different from the setting of this study. Important differences include access to technology, demographic differences, and availability of sufficient time and resources for teachers to learn to use ASSISTments.  In interpreting the effect size relative to other studies, educators should be aware that effect sizes vary with the quality of the research design so it is important to compare equally rigorous and large-scale studies.

## Appendix A: Outliers and the Three Level Model

In looking at the distribution of student scores on the TerraNova, we noticed outliers in the data set. A group of 81 students received a score of 487 on the TerraNova. A histogram of student scores showed a bimodal pattern (Figure 3). The students who had scores of 487 formed a second peak in the distribution, and this score was about 60 points less than the next nearest score of 547. We contacted the test publisher about this score and learned the following:

> the approach used to score these students cannot produce scale score estimates for examinees with scores below the level expected from guessing or chance. In addition, the estimates that are available for examinees with extremely low scores, may have estimates with larger conditional standard errors of measurement, and thus larger gaps in scale scores, with differences between these extreme values having little meaning. Therefore, these low scores are established for these examinees based on a specific set of rules. These values, which are set separately by level, are called the Lowest Obtainable Scale Score (LOSS). (personal communication,  July 11, 2016)

**Figure 3: A histogram reveals a bimodal distribution of scores, with a large number of students having the unusually low score of 487.**



Basically, the score of 487 was assigned to students who turned in blank papers or did very little work on the assessment. In this study, students were volunteers and could choose not to participate at any time. Thus, it is reasonable to consider these low-scoring students as nonparticipants. This does not change the overall statistical significance of the study. It does change the denominator in the effect size calculation because removing these students decreases the standard deviation.

We also introduced a three level HLM model to analyze the data and we will report in full this model in a publication that is currently in preparation. In short, the three levels in this model are the school level, the classroom level, and the student level. Teachers in the study sometimes taught multiple classrooms, and classroom cohorts may differ within the same teacher, which is why we used "classroom" as the middle level.

Other variables in the three-level HLM model include:

- Mean math scores at the school level
- Mean free and reduced price lunch status at the school level
- Number of students enrolled in 7th grade at the school level
- Mean math scores at the classroom level
- Number of students in the classroom
- Student prior math scores
- Student gender
- Student individualized education plan status
- Student free and reduced priced lunch status

Controlling for these student, classroom, and school-level covariates, this model found a significant treatment effect for those students who used ASSISTments ($\gamma 001=8.492$, $t(18)=465.096$, $p<0.001$). This corresponded to an effect size (Hedge's *g*) of 0.22. The 95% confidence interval of the effect size is [0.15, 0.30]. The mean TerraNova score for the control group in this model is 685.230 (which we round to 685 for grade level equivalent comparisons) and the mean score is 693.731 in the the control group (which we round to 694).

## SRI Education

SRI Education, a division of SRI International, is tackling the most complex issues in education to identify trends, understand outcomes, and guide policy and practice. We work with federal and state agencies, school districts, foundations, nonprofit organizations, and businesses to provide research-based solutions to challenges posed by rapid social, technological and economic change. SRI International is a nonprofit research institute whose innovations have created new industries, extraordinary marketplace value, and lasting benefits to society.

**Silicon Valley**
(SRI International Headquarters
333 Ravenswood Avenue
Menlo Park, CA 94025
+1.650.859.2000
education@sri.com

**Washington, D.C.**
1100 Wilson Boulevard, Suite 2800
Arlington, VA 22209
+1.703.524.2053

***www.sri.com/education***

**STAY CONNECTED**