THE UNIVERSITY OF MEMPHIS | The Center for Research in Educational Policy (CREP)

The LASER Model: A Systemic and Sustainable Approach for Achieving High Standards in Science Education

SSEC i3 Validation Final Report of Confirmatory and Exploratory Analyses

Todd Zoblotsky, Ed.D.

Christine Bertz, Ph.D.

Brenda Gallagher, Ed.D.

Marty Alberg, Ph.D.
Principal Investigator

The University of Memphis

8/31/2016

Acknowledgments

The success of this evaluation would not have been possible without the herculean efforts built on strong partnerships among the Center for Research in Educational Policy (CREP), the Smithsonian Science Education Center (SSEC), Abt Associates, Bernalillo Public Schools, Chama Public Schools, Cleveland County Schools, Greene County Schools, Houston Independent School District, Jemez Valley Public Schools, Johnston County Schools, Los Alamos Public Schools, McDowell County Schools, Moore County Schools, Mora Public Schools, Pecos Independent School District, Rio Rancho Public Schools, Santa Fe Public Schools, Warren County Schools, and Wilson County Schools. We extend our heartfelt thanks and appreciation to all who contributed to this amazing endeavor, and sought – and still seek – to improve the state of science education in America.

CREP Project Staff:

Marty Alberg	Principal Investigator
Carolyn Kaldon	Co-Principal Investigator
Dan Strahl	Co-Principal Investigator
Michael Rowe	Project Manager
John Burgette	Qualitative Analysis
Todd Zoblotsky	Statistics
Brenda Gallagher	Statistics
Yun Tang	Statistics
Lou Franceschini	Statistics
Haixia Qian	Statistics
Bryan Winter	Statistics
Ying Huang	Statistics
Adrian Young	School Liaison
Cindy Muzzi	School Liaison
Dallas Burkhardt	Site Researcher Liaison
Margie Stevens	SMS Administration
Ruby Booth	SMS Administration

Table of Contents

Introduction

In August 2010, the Smithsonian Science Education Center (SSEC), a division of the Smithsonian Institution formerly known as the National Science Resources Center (NSRC), received a grant of more than $25 million from the U.S. Department of Education's Investing in Innovation (i3) program for a five-year study to validate its Leadership and Assistance for Science Education Reform (LASER) model in three regions of the United States: rural North Carolina, northern New Mexico, and the Houston Independent School District (HISD). Matching funds to support the study in the amount of more than $5 million were obtained from partners in the three regions as required by the Department of Education.

The independent third-party research evaluation of the LASER model was conducted by the Center for Research in Educational Policy (CREP) with technical assistance from Westat and Abt Associates, who were provided to i3 grantees' evaluation partners by the US Department of Education (USDOE). CREP, a Tennessee Center of Excellence, is a research and evaluation unit based at the College of Education at the University of Memphis.

Interim evaluation results were reported to SSEC by CREP in annual formal technical reports as well as more informally through presentations and written materials. In July 2015, a comprehensive report (The LASER Model: A Systemic and Sustainable Approach for Achieving High Standards in Science Education Summative Report) was submitted to SSEC containing overall findings, conclusions, and recommendations in summary form based on analysis of the final year of available quantitative and qualitative data for students in two cohorts of schools: an elementary cohort and a middle school cohort. Supporting materials comprising the complete final report included an overview of implementation findings related to the five pillars of the LASER model and a report of findings from case studies in addition to quantitative analyses of achievement data related to both confirmatory and exploratory research questions.

The current report focuses on the confirmatory and exploratory research questions submitted to i3 for the two studies conducted for the LASER i3 validation grant, providing clarifying detail related to methodology and instrumentation. The studies were conducted to answer two confirmatory research questions:

*After three years of participation in the study (i.e., after Year 3), do schools containing the Grade 3 elementary school cohort that receive the LASER intervention (i.e., **Phase 1 schools**) attain higher levels of science achievement than schools that do not receive this intervention (i.e., **Phase 2 schools**) as measured by the PASS?*

*After three years of participation in the study (i.e., after Year 3), do schools containing the Grade 6 middle school cohort that receive the LASER intervention (i.e., **Phase 1 schools**) attain higher levels of science achievement than schools that do not receive this intervention (i.e., **Phase 2 schools**) as measured by the PASS?*

In addition, the studies were conducted to answer two exploratory research questions:

After three years of participation in the study (i.e., after Year 3), do schools containing the Grade 3 elementary school cohort that receive the LASER intervention (i.e., ***Phase 1 schools****) attain higher levels of science achievement than schools that do not receive this intervention (i.e.,* ***Phase 2 schools****) as measured by the PASS for the following underrepresented students in STEM?*

a. *Students with Disabilities*

b. *English Language Learners*

c. *Economically Disadvantaged*

d. *Females*

After three years of participation in the study (i.e., after Year 3), do schools containing the Grade 6 middle school cohort that receive the LASER intervention (i.e., ***Phase 1 schools****) attain higher levels of science achievement than schools that do not receive this intervention (i.e.,* ***Phase 2 schools****) as measured by the PASS for the following underrepresented students in STEM?*

a. *Students with Disabilities*

b. *English Language Learners*

c. *Economically Disadvantaged*

d. *Females*

Two new versions of the *What Work Clearinghouse Procedures and Standards Handbook* (Version 2.1 and Version 3.0) were published after the research design of the study was approved under i3 and data collection had begun, in addition to the *Reviewer Guidance for Use with the Procedures and Standards Handbook Version 3.0, which* was published in March 2016, subsequent to completion of this summative report submitted to SSEC in July 2015. The current report contains language intended to clarify findings and ensure alignment with the most recent version of the handbook (Version 3.0) and the *Reviewer Guidance* document. For example, issues such as level of inference (individual or cluster) and joiners vs. stayers were not part of the WWC Handbook at the time this study was first implemented under i3. As a result, considerations such as these were not factored into the original presentation of results. The current version of the report reflects language that the WWC would evaluate as part of their review of the study findings under the most current version of the handbook.

Note: As previously indicated, data for the elementary and middle school samples reported in this manuscript should be treated as two separate studies for reporting and interpretation purposes--an elementary school study and a middle school study (as confirmed by the WWC in the letter in ***Figure A-1***), both with cluster-level research questions and inference, but with analysis at the individual (i.e., student) level. Subgroup analyses focused on under-represented groups in STEM: Students with a Disability (IEP), English Language Learners (ELL), Economically Disadvantaged (FRL), and females.

Methodology

The LASER i3 Validation study utilized a matched-pair, randomized controlled trial (RCT) and was designed to meet the What Works Clearinghouse (WWC) criteria "without reservations," which is the highest possible rating. Schools with intact elementary (grades 3-5) and middle school (grades 6-8) cohorts were paired and randomly assigned to Phase 1 (immediate implementation) or Phase 2 (delayed implementation). Schools in Phase 1 began implementing LASER in the fall of 2011; Phase 2 schools served as the control group, receiving a reduced version of LASER following the conclusion of the research study. The matched-pair design was utilized to ensure equivalency between groups.

Baseline equivalence was established with the analytic sample for elementary schools overall and for all subgroups and in all but two cases for middle schools (PASS multiple choice and open-ended for the ELL subgroup, both of which favored Phase 1 schools). For the middle school study, all sixth graders in the impact analyses attended elementary schools the year prior to the inception of this study and were therefore not in their middle schools at the time of random assignment of clusters. By WWC definition, they were considered joiners (students not enrolled in the school at the time of random assignment). There were also joiners included in the elementary study (students whose first year in the study school/cluster was the third grade), although this number was relatively small (17.9% of the Phase 1 and 9.9% of the Phase 2 sample, 14.5% of the total sample). Please also note that while the word "student" is used throughout this manuscript in reporting some incidental findings (e.g., percentile ranks) due to the individual level analyses, and while tables of sample sizes and outcomes reference samples at both the school and individual levels due to the individual level analyses, inferences for both the elementary and middle school results should be at the cluster (i.e., school) level as whole schools, not individual students, received the intervention.

Overall and differential cluster-level attrition levels were calculated for the full elementary and middle school samples as well as for each subgroup for comparison to the What Works Clearinghouse (WWC) cluster-level attrition standards (See Table A-3,Table A-4, and Table A-5 in the Appendix), and analyses of aggregate and subgroup data included ANCOVAs with a cluster correction (See Table A-1 in the Appendix) as well as the Benjamini-Hochberg correction for multiple comparisons (See Table A-2 in the Appendix) the WWC applies to primary (i.e., confirmatory) and secondary (i.e., exploratory) contrasts (What Works Clearinghouse, 2014 and 2016). Details are provided in subsequent sections of this report and in the appendices.

Instrumentation

The evaluation team elected to use the WestEd-developed Partnership for Standards-based Science Assessment (PASS at WestEd$^{®}$) test as the primary measure of student learning. PASS is a standards-based test, developed with funding from the National Science Foundation (NSF). The work of PASS builds upon research on the properties of science assessment and current approaches for assessment development and scoring. For the purposes of this study, PASS was administered at the elementary and middle school levels. It consists of three assessment components at each grade level: 1) selected-response/multiple choice items (hereafter referred to in this report as multiple choice or MC), 2) constructed response investigations (grades 3-5) or open-ended questions (grades 6-8 (hereafter referred to in this report as open-ended or OE), and 3) hands-on performance tasks (PT). All students in the study completed the MC component; the OE and PT sections were completed by a sub-group of students in focal schools. Table 1 provides a description of each PASS Assessment Component.

Table 1: PASS Assessment Components

PASS Assessment Component	Description
Selected Response or Multiple Choice Items (MC) (29 items for both Elementary and Middle school)	Items assess students' understanding of important scientific facts, concepts, principles, laws, and theories.
Constructed Response Investigations and Open-Ended Questions (OE) (2 items for Elementary and 6 items for Middle school)	Students analyze a problem, think critically, conduct a secondary analysis, and apply learning. They construct explanations using evidence.
Hands-on Performance Tasks (PT) (6 items for both Elementary and Middle school)	Investigations identifying a problem to solve. Students use equipment to perform investigations; make observations; generate, organize, and analyze data; communicate understandings; and apply learning.

Measurement specialists on the PASS development teams conducted equating studies in the design of the forms and established the validity and reliability of the assessments. Table 2 below, obtained from WestEd, shows score reliabilities and inter-rater agreement calculated for the PASS assessments, including selected response/multiple choice items (MC), constructed response/open-ended investigations (OE), and hands-on performance tasks (PT). The reliabilities of the item calibrations are given by the Rasch equivalent of the Cronbach alpha statistic and are derived from the ratio of the spread of the items over the scale to their own root mean squared error (RMSE). The statistic is scaled to stretch from 0 to 1.0. Inter-rater agreement is the correlation between the first and second reader on each answer within a task.

Table 2: PASS Score Reliabilities and Inter-Rater Reliability

	Number of Students	Overall Score Reliabilities (Cronbach Alpha)	Inter-Rater Agreement (Constructed Response & Performance Tasks)		
			CR-1	CR-2	PT
Grade 5: Administered at grades 3, 4, & 5	7, 429	.87	.84	.87	.85
Grade 8: Administered at grades 6, 7, & 8	7, 777	.92	.95	.88	.91

To ensure that only students present at the beginning of LASER implementation would continue to be tested throughout the course of the study (and would therefore be considered as accurately representing their school in terms of science achievement), CREP researchers pre-slugged answer sheets following the baseline administration of the PASS MC component. Data from each administration were carefully monitored to ensure accuracy of the data set.

In both the elementary and middle school cohorts, there were four alternate forms of the PASS MC subtest administered over the course of the study: Fall 2011 (pretest), Spring 2012 (posttest), Spring 2013 (posttest), and Spring 2014 (posttest). The three posttest forms were equated, separately by grade level band, with the Fall 2011 (base) form using common items and the Mean/Sigma method based on single parameter (item difficulty) IRT Rasch models estimated under the assumption of random items. As the OE and PT subtests for both grade band tests were the same on all test forms, no equating was required for those subtests. It should be noted that the OE and PT sections were administered beginning in Spring 2012, so there were only three vs. the four administrations of the MC section. In addition to the equating, scaled scores (0-600 scale) were established for the MC subtest to enable analyses of outcomes across the different test forms.

CREP recruited, trained and calibrated scorers for the PASS OE and PT components of the assessment using WestEd's validated training and calibration materials and conducted an adaptation of WestEd's training and calibration process starting with the Spring 2013 administration. Scorers first participated in a five-hour training session, focused on either elementary or middle school scoring, where the open-ended questions and performance tasks, along with their associated scoring rubrics, were presented and discussed. At the end of the training session and prior to scoring any of the actual PASS OE/PT student materials, participants independently scored a calibration set of eight OE/PT calibration booklets (validated and provided from WestEd). The participants' results were compared to the 80% percent agreement benchmark set to ensure their readiness to become a certified scorer.

After qualifying as certified scorers, each was randomly assigned to score student test booklets, based on the elementary or middle school training they had received. Each time a scorer received a set of 100 booklets, three unidentified calibration booklets were randomly inserted into the set. When scorers returned each set of papers, CREP checked the calibration papers against the pre-determined scores. Scorers could not receive a new set of booklets unless they met or exceeded the benchmark calibration level.

Certified scorers signed a non-disclosure agreement and kept booklets secure and confidential until they were scored and returned. They scored each question individually and recorded results on student scan documents. The same scorers returned to score the following year's PASS OE/PT materials, and completed a second training and calibration process prior to scoring those assessments.

Population and Context

The population from which the sample for this study was drawn encompasses three regions (Houston Independent School District, rural North Carolina, and northern New Mexico), and represents a total of 16 school districts. This population of school districts includes more than 325,000 students and 20,000 teachers, and over 150 district and building level instructional leaders, with more than half of students (56.2%) identified as "economically disadvantaged" by the National Center for Educational Statistics (NCES) based on free and reduced lunch status. From this total, schools were nominated for participation in the study, and from that nominated list the study sample was created.

Although these regions, and schools within the regions, are very diverse, commonalities exist across the regions that stem from conversations, trends and initiatives taking place at the national level. These have had varying levels of impact on the teaching of science in elementary and middle schools, and therefore on the implementation of the LASER model during the course of this study. Most notable are the national debate around Common Core State Standards and associated testing; the Next Generation Science Standards (final draft released in April 2013); new teacher evaluation models; and identification and implementation of programs for low performing schools.

Within each region, unique conditions also existed during the LASER implementation window that had potential to impact implementation of the model. In North Carolina, a program called Read to Achieve was adopted in July 2012 via a state budget act and became effective during the 2013-2014 school year. With mandated summer reading camps and possible retention imposed on students who did not achieve acceptable reading levels by third grade, more instructional time was devoted to the teaching of reading in the lower grades in many NC schools. In Houston, 11 of the LASER i3 schools (25.6%) were also part of that district's Apollo 20 initiative, with multiple strategies in place to close gaps in achievement and more time devoted to testing than was required in other schools. Implementation in Northern New Mexico was impacted by the geography of the region: remote mountain school locations affected teacher travel to professional development sessions; delivery of materials, particularly live specimens; and access to SSEC's regional support system. Partnership with the Los Alamos National Laboratory (LANL) Foundation provided support for schools in that area.

Baseline data collected from teachers prior to LASER i3 implementation in fall 2011 revealed that most students in all three regions received science instruction from their classroom teachers rather than from a science specialist (reported by more than 90% of teachers in all three regions) and that these teachers did not major in science or science education (Approximately 10% of respondents from New Mexico and HISD and 8% from North Carolina reported holding majors in science). Science laboratories were more prevalent in Houston schools than in the other regions, with nearly half (49.8%) of teachers reporting that their students went to labs to receive science instruction compared to 10.5% in North Carolina and 1.5% in New Mexico. Time devoted to the teaching of science was also greater in Houston, with teachers reporting an average of 3.3 hours per week of science instruction compared to slightly under 2.5 hours in North Carolina and slightly over 2 hours in New Mexico.

When asked at baseline about challenges associated with teaching science in elementary and middle schools, teachers' responses were consistent across all three regions. The greatest challenge (reported as "substantial" or "significant" by 70% of NM teachers, 63% of NC teachers, and 59% of HISD teachers) was limited time for science instruction. Limited funds for purchasing equipment and supplies and more emphasis on English/language arts and mathematics than science instruction were also reported challenges in all regions.

Sample

The study sample originally included 139 schools across the three regions. These schools went through a nomination and qualification process, were matched based on several school-level demographic and achievement variables, and then randomly assigned to Phase 1 (immediate implementation) or Phase 2 (delayed implementation). After random assignment, changes in school participation occurred within each region, and the final sample contained 125 study schools within the 16 districts and encompassed approximately 60,000 students, 1,900 teachers, and over 140 district administrators and principals. While LASER is a school-level intervention in which all students in the participating schools received the treatment, a subsample of 9,000 students in two cohorts (third and sixth graders in 2011-12) was followed longitudinally over the three years of the study, and a further subset of focal schools was randomly selected to participate in additional components of data collection. A breakdown of the study schools by region, phase, and focal status as well as a detailed description of selection methodology has been provided in previous annual reports.

HISD is the largest school district in the study. Participating schools generally served Hispanic and African American populations (63.8% and 28.9%, respectively), with most students (88.1%) identified as eligible for free and reduced lunch.

New Mexico LASER schools in the participating school districts ranged in size from 26 to 984 students. New Mexico districts served mostly Hispanic, White, and American Indian/ Alaskan populations (48.0%, 34.8%, and 11.7%, respectively), with over half of students (58.0%) qualifying for free and reduced lunch status.

The sizes of LASER schools in the participating school districts in North Carolina (NC) ranged from 186 to 930 students. Most of the districts served primarily White and African American populations (60.7% and 18.1%, respectively), with almost two-thirds of students (63.1%) identified as eligible for free or reduced lunch by the North Carolina Public Schools.

Findings

Although student achievement gains as measured by traditional standardized tests comprise only one component of a successful intervention, it is the single outcome of most interest to many constituencies. To obtain valid achievement outcome data, CREP researchers analyzed scores from only students in the elementary and middle school cohorts for whom both pretest and posttest (baseline and spring 2014) PASS scores were available and established baseline equivalence using these analytic samples. It is important to note that all schools identified as Phase 1 are considered to be in the "treatment" group regardless of their level of implementation of the LASER model, and that fidelity of implementation varied widely across regions and across schools within regions. It should also be noted that the statistical analyses utilized an "Intent-to-Treat" model, where students were included in the Phase 1 or Phase 2 groups for analysis based on their treatment status at the time of random assignment.

Important and positive trends between Phase 1 and Phase 2 schools are evidenced in exploratory subgroup outcomes related to characteristics commonly agreed upon as most valued by employers. Both the OE and PT sections of the PASS call upon students to communicate their knowledge in written form. They also engage students in activities associated with critical thinking and problem solving. These twenty-first century skills are associated with college and career readiness; it is therefore noteworthy that these are the areas of achievement in which Phase 1 schools excelled. It is also important to note that the underserved populations of economically disadvantaged and special needs students, as well as those for whom English is a second language, seem to have benefited from their experiences with LASER as reflected in scores on the PASS.

PASS results for all three regions combined follow, first for the MC, then the OE and PT sections. A summary of the key findings for each set of analyses is presented at the beginning of each section, followed by information on the samples included, baseline equivalence between the Phase 1 and Phase 2 groups, and the detailed outcomes by grade level (i.e., elementary cohort and middle school cohort) and subgroup. In keeping with guidelines in the most recent *What Work Clearinghouse Procedures and Standards Handbook* (Version 3.0), a clustering correction and Benjamini-Hochberg correction for multiple comparisons was applied to all statistically significant findings, including secondary contrasts (i.e., exploratory analyses). In addition, the overall and differential cluster-level attrition rate for each outcome was calculated and compared to the WWC allowable standards.

- Based on guidelines in the most recent *What Work Clearinghouse Procedures and Standards Handbook* (Version 3.0), guidelines in *Reviewer Guidance for Use with the Procedures and Standards Handbook (version 3.0)*, consultation with representatives from the WWC (See Figure A-2 and Figure A-3 in the Appendix), and the fact that the elementary school study was an RCT with low cluster-level attrition, it should receive a WWC Study Rating of "Meets WWC Group Design Standards without Reservations." While the middle school study was also an RCT, none of the outcomes met the WWC cluster-level attrition standards. For the OE and PT sections, this was due to issues of differential attrition related to Phase 2 schools (no Phase 1 middle schools were lost to attrition). However, all but two outcomes demonstrated baseline equivalence, meaning the middle school study should receive a WWC Study Rating of "Meets WWC Group Design Standards with Reservations."

There were three exploratory outcomes for the elementary study that were statistically significant and positive after the cluster correction: The ELL subgroup on PASS OE and PT, and the IEP subgroup on the PT. However, there were no statistically significant positive confirmatory or exploratory findings in either the elementary or middle school studies after applying both the cluster and Benjamini-Hochberg correction for multiple comparisons that the WWC, but not IES, applies to secondary contrasts. As stated by The Institute of Education Sciences (IES) at the U.S. Department of Education (Schochet, 2008), "multiplicity adjustments are not required for exploratory analyses". Furthermore, it should be noted that there were no statistically significant or substantively important confirmatory or exploratory negative findings for the elementary study. Therefore, based on the IES guidance, these three exploratory outcomes for the elementary study in favor of Phase 1 schools, which remained statistically significant after the cluster correction, should still be considered meaningful and positive findings. In addition, the IEP (*g* = 0.39) and ELL (*g* = 0.30) PT exploratory outcomes had substantively important effect sizes. According to the *What Work Clearinghouse Procedures and Standards Handbook* (Version 3.0):

> For the WWC, effect sizes of 0.25 standard deviations or larger are considered to be **substantively important**. Effect sizes at least this large are interpreted as a qualified positive (or negative) effect, even though they may not reach statistical significance in a given study (What Works Clearinghouse, 2014, p. 23).

Furthermore, the American Statistical Association (ASA) recently released a statement on the use and interpretation of *p*-values in which they concluded:

> Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold….A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result….By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis (Wasserstein and Lazar, 2016, pp. 131-132).

Given that the WWC considers substantively important effect sizes as a qualified positive effect, even if not statistically significant, the ASA's recent guidance on *p*-values, as well as the IES guidance on multiple comparison corrections for exploratory analyses, the SSEC's LASER model can claim success in meeting its goal of improving student achievement.

All Regions:
Results for Spring 2014 PASS
Multiple Choice

All Regions Spring 2014 PASS Multiple Choice Key Findings for Phase 1

After applying the cluster correction and Benjamini-Hochberg correction for multiple comparisons that the WWC applies to secondary contrasts (i.e., exploratory analyses), and calculating the overall and subgroup cluster-level attrition, there were no statistically significant or substantively important PASS scaled score outcomes favoring Phase 1 elementary schools on the Spring 2014 PASS multiple choice section. It should be noted that while none of the middle school analyses (overall or by subgroup) met the WWC cluster-level attrition standard, all outcomes demonstrated baseline equivalence using the analytic samples. For the elementary cohort, only the IEP subgroup did not meet the WWC cluster-level attrition standard, but the subgroup did demonstrate baseline equivalence with the analytic sample.

Fall 2011 to Spring 2014 PASS Results:
All Regions

There were a total of 29 multiple choice questions on both the Fall 2011 and Spring 2014 forms of the PASS (PASS MC) addressing five broad science content standard categories for the elementary cohort and six broad science content standard categories for the middle school cohort. Scaled scores on the PASS MC for both elementary and middle schools range from 0-600. Only students who answered at least one multiple choice achievement question at both time points were included in the analyses for each respective area of analysis.

PASS Multiple Choice: All Regions

Table 3 shows the final cluster (i.e., school) and student sample sizes employed in the elementary cohort analyses (5th graders in 2013-2014) once students missing data on all 29 PASS MC questions at either time point were excluded.

Table 3: PASS MC, Spring 2014: School and Student Samples for the PASS MC Analyses for the Elementary Cohort: All Regions

Sample	Phase 1	Phase 2
Schools available for the PASS MC achievement analysis	51	43
Students available for the PASS MC achievement analysis	2,338	1,785

Table 4 shows the final school and student sample sizes employed in the middle school cohort analyses (8th graders in 2013-2014) once students missing all 29 PASS MC questions at either time point were excluded.

Table 4: PASS MC, Spring 2014: School and Student Samples for the PASS MC Analyses for the Middle School Cohort: All Regions

Sample	Phase 1	Phase 2
Schools available for the PASS MC achievement analysis	11	11
Students available for the PASS MC achievement analysis	1,036	1,132

To determine baseline equivalence on the Fall 2011 PASS MC scaled scores between Phase 1 and Phase 2 for elementary and middle schools included in the present analysis, a series of independent *t*-tests were conducted using the analytic samples for all elementary and middle schools in the aggregate as well as for subgroups of designated by their Special Education (IEP) status, English Language Learner (ELL) status, Economically Disadvantaged (FRL) status, and Gender. In addition, an effect size was also calculated as a measure of baseline equivalence.

As an indicator of the impact or "practical significance" of the treatment, the "effect size" (calculated as Hedges' *g*) is a descriptive statistic that indicates the magnitude of the difference (in standard deviation units) between two measures. For example, a positive effect size would indicate a higher (i.e., better)

Phase 1 mean, while a negative effect size would indicate a higher (i.e., better) Phase 2 mean. Based on guidelines from the What Works Clearinghouse (WWC), a unit within the research division of the U.S. Department of Education, an effect size of +/- 0.25 is considered to be “substantively important” (What Works Clearinghouse, 2014).

With respect to the elementary cohort (Table 5), in the aggregate (the “All” group), there was no statistically significant difference by Phase between schools in baseline achievement levels (*t* (4121) = -0.75, *p* =.45, *g* = -0.02, PR = 49) based on the analytic samples. At the same time, the ELL subgroup was the only one that demonstrated a statistically significantly difference in baseline achievement, with Phase 2 ELL schools outperforming their Phase 1 counterparts, although based on the effect size (*g*), not to a substantively meaningful degree (*t* (926.9) = -2.36, *p* = .02, *g* = -0.15, PR = 44). Overall, there were no substantively important effect size differences for the elementary cohort schools, meaning there was baseline equivalence for all groups based on the analytic samples.

Table 5: Baseline Comparison of Fall 2011 PASS MC Scaled Scores for Elementary Cohort Phase 1 (Treatment) and Phase 2 (Control) Schools (N = 94): All Regions

Group	Treatment (Phase 1)				Control (Phase 2)				*t*	*g*	PR
	School n	*Student n*	*M*	*SD*	*School n*	*Student n*	*M*	*SD*			
Elementary Cohort											
All	51	2,338	312.02	101.33	43	1,785	314.39	98.11	-0.75	-0.02	49
IEP	42	209	267.05	93.13	30	152	260.83	98.96	0.61	0.06	53
ELL	45	537	263.82	91.57	38	418	277.31	84.24	-2.36*	-0.15	44
FRL	47	1,416	284.79	94.85	40	1,060	289.39	90.29	-1.23	-0.05	48
Female	51	1,157	310.55	97.49	43	887	311.97	96.89	-0.33	-0.01	49

Note: PR = The percentile rank of the average Phase 1 student in the control group based on the effect size (*g*). For example, if the PR is 60, then the average Phase 1 student scored at the 60th percentile of the control group.
Note: PASS MC scaled scores range from 0-600.
* *p* < .05.

Likewise, with respect to schools in the middle school cohort (Table 6), there was no statistically significant difference between schools in baseline achievement by Phase (*t* (2166) = 1.17, *p* =.24, *g* = 0.05, PR = 52) in the aggregate based on the analytic samples. When the outcomes for the FRL subgroup were compared by Phase, there was a statistically significant difference in Fall 2011 PASS scores that favored Phase 1 schools, but the effect size linked to the comparison did not meet WWC criteria for substantive importance (i.e., *g* ≥ 0.25) (*t* (1223.2) = 3.62, *p* < .01, *g* = 0.20, PR = 58). On the other hand, there was a statistically significant difference in Fall 2011 PASS scores for the ELL subgroup, and the effect size associated with the difference met the WWC threshold for substantive importance, favoring Phase 1 schools (*t* (181) = 3.30, *p* < .01, *g* = 0.49, PR = 69). Therefore, the outcome for the ELL subgroup comparison for the middle school cohort should be interpreted in light of the substantively important difference in baseline achievement between Phase 1 and Phase 2 schools.
Employing these Fall 2011 data as covariates to statistically adjust the outcomes for baseline differences in achievement for the analytic sample, analyses were conducted on Spring 2014 PASS MC scaled scores

to determine differences between Phase 1 and Phase 2 elementary and middle schools, with scaled scores on the Spring 2014 PASS MC used as the outcome measure.

Table 6: Baseline Comparison of Fall 2011 PASS MC Scaled Scores for Middle School Cohort Phase 1 (Treatment) and Phase 2 (Control) Schools (N = 22): All Regions

Group	Treatment (Phase 1)				Control (Phase 2)				*t*	*g*	PR
	School n	*Student n*	*M*	*SD*	*School n*	*Student n*	*M*	*SD*			
Middle School Cohort											
All	11	1,036	364.51	102.66	11	1,132	359.10	112.40	1.17	0.05	52
IEP	10	111	282.22	96.57	6	114	276.25	112.07	0.43	0.06	52
ELL	10	83	290.70	82.07	11	100	248.08	90.75	3.30*	0.49	69
FRL	11	644	339.08	98.64	11	614	317.63	110.88	3.62*	0.20	58
Female	11	531	367.68	99.85	11	562	357.67	108.00	1.59	0.10	54

Note: PR = The percentile rank of the average Phase 1 student in the control group based on the effect size (*g*). For example, if the PR is 60, then the average Phase 1 student scored at the 60th percentile of the control group.
Note: PASS MC scaled scores range from 0-600.
* *p* < .05.

Elementary and Middle School Cohort PASS Multiple Choice Analyses: All Regions

With respect to the cohort of elementary schools in Phase 1 (*n* = 51) and Phase 2 (*n* = 43) and the cohort of middle schools in Phase 1 (*n* = 11) and Phase 2 (*n* =11), a set of ANCOVA analyses intended to generate pairs of adjusted scaled score means and to compute the treatment effect sizes (*g*) were conducted on the PASS MC outcomes for all elementary and middle schools by Phase within cohort, as well as for subgroups, categorized by their IEP status, ELL status, FRL status, and Gender (see ***Table 7*** and ***Table 8***).

Elementary Cohort PASS Multiple Choice Spring 2014 Results: All Regions

For the elementary cohort schools across the three regions, while the overall (i.e., the "All" group) ANCOVA adjusted scaled score mean presented in ***Table 7*** was higher for Phase 1 schools (*n* = 51, Adjusted Mean = 435.80) compared to Phase 2 schools (*n* = 43, Adjusted Mean = 434.88), it also fell short of being statistically significant (*F* (1, 4116) = 0.15, *p* = 0.698, *g* = 0.01, *PR* = 50), and the effect size (*g* = 0.01) was not substantively important. Consistent with these overall outcomes, two subgroup analyses (IEP and ELL) were linked to positively signed effect sizes that favored Phase 1 schools in the elementary cohort (see ***Table 7***). Meanwhile, while the ELL subgroup in Phase 2 schools statistically significantly outperformed the ELL subgroup in Phase 1 schools at baseline, the ELL subgroup in Phase 1 schools had a higher adjusted mean scaled score on the posttest that ultimately fell short of being statistically significant or substantively important. Overall, none of the effect sizes for the ANCOVA analyses were large enough to be substantively important, ranging from a low of -0.03 (FRL and Female) to a high of 0.19 (IEP).

Table 7: PASS MC, Spring 2014: Subgroup Mean Scaled Score Comparison for Elementary Cohort Phase 1 (Treatment) and Phase 2 (Control) Schools (N = 94): All Regions

	Treatment (Phase 1)					Control (Phase 2)								
Group	*School n*	*Student n*	*M*	*SD*	*Adj. M*	*School n*	*Student n*	*M*	*SD*	*Adj. M*	*F*	*p*	*g*	PR
All	51	2,338	435.28	88.72	435.80	43	1,785	435.56	88.76	434.88	0.15	0.698	0.01	50
IEP	42	209	392.08	104.38	390.08	30	152	366.49	116.11	369.23	3.86	0.050	0.19	58
ELL	45	537	402.87	100.18	405.64	38	418	403.77	104.85	400.21	0.80	0.370	0.05	52
FRL	47	1,416	415.28	94.98	415.85	40	1,060	419.38	94.50	418.62	0.66	0.416	-0.03	49
Female	51	1,157	434.92	84.14	435.33	43	887	438.12	84.10	437.59	0.50	0.481	-0.03	49

Note: PR = The percentile rank of the average Phase 1 student in the control group based on the effect size (*g*). For example, if the PR is 60, then the average Phase 1 student scored at the 60th percentile of the control group.
Note: PASS MC scaled scores range from 0-600.
* *p* < .05

Middle School Cohort PASS MC Spring 2014 Results: All Regions

For the schools across the three regions in the middle school cohort, unlike the outcomes observed for the elementary cohort, the overall scaled score performance result for the ANCOVA analysis (i.e., the "All" group) shown in ***Table 8*** was negative for middle school cohort Phase 1 schools (*n* = 11, Adjusted Mean = 323.02) compared to middle school cohort Phase 2 schools (*n* = 11, Adjusted Mean = 327.22), and was not statistically significant (*F* (1, 2161) = 1.35, *p* = 0.246). In addition, the effect size (*g* = -0.04, *PR* = 48) was not substantively important. Furthermore, despite advantage of Phase 1 schools on the pretest for all subgroups, including a substantively important advantage of Phase 1 schools on the Fall 2011 baseline for the ELL subgroup, Phase 2 outperformed Phase 1 for all subgroups. However, the effect size favoring the Phase 2 IEP subgroup (*g* = -0.28) was the only substantively important subgroup effect found, and indicated that the average Phase 1 student scored at the 39th percentile of the control group. Furthermore, none of the outcomes was statistically significant, including the outcome for the IEP subgroup, which was not statistically significant after applying the cluster correction.

Table 8: PASS MC, Spring 2014: Subgroup Mean Scaled Score Comparison for Middle School Cohort Phase 1 (Treatment) and Phase 2 (Control) Schools (N = 22): All Regions

	Treatment (Phase 1)					Control (Phase 2)									
Group	*School n*	*Student n*	*M*	*SD*	*Adj. M*	*School n*	*Student n*	*M*	*SD*	*Adj. M*	*F*	*p*	*p^*	*g*	PR
All	11	1,036	323.75	110.85	323.02	11	1,132	326.55	106.00	327.22	1.35	0.246		-0.04	48
IEP	10	111	220.02	124.36	217.99	6	114	250.53	124.45	252.50	5.91	0.016*	0.196	-0.28	39
ELL	10	83	235.75	107.78	224.46	11	100	232.97	106.64	242.34	1.39	0.240		-0.17	43
FRL	11	644	299.72	111.57	293.04	11	614	293.58	109.45	300.58	2.25	0.134		-0.07	47
Female	11	531	333.00	104.39	330.93	11	562	330.84	99.39	332.80	0.16	0.690		-0.02	49

Note: PR = The percentile rank of the average Phase 1 student in the control group based on the effect size (*g*). For example, if the PR is 60, then the average Phase 1 student scored at the 60th percentile of the control group.
Note: PASS MC scaled scores range from 0-600.
* *p* < 0.05.
*p^* = Clustering-corrected statistical significance

All Regions:
Results for Spring 2014 PASS
Open-Ended and Performance Task

All Regions Spring 2014 PASS Open-Ended and Performance Task Key Findings for Phase 1

After applying both the cluster correction and Benjamini-Hochberg correction for multiple comparisons that the WWC applies to secondary contrasts (exploratory analyses), and calculating the overall and subgroup cluster-level attrition, for all students combined (the "All" group) and the specified subgroups, the following outcomes (percentage correct) favoring Phase 1 elementary and middle schools were found on the Spring 2014 PASS Open-Ended/Constructed Response (OE), and Performance Task (PT) sections. It should be noted that while none of the middle school analyses (overall or by subgroup) met the WWC cluster-level attrition standard (due to differential attrition rates related to Phase 2 schools, as no Phase 1 schools were lost to attrition), all outcomes demonstrated baseline equivalence with the analytic samples except for the ELL subgroup on the Open-Ended section. For the elementary cohort, only the IEP subgroup did not meet the WWC cluster-level attrition standard, but the subgroup did demonstrate baseline equivalence with the analytic sample.

ELL

Elementary Cohort Open-Ended: Phase 1 schools had statistically significantly higher achievement than Phase 2 schools after the cluster correction. However, the difference was not statistically significant after the Benjamini-Hochberg correction for multiple comparisons that the WWC, but not IES, applies to secondary contrasts (i.e., exploratory analyses).

Elementary Cohort Performance Task: After controlling for the statistically significant advantage Phase 2 schools demonstrated on the pretest (*g* = -0.18), Phase 1 schools demonstrated a statistically significant advantage over Phase 2 schools after the cluster correction. The difference, however, was not statistically significant after the Benjamini-Hochberg correction for multiple comparisons that the WWC, but not IES, applies to secondary contrasts (exploratory analyses), but was substantively important (*g* = 0.30).

Middle School Cohort Performance Task: Phase 1 schools had a substantially important advantage over Phase 2 schools on the posttest (*g* = 0.37).

Economically Disadvantaged (FRL)

Middle School Cohort Performance Task: Phase 1 schools outperformed Phase 2 schools with an effect size that was substantively important (*g* = 0.27).

IEP

Elementary Cohort Performance Task: Phase 1 schools demonstrated statistically significantly higher achievement than Phase 2 schools after the cluster correction, but the difference was not statistically significant after the Benjamini-Hochberg correction for multiple comparisons that the WWC, but not IES, applies to secondary contrasts (i.e., exploratory analyses). In addition, the effect size was substantially important (*g* = 0.39).

Female

Middle School Cohort Performance Task: Phase 1 schools outperformed Phase 2 schools with a nearly substantively important effect size (*g* = 0.23).

Spring 2014 PASS Open-Ended and Performance Task Results: All Regions

Introduction

A random sample of schools in the three regions took the PASS Open-Ended and Performance Task assessment for the first time in Spring 2012, (end of first posttest year) and again in Spring 2013 and Spring 2014 (second and third posttest years, respectively). Students in the elementary cohort (5th graders in 2013-2014) responded to two Open-Ended (OE) and six Performance Task (PT) items, while middle schools (8th graders in 2013-2014) responded to six OE and six PT items. It should be noted that a random sample of schools in the HISD middle school cohort took the OE and PT sections for the first time in Spring 2013, and are therefore not included in these analyses.

PASS Open Ended and Performance Task Scoring

For the elementary cohort, there are a total of six points possible for the OE section and 17 total points possible for the PT section. For the middle school cohort, there are a total of 15 points possible for the OE section and 17 total points possible for the PT section. The items are scored using a rubric, with the number of points available for each item in each section shown in Table 9 below. In order to score a section, the student had to answer at least one item (i.e., gave a response that received a score of zero or higher). Otherwise, the section was dropped from the analysis if all the items were either missing, scored a “B” (blank), or had a combination of missing data and scores of “B”. If the section was scored, any item with a “B” and any missing items were given a value of zero. As a result, when a section was scored and a student had missing items or items scored with a “B”, those items were treated the same as the case where a student actually responded to an item, but received a score of zero, indicating the response did not contain any correct elements or was irrelevant. For both the OE and PT sections, the outcome score used in the analyses was the percentage correct out of the total number of points possible.

Table 9: PASS OE and PT Scoring Scales, Spring 2012, Spring 2013, and Spring 2014

Elementary Cohort				Middle School Cohort			
Open-ended Question		Performance Task		Open-ended Question		Performance Task	
Item	Scale	Item	Scale	Item	Scale	Item	Scale
1	B, 0, 1, 2, 3	1	B, 0, 1, 2, 3	1	B, 0, 1, 2	1	B, 0, 1, 2, 3
2	B, 0, 1, 2, 3	2	B, 0, 1, 2, 3	2	B, 0, 1, 2	2	B, 0, 1, 2, 3
		3	B, 0, 1, 2, 3	3	B, 0, 1, 2	3	B, 0, 1, 2, 3
		4	B, 0, 1, 2, 3	4	B, 0, 1, 2, 3	4	B, 0, 1, 2, 3
		5	B, 0, 1, 2, 3	5	B, 0, 1, 2, 3	5	B, 0, 1, 2, 3
		6	B, 0, 1, 2	6	B, 0, 1, 2, 3	6	B, 0, 1, 2
Total Points	6	Total Points	17	Total Points	15	Total Points	17

B = Blank

A summary of the Key Findings for each set of analyses is presented at the beginning of the report, followed by information on the samples included, baseline equivalence between the Phase 1 and Phase 2 schools, and the detailed outcomes by grade level (i.e., elementary cohort and middle school cohort), outcome (PASS OE and PASS PT) and subgroup.

A preliminary analysis was conducted on the Spring 2012 OE and PT sections of the PASS for students in the analytic sample who had a Spring 14 OE or PT percent correct score to determine baseline equivalence between Phase 1 and Phase 2 for elementary and middle schools included the present analysis (see Table 10) as the PASS OE and PT sections were not administered until the end of the first posttest year, meaning there was no Fall 2011 baseline scores available. In addition, an effect size was also calculated as a measure of baseline equivalence.
As an indicator of the impact or "practical significance" of the treatment, the "effect size" (calculated as Hedges' *g*) is a descriptive statistic that indicates the magnitude of the difference (in standard deviation units) between two measures. For example, a positive effect size would indicate a higher (i.e., better) Phase 1 mean, while a negative effect size would indicate a higher (i.e., better) Phase 2 mean. Based on guidelines from the What Works Clearinghouse, a unit within the research division of the U.S. Department of Education, an effect size of +/- 0.25 is considered to be "substantively important" (What Works Clearinghouse, 2014).

Results indicated that for the elementary cohort aggregate scores (i.e., for all students combined), there was no statistically significant difference between Phase 1 and Phase 2 schools on the Spring 2012 OE or PT percent correct, along with no substantially important effect sizes according to What Work Clearinghouse (WWC) standards. For the middle school cohort aggregate scores, Phase 1 schools had a statistically significantly higher mean Spring 2012 OE percent correct, as well as Spring 2012 PT percent correct, with the magnitude of the effects for both being substantially important.

Table 10: PASS OE and PT, Spring 2012, Treatment (Phase 1) and Control (Phase 2) School Percent Correct Means Comparison: All Regions

			Treatment (Phase 1)				Control (Phase 2)				
Section	Cohort	*School n*	*Student n*	*M*	*SD*	*School n*	*Student n*	*M*	*SD*	*t*	*g*
Open-Ended	Elementary	35	1,159	43.3	20.66	31	991	44.43	18.69	-1.37	-0.06
Performance Task	Elementary	35	1,326	53.68	19.76	32	1,099	54.41	17.35	-0.97	-0.04
Open-Ended	Middle School	8	795	72.6	16.32	7	578	68.06	19.43	4.56*	0.26
Performance Task	Middle School	8	697	52.11	20.08	7	514	42.23	23.45	7.69*	0.46

* $p < 0.05$

Due to the fact that the PASS OE and PT were not administered until the end of the first posttest year, meaning there were no true baseline scores available, and due to substantively meaningful differences on the Spring 2012 scores, correlation analyses were conducted to examine the relationship between the Spring 2014 PASS OE and PT percent correct and (1) the Spring 2012 PASS OE and PT percent correct, as well as (2) the Fall 2011 PASS Multiple Choice (MC) scaled score results, to determine which scores would serve as the better baseline measure of achievement. The analyses revealed statistically significant, but low correlations among each of the measures of achievement (see ***Table 11***). For the schools in both the elementary and middle school cohorts, the Fall 2011 PASS MC scaled scores had higher statistically significant correlations with the Spring 2014 PASS OE and PT, compared to the Spring 2012 OE and PT.

Table 11: Correlations on the Percent Correct for Spring 2014 PASS OE and PT with Spring 2012 PASS OE and PT, and Fall 2011 PASS Multiple Choice for Phase 1 and Phase 2 Schools: All Regions

Spring 2014 PASS	Cohort	Fall 2011 PASS Multiple Choice	Spring 2012 Open-Ended	Spring 2012 Performance Task
Spring 2014 Open-Ended	Elementary	0.37*	0.33*	NA
	Middle School	0.45*	0.38*	NA
Spring 2014 Performance Task	Elementary	0.36*	NA	0.35*
	Middle School	0.39*	NA	0.34*

* $p < 0.05$

To determine baseline equivalence on the Fall 2011 PASS MC scaled score between the analytic samples in Phase 1 and Phase 2 elementary and middle schools included the present analyses, a series of independent *t*-tests were conducted for all elementary and middle schools in the aggregate as well as for subgroups identified by their Special Education (IEP) status, English Language Learner (ELL) status, Economically Disadvantaged (FRL) status, and Gender (see Table 12). For the analytic sample in the aggregate elementary OE cohort (i.e., the “All” group), Phase 2 schools demonstrated a statistically significant advantage over Phase 1 schools in their baseline achievement levels ($t(2583) = -2.53$, $p = 0.011$, $g = -0.10$, PR = 46), but the effect size linked to this advantage did not meet WWC criteria for substantive importance (i.e., $g \geq 0.25$). In addition to this overall difference in performance, a statistically significant, but not substantively important advantage was also observed to favor the analytic sample in the Female subgroup in Phase 2 schools in the elementary cohort.

For schools in the middle school OE cohort, no statistically significant difference in aggregate performance (i.e., the “All” group) between the analytic sample in Phase 1 and Phase 2 schools was observed ($t(1525) = 1.02$, $p = 0.309$, $g = 0.05$, PR = 52), and the associated effect size did not meet the WWC criteria for substantive importance. Meanwhile, a statistically significant, but not substantively important advantage in baseline performance for the analytic sample was observed for the FRL subgroup in Phase 1 middle schools. Additionally, the ELL subgroup in Phase 1 schools had an advantage over Phase 2 schools that was not statistically significant, but was substantively important ($g = 0.31$).

With respect to the elementary PT cohort in the aggregate (i.e., the “All” group), Phase 2 schools demonstrated a statistically significant advantage over Phase 1 schools in their baseline achievement levels for the analytic sample ($t(2599) = -2.20$, $p = 0.028$, $g = -0.09$, PR = 47), but the effect size linked to this advantage did not meet WWC criteria for substantive importance (i.e., $g \geq 0.25$). Consistent with this overall difference in performance, a statistically significant, but not substantively important advantage was observed to favor the ELL subgroup in Phase 2 schools.

With respect to the middle school PT cohort, no statistically significant difference in aggregate performance (i.e., the “All” group) between the analytic samples in Phase 1 and Phase 2 schools was observed ($t(1406) = -0.55$, $p = 0.582$, $g = -0.03$, PR = 49), and the associated effect size did not meet the

WWC criteria for substantive importance. No statistically significant or substantively important advantages in baseline performance were observed for the analytic samples in any of the four subgroups for middle schools. While neither the Spring 2012 PASS OE and PT nor the Fall 2011 PASS Multiple Choice provided complete baseline equivalence between Phase 1 and Phase 2 schools, the Fall 2011 PASS Multiple Choice was administered as a true baseline assessment vs. the Spring 2012 PASS OE and PT, which was not administered until the end of the first posttest year. Therefore, due to its stronger relationship to the Spring 2014 PASS OE and PT outcomes, and because it was a true baseline measure, the Fall 2011 PASS Multiple Choice scaled score was chosen as the covariate (i.e., pretest measure) for both the elementary and middle school cohort analyses.

Table 12: Fall 2011 PASS Multiple Choice, Treatment (Phase 1) and Control (Phase 2) School Mean Scaled Score Comparison: All Regions

	Treatment (Phase 1)				Control (Phase 2)						
Group	*School n*	*Student n*	*M*	*SD*	*School n*	*Student n*	*M*	*SD*	*t*	*g*	PR
Elementary Cohort - Open-Ended											
All	35	1,409	306.03	100.39	31	1,176	316	98.81	-2.53*	-0.10	46
IEP	28	133	265.75	92.62	21	90	258.42	102.86	0.55	0.08	53
ELL	32	370	263.68	92.01	28	247	277.96	86.41	-1.94	-0.16	44
FRL	32	890	280.08	95.69	28	659	287.52	89.2	-1.56	-0.08	47
Female	35	693	302.23	94.44	31	584	313.05	97.34	-2.01*	-0.11	46
	Treatment (Phase 1)				Control (Phase 2)						
Group	*School n*	*Student n*	*M*	*SD*	*School n*	*Student n*	*M*	*SD*	*t*	*g*	PR
Middle School Cohort - Open-Ended											
All	8	832	368.45	103.45	7	695	362.84	111.7	1.02	0.05	52
IEP	7	83	275.46	90	6	71	269.59	120.65	0.34	0.06	52
ELL	7	44	274.25	80.46	7	64	246.67	91.15	1.62	0.31	62
FRL	8	490	338.92	99.53	7	380	318.23	108.71	2.92*	0.20	58
Female	8	434	372.06	99.2	7	352	359.58	108.43	1.68	0.12	55
	Treatment (Phase 1)				Control (Phase 2)						
Group	*School n*	*Student n*	*M*	*SD*	*School n*	*Student n*	*M*	*SD*	*t*	*g*	PR
Elementary Cohort - Performance Task											
All	35	1,429	308.05	101.59	32	1,172	316.73	98.39	-2.20*	-0.09	47
IEP	28	132	266.52	91.6	22	94	254.68	100.22	0.92	0.12	55
ELL	32	371	263.52	92.41	29	238	279.6	85.71	-2.15*	-0.18	43
FRL	32	895	280.24	95.46	29	654	288.61	88.3	-1.76	-0.09	46
Female	35	703	303.87	94.93	32	581	314.23	96.22	-1.93	-0.11	46
	Treatment (Phase 1)				Control (Phase 2)						
Group	*School n*	*Student n*	*M*	*SD*	*School n*	*Student n*	*M*	*SD*	*t*	*g*	PR
Middle School Cohort - Performance Task											
All	8	772	365.64	104.81	7	636	368.78	107.97	-0.55	-0.03	49
IEP	7	84	271.93	90.84	5	61	280.89	114.37	-0.53	-0.09	46
ELL	7	42	274.98	82.04	7	50	259.06	91.01	0.87	0.08	57
FRL	8	465	338.95	100.62	7	338	325.94	106.72	1.76	0.13	55
Female	8	405	368.36	102.13	7	328	362.54	105.82	0.75	0.06	52

* $p < 0.05$

Note: PASS MC scaled scores range from 0-600.

Employing the Fall 2011 PASS MC data as a covariate to statistically adjust the outcomes for baseline differences in achievement, preliminary analyses were conducted on Spring 2014 PASS OE and PT percent correct scores to determine any differences between Phase 1 and Phase 2 elementary and middle schools. As noted earlier, for the elementary cohort, there were statistically significant differences between Phase 1 and Phase 2 schools on the baseline measures for both the OE and PT, with Phase 2 schools having an advantage both overall and for several subgroups. For the middle school cohort, the Phase 1 ELL subgroup had a substantively important advantage and the FRL subgroup had a statistically significant advantage on the OE section. Due to these baseline differences, results for these particular groups should be interpreted with these advantages in mind.

Elementary and Middle School Cohorts PASS Open-Ended Analyses: All Regions

A set of ANCOVA analyses intended to generate pairs of adjusted percentage correct scores and to compute the treatment effect sizes (*g*) was conducted by Phase within cohort on the PASS OE outcomes for all elementary (Phase 1 *n* = 35, Phase 2 *n* = 31) and middle schools (Phase 1 *n* = 8, Phase 2 *n* = 7), as well as for subgroups, categorized by their Special Education (IEP) status, English Language Learner (ELL) status, Economically Disadvantaged (FRL) status, and Gender.

Elementary Cohort Spring 2014 PASS Open-Ended Results: All Regions

For the elementary cohort schools across the three regions, after applying the cluster correction, the ANCOVA adjusted means presented in Table 13 indicated no statistically significant difference between Phase 1 and Phase 2 schools overall (i.e., the "All" group) (*F* (1, 2578) = 6.32, *p*^ = 0.240, *g* = 0.09, *PR* = 54). In addition, the magnitude of the effect size for the All group (*g* = 0.09) was not considered to be substantively important. It should be noted that on the pretest for this group, Phase 2 schools had a statistically significant advantage over Phase 1 schools, although it was not considered substantively important. Only the ELL subgroup demonstrated a statistically significant difference after the cluster correction, which favored Phase 1 schools (*F* (1, 610 = 6.70, *p*^ = 0.043, *g* = 0.20, *PR* = 58). The difference, though, was not statistically significant after the Benjamini-Hochberg correction for multiple comparisons that the WWC applies to secondary contrasts (exploratory analyses). It should be noted, however, that according to The Institute of Education Sciences (IES) at the U.S. Department of Education (Schochet, 2008), "multiplicity adjustments are not required for exploratory analyses." Therefore, based on the IES guidance, the ELL outcome, which was still statistically significant after the cluster correction, should still be considered a meaningful finding. Furthermore, even though Phase 2 schools had an advantage on the pretest overall and for all but the IEP subgroup, after controlling for pretest differences, Phase 1 schools outperformed Phase 2 schools on the posttest for all groups, although no posttest effect size was substantively important.

Table 13: PASS Open-Ended Questions, Spring 2014: Mean Percent Correct Comparison of Phase 1 (Treatment) and Phase 2 (Control) Elementary Cohort Schools (N = 66): All Regions

		Treatment (Phase 1)				Control (Phase 2)										
Group	*School n*	*School n*	*M*	*SD*	*Adj. M*	*School n*	*School n*	*M*	*SD*	*Adj. M*	*F*	*p*	*p*^	*P*^^	*g*	PR
All	35	1,409	65.87	21.09	66.39	31	1,176	65.12	20.18	64.50	6.32	0.012*	0.240		0.09	54
IEP	28	133	56.27	22.99	55.97	21	90	51.48	22.16	51.92	1.91	0.168			0.18	57
ELL	32	370	61.89	22.38	62.38	28	247	58.77	20.54	58.05	6.70	0.010*	0.043*	0.013	0.20	58
FRL	32	890	63.43	21.69	63.69	28	659	61.51	20.35	61.15	6.34	0.012*	0.134		0.12	55
Female	35	693	67.68	20.07	68.21	31	584	67.01	19.47	66.37	3.22	0.073			0.09	54

* *p* < 0.05.

p^ = Clustering-corrected statistical significance

p^^ = Benjamini-Hochberg correction for multiple comparisons correction of the clustering-corrected statistical significance. To remain statistically significant after the multiple comparison correction, *p*^ ≤ *p*^^

Middle School Cohort Spring 2014 PASS Open-Ended Results: All Regions

Across the middle schools in the three regions, the ANCOVA adjusted means presented in Table 14 indicated no statistically significant difference between Phase 1 and Phase 2 schools overall (i.e., the "All" group) (*F* (1, 1520) = 0.51, *p* = 0.477, *g* = 0.03, *PR* = 51). While no subgroup comparison was statistically significant, the ELL subgroup produced an effect size that was substantively important favoring Phase 2 schools (*g* = -0.32) (with the ELL subgroup in Phase 1 schools having an advantage (*g* = 0.31) on the pretest).

Table 14: PASS Open-Ended Questions, Spring 2014: Mean Percent Correct Comparison of Phase 1 (Treatment) and Phase 2 (Control) Middle School Cohort Schools (N = 15): All Regions

	Treatment (Phase 1)					Control (Phase 2)								
Group	*School n*	*Student n*	*M*	*SD*	*Adj. M*	*School n*	*Student n*	*M*	*SD*	*Adj. M*	*F*	*p*	*g*	PR
All	8	832	85.32	15.49	85.08	7	695	84.32	15.32	84.60	0.51	0.477	0.03	51
IEP	7	83	70.92	20.04	70.28	6	71	69.48	23.55	70.24	0.00	0.992	0.00	50
ELL	7	44	66.67	19.40	66.25	7	64	72.40	20.12	72.68	2.77	0.099	-0.32	37
FRL	8	490	82.91	16.74	63.69	7	380	80.02	17.01	61.15	1.09	0.297	0.15	56
Female	8	693	67.68	20.07	68.21	7	584	67.01	19.47	66.37	0.36	0.549	0.09	54

* $p < 0.05$.

Elementary and Middle School Cohorts PASS Performance Task Analyses: All Regions

A set of ANCOVA analyses intended to generate pairs of adjusted percentage correct scores and to compute the treatment effect sizes (*g*) was conducted by Phase within cohort on the PASS PT outcomes for all elementary (Phase 1 *n* = 35, Phase 2 *n* = 32) and middle schools (Phase 1 *n* = 8, Phase 2 *n* = 7), as well as for subgroups categorized by their Special Education (IEP) status, English Language Learner (ELL) status, Economically Disadvantaged (FRL) status, and Gender.

Elementary Cohort Spring 2014 PASS Performance Task Results: All Regions

For the elementary cohort schools across the three regions, the ANCOVA adjusted means presented in Table 15 demonstrate no statistically significantly difference between Phase 1 and Phase 2 schools overall (i.e., the "All" group) after applying the cluster correction (*F* (1, 2594) = 6.28, *p^* = 0.355, *g* = 0.09, *PR* = 54), indicating that the average Phase 1 student scored at the 54th percentile of the control group. In addition, the effect size was not considered to be substantively important according to WWC standards. On the other hand, both the IEP (*F* (1, 219) = 10.16, *p^* = 0.014, *g* = 0.39, *PR* = 65) and ELL (*F* (1, 602) = 15.54, *p^* = 0.039, *g* = 0.30, *PR* = 62) subgroups in Phase 1 schools demonstrated statistically significant outcomes after the cluster correction, but neither difference remained statistically significant after the Benjamini-Hochberg correction for multiple comparisons that the WWC applies to secondary contrasts (exploratory analyses). Furthermore, effect sizes for both subgroups were substantively important. The IEP subgroup, however, did not meet the WWC cluster-level attrition standard, but did demonstrate baseline equivalence with the analytic sample. Again, based on the IES guidance that adjustments for multiple comparisons are not required for exploratory analyses (Schochet, 2008), the IEP and ELL outcomes, which remained statistically significant after the cluster correction, should still be considered meaningful findings.

Meanwhile, even though Phase 2 schools had a statistically significant advantage on the pretest overall and for the ELL subgroup, for both groups, after controlling for statistically significant pretest differences (All, *g* = -0.09, and ELL, *g* = -0.18), there was no statistically significant difference between Phase 1 and Phase 2 schools on the posttest. Furthermore, both the IEP (*g* = 0.39) and ELL (*g* = 0.30) subgroups had substantively important posttest effect sizes. In addition, after controlling for the advantage of the Phase 2 FRL (*g* = -0.09) and Female (*g* =-0.11) subgroups on the pretest, the Phase 1 FRL and Female subgroups were able to demonstrate small, but positive effect sizes on the posttest (*g* = 0.14 and *g* = 0.06 respectively).

Table 15: PASS Performance Task Questions, Spring 2014: Mean Percent Correct Comparison of Phase 1 (Treatment) and Phase 2 (Control) Elementary Cohort Schools (N = 67): All Regions

	Treatment (Phase 1)					Control (Phase 2)										
Group	*School n*	*Student n*	*M*	*SD*	*Adj. M*	*School n*	*Student n*	*M*	*SD*	*Adj. M*	*F*	*p*	*p^*	*P^^*	*g*	PR
All	35	1,429	66.16	15.50	66.55	32	1,172	65.56	16.82	65.09	6.28	0.012*	0.355		0.09	54
IEP	28	132	60.29	16.26	59.96	22	94	52.25	21.16	52.72	10.16	0.002*	0.014*	0.004	0.39	65
ELL	32	371	63.20	14.85	63.55	29	238	59.17	17.95	58.63	15.54	<0.001*	0.039*	0.008	0.30	62
FRL	32	895	63.75	15.47	63.95	29	654	61.93	17.56	61.66	8.37	0.004*	0.157		0.14	56
Female	35	703	67.15	14.82	67.57	32	581	67.22	16.26	66.70	1.18	0.278			0.06	52

* $p < 0.05$.

p^ = Clustering-corrected statistical significance

p^^ = Benjamini-Hochberg correction for multiple comparisons correction of the clustering-corrected statistical significance. To remain statistically significant after the multiple comparison correction, *p*^ ≤ *p*^^

Middle School Cohort Spring 2014 PASS Performance Task Results: All Regions

Across the middle schools in the three regions, the ANCOVA adjusted means presented in ***Table 16*** demonstrate no statistically significant difference between Phase 1 and Phase 2 schools overall (i.e., the "All" group) after applying the cluster correction (*F* (1, 1401) = 19.09, *p* <0.001, *p*^ = 0.516, *g* = 0.12, *PR* = 58), indicating that the average Phase 1 student scored at the 58th percentile of the control group. In addition, the effect size was not considered to be substantively important according to WWC standards. Furthermore, while there were no statistically significant differences for any subgroups after applying the cluster correction, for the All and IEP subgroups, after controlling for a Phase 2 advantage on the pretest (*g* = -0.55 and *g* = -0.53 respectively), Phase 1 schools outperformed Phase 2 schools on the posttest (*g* = 0.12 and *g* = 0.19 respectively). Meanwhile, although Phase 1 schools had advantages on the pretest for three additional subgroups (ELL, FRL, and Female) that were neither statistically significant nor substantially important (ELL, *g* = 0.08, FRL, *g* = 0.13, and Female, *g* = 0.06) the Phase 1 advantages were even stronger on the posttest with substantially important effects for the ELL (*g* = 0.37) and FRL subgroups (*g* = 0.27), and nearly substantially important effects for the Female subgroup (*g* = 0.23).

Table 16: PASS Performance Task Questions, Spring 2014: Mean Percent Correct Comparison of Phase 1 (Treatment) and Phase 2 (Control) Middle School Cohort Schools (N = 15): All Regions

	Treatment (Phase 1)					Control (Phase 2)									
Group	*School n*	*Student n*	*M*	*SD*	*Adj. M*	*School n*	*Student n*	*M*	*SD*	*Adj. M*	*F*	*p*	*p^*	*g*	PR
All	8	772	58.64	24.49	58.81	7	636	53.95	23.01	53.74	19.09	<0.001*	0.516	0.12	58
IEP	7	84	40.97	20.43	41.02	5	61	37.32	19.65	37.25	1.28	0.260		0.19	57
ELL	7	42	45.10	20.45	44.48	7	50	36.94	18.45	37.21	3.46	0.066		0.37	65
FRL	8	465	55.75	23.31	55.26	7	338	48.45	21.70	49.13	17.19	<0.001*	0.170	0.27	61
Female	8	405	61.68	24.33	61.40	7	328	55.69	22.94	56.04	11.21	0.001*	0.287	0.23	59

* $p < 0.05$

*p^* = Clustering-corrected statistical significance

References

Schochet, P. Z. (2008). *Technical Methods Report*: *Guidelines for Multiple Testing in Impact Evaluations* (NCEE 2008-4018). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/pdf/20084018.pdf

Wasserstein, R.L. & Lazar, N.A. (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician, 70:2*, 129-133. doi: 10.1080/00031305.2016.1154108

What Works Clearinghouse (2014). Procedures and standards handbook (Version 3.0). Washington, DC: Author. Retrieved from ies.ed.gov/ncee/wwc/pdf/reference_resources/ wwc_procedures_v3_0_standards_handbook.pdf

What Works Clearinghouse (2016). Reviewer Guidance for Use with the Procedures and Standards Handbook (version 3.0). Washington, DC: Author. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_reviewer_guidance_030416.pdf

Appendix

Table A-1: Clustering Correction for Mismatched Analyses

Due to the fact that the random assignment was carried out at the cluster level (i.e., school level), but the analyses were conducted at the student level, a clustering correction was applied to the *p*-values for statistically significant outcomes of the ANCOVA analyses to calculate clustering-corrected statistical significance levels (*p*-values).

Subtest	Group	Treatment (Phase 1) *n*	Control (Phase 2) *n*	*g*	Uncorrected *p*	M	ICC	*t*	t_a	*df*	Corrected *p*
Elementary Cohort											
PASS OE	All	1,409	1,176	0.09	0.012	94	0.10	2.278617	1.1744776	2014.564	0.240
PASS OE	ELL	370	247	0.20	0.010	83	0.07	2.434091	2.0241855	596.9861	0.043*
PASS OE	FRL	890	659	0.12	0.012	87	0.08	2.335034	1.5001646	1384.108	0.133
PASS PT	All	1,429	1,172	0.09	0.012	94	0.19	2.283769	0.9253471	1332.025	0.354
PASS PT	IEP[1]	132	94	0.39	0.002	72	0.17	2.889757	2.4783267	211.5658	0.013*
PASS PT	ELL	371	238	0.30	<0.001	83	0.32	3.612335	2.0716161	369.8459	0.038*
PASS PT	FRL	895	654	0.14	0.004	87	0.16	2.721465	1.4145159	1086.117	0.157
Middle School Cohort											
PASS MC	IEP[1]	111	114	-0.28	0.016	16	0.12	-2.09981	-1.2978464	189.7401	0.195
PASS PT	All[1]	772	636	0.12	<0.001	22	0.17	2.240872	0.6498379	518.5035	0.516
PASS PT	FRL[1]	465	338	0.27	<0.001	22	0.18	3.777381	1.3750417	381.8819	0.169
PASS PT	Female[1]	405	328	0.23	0.001	22	0.23	3.09628	1.0658156	285.8985	0.287

[1] Subgroup did not meet the WWC cluster-level attrition standard. All subgroups demonstrated baseline equivalence.

* *p* < 0.05 after clustering correction

Table A-2: Benjamini-Hochberg Correction for Multiple Comparisons

Due to the fact that the What Works Clearinghouse applies a multiple comparison correction for both primary (confirmatory) and secondary (exploratory) contrasts, the Benjamini-Hochberg correction was applied to the following secondary contrasts that remained statistically significant after the cluster correction.

Subtest	Group	Clustering Corrected *p*-value (p_x)	*p*-value Rank (*x*)	Alpha	*x**Alpha	Total Number of Tests	New Critical *p*-value (p_x'= *x**Alpha/Total Number of Tests)	Finding *p*-value < New Critical *p*-value? ($p_x \leq p_x'$)	Statistical Significance after BH Correction?
Elementary Cohort									
PASS OE	ELL	0.043	3	0.05	0.15	12	0.013	No	No
PASS PT	IEP[1]	0.014	1	0.05	0.05	12	0.004	No	No
PASS PT	ELL	0.039	2	0.05	0.10	12	0.008	No	No

[1] Subgroup did not meet the WWC cluster-level attrition standard. All subgroups demonstrated baseline equivalence.

Note: To remain statistically significant after the multiple comparison correction, the Clustering Corrected *p*-value ≤ New Critical *p*-value *(px ≤ px')*

Table A-3: Elementary School Cohort Cluster-Level Attrition

Group	Cluster Category	MC			OE			PT		
		Phase 1	**Phase 2**	**Overall**	**Phase 1**	**Phase 2**	**Overall**	**Phase 1**	**Phase 2**	**Overall**
All	Total	54	51	105	35	34	69	35	34	69
	Dropped	3	8	11	0	3	3	0	2	2
	Attrition Rate	**5.6%**	**15.7%**	**10.5%**	**0.0%**	**8.8%**	**4.3%**	**0.0%**	**5.9%**	**2.9%**
	Differential Attrition		**10.1%**			**8.8%**			**5.9%**	
	Final Count	**51**	**43**	**94**	**35**	**31**	**66**	**35**	**32**	**67**
IEP	Total	51	47	98	33	32	65	33	32	65
	Dropped	9	17	26	5	11	16	5	10	15
	Attrition Rate	**17.6%**	**36.2%**	**26.5%**	**15.2%**	**34.4%**	**24.6%**	**15.2%**	**31.3%**	**23.1%**
	Differential Attrition		**18.5%**			**19.2%**			**16.1%**	
	Final Count	**42**	**30**	**72**	**28**	**21**	**49**	**28**	**22**	**50**
ELL	Total	52	48	100	35	33	68	35	33	68
	Dropped	7	10	17	3	5	8	3	4	7
	Attrition Rate	**13.5%**	**20.8%**	**17.0%**	**8.6%**	**15.2%**	**11.8%**	**8.6%**	**12.1%**	**10.3%**
	Differential Attrition		**7.4%**			**6.6%**			**3.5%**	
	Final Count	**45**	**38**	**83**	**32**	**28**	**60**	**32**	**29**	**61**
FRL	Total	51	48	99	33	31	64	33	31	64
	Dropped	4	8	12	1	3	4	1	2	3
	Attrition Rate	**7.8%**	**16.7%**	**12.1%**	**3.0%**	**9.7%**	**6.3%**	**3.0%**	**6.5%**	**4.7%**
	Differential Attrition		**8.8%**			**6.6%**			**3.4%**	
	Final Count	**47**	**40**	**87**	**32**	**28**	**60**	**32**	**29**	**61**
Females	Total	54	51	105	35	34	69	35	34	69
	Dropped	3	8	11	0	3	3	0	2	2
	Attrition Rate	**5.6%**	**15.7%**	**10.5%**	**0.0%**	**8.8%**	**4.3%**	**0.0%**	**5.9%**	**2.9%**
	Differential Attrition		**10.1%**			**8.8%**			**5.9%**	
	Final Count	**51**	**43**	**94**	**35**	**31**	**66**	**35**	**32**	**67**

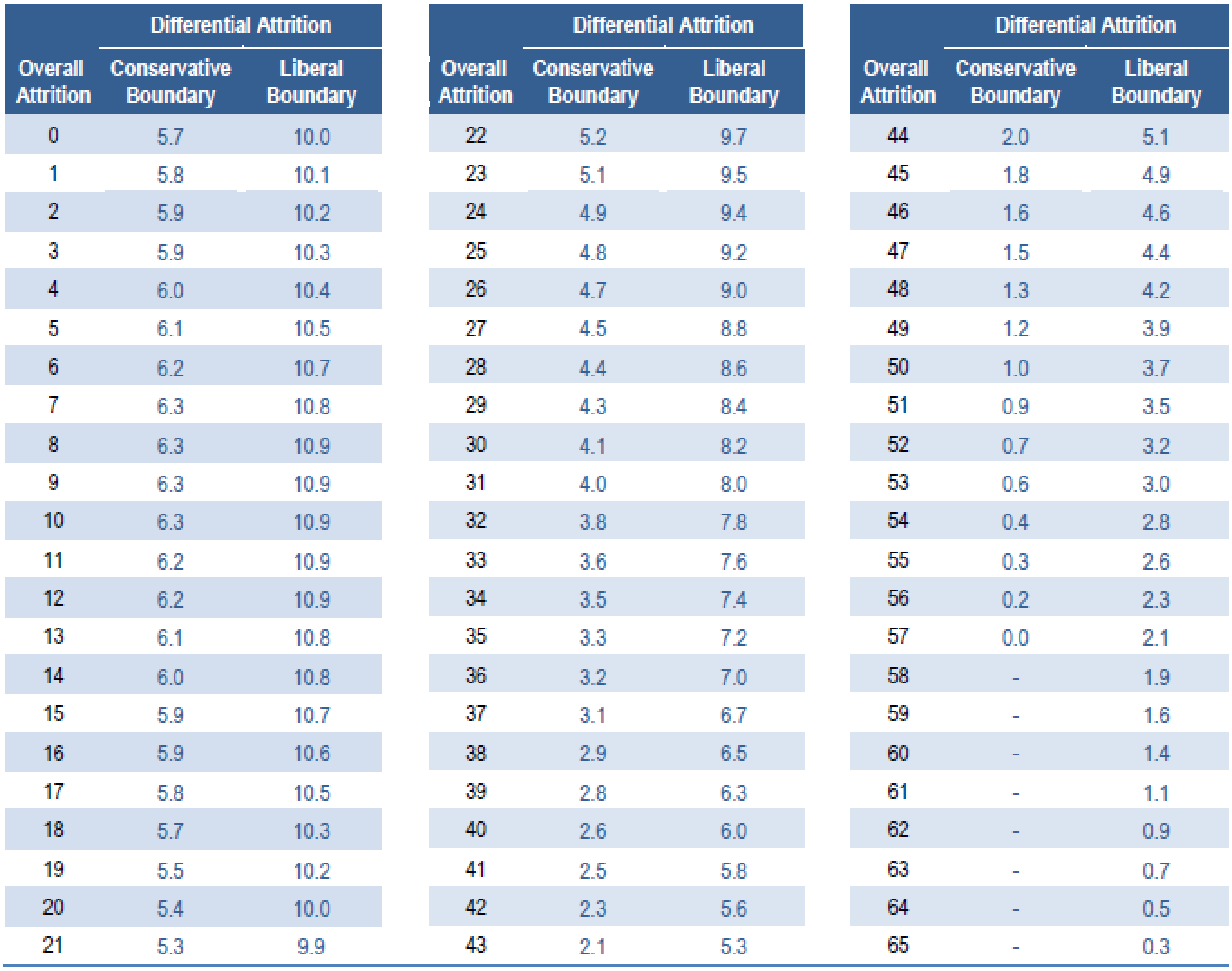
Note: Only the IEP subgroup did not meet the WWC cluster-level attrition standard.

Table A-4: Middle School Cohort Cluster-Level Attrition

Group	Cluster Category	MC			OE			PT		
		Phase 1	**Phase 2**	**Overall**	**Phase 1**	**Phase 2**	**Overall**	**Phase 1**	**Phase 2**	**Overall**
All	Total	13	17	30	8	8	16	8	8	16
	Dropped	2	6	8	0	1	1	0	1	1
	Attrition Rate	**15.4%**	**35.3%**	**26.7%**	**0.0%**	**12.5%**	**6.3%**	**0.0%**	**12.5%**	**6.3%**
	Differential Attrition		**19.9%**			**12.5%**			**12.5%**	
	Final Count	**11**	**11**	**22**	**8**	**7**	**15**	**8**	**7**	**15**
IEP	Total	12	12	24	7	7	14	7	6	13
	Dropped	2	6	8	0	1	1	0	1	1
	Attrition Rate	**16.7%**	**50.0%**	**33.3%**	**0.0%**	**14.3%**	**7.1%**	**0.0%**	**16.7%**	**7.7%**
	Differential Attrition		**33.3%**			**14.3%**			**16.7%**	
	Final Count	**10**	**6**	**16**	**7**	**6**	**13**	**7**	**5**	**12**
ELL	Total	12	17	29	7	8	15	7	8	15
	Dropped	2	6	8	0	1	1	0	1	1
	Attrition Rate	**16.7%**	**35.3%**	**27.6%**	**0.0%**	**12.5%**	**6.7%**	**0.0%**	**12.5%**	**6.7%**
	Differential Attrition		**18.6%**			**12.5%**			**12.5%**	
	Final Count	**10**	**11**	**21**	**7**	**7**	**14**	**7**	**7**	**14**
FRL	Total	13	17	30	8	8	16	8	8	16
	Dropped	2	6	8	0	1	1	0	1	1
	Attrition Rate	**15.4%**	**35.3%**	**26.7%**	**0.0%**	**12.5%**	**6.3%**	**0.0%**	**12.5%**	**6.3%**
	Differential Attrition		**19.9%**			**12.5%**			**12.5%**	
	Final Count	**11**	**11**	**22**	**8**	**7**	**15**	**8**	**7**	**15**
Females	Total	13	17	30	8	8	16	8	8	16
	Dropped	2	6	8	0	1	1	0	1	1
	Attrition Rate	**15.4%**	**35.3%**	**26.7%**	**0.0%**	**12.5%**	**6.3%**	**0.0%**	**12.5%**	**6.3%**
	Differential Attrition		**19.9%**			**12.5%**			**12.5%**	
	Final Count	**11**	**11**	**22**	**8**	**7**	**15**	**8**	**7**	**15**

Note: Neither the Overall sample or any subgroup met the WWC cluster-level attrition standard.

Table A-5: What Works Clearinghouse (WWC) Allowable Overall and Differential Attrition Rates



	Differential Attrition			Differential Attrition			Differential Attrition	
Overall Attrition	Conservative Boundary	Liberal Boundary	Overall Attrition	Conservative Boundary	Liberal Boundary	Overall Attrition	Conservative Boundary	Liberal Boundary
0	5.7	10.0	22	5.2	9.7	44	2.0	5.1
1	5.8	10.1	23	5.1	9.5	45	1.8	4.9
2	5.9	10.2	24	4.9	9.4	46	1.6	4.6
3	5.9	10.3	25	4.8	9.2	47	1.5	4.4
4	6.0	10.4	26	4.7	9.0	48	1.3	4.2
5	6.1	10.5	27	4.5	8.8	49	1.2	3.9
6	6.2	10.7	28	4.4	8.6	50	1.0	3.7
7	6.3	10.8	29	4.3	8.4	51	0.9	3.5
8	6.3	10.9	30	4.1	8.2	52	0.7	3.2
9	6.3	10.9	31	4.0	8.0	53	0.6	3.0
10	6.3	10.9	32	3.8	7.8	54	0.4	2.8
11	6.2	10.9	33	3.6	7.6	55	0.3	2.6
12	6.2	10.9	34	3.5	7.4	56	0.2	2.3
13	6.1	10.8	35	3.3	7.2	57	0.0	2.1
14	6.0	10.8	36	3.2	7.0	58	-	1.9
15	5.9	10.7	37	3.1	6.7	59	-	1.6
16	5.9	10.6	38	2.9	6.5	60	-	1.4
17	5.8	10.5	39	2.8	6.3	61	-	1.1
18	5.7	10.3	40	2.6	6.0	62	-	0.9
19	5.5	10.2	41	2.5	5.8	63	-	0.7
20	5.4	10.0	42	2.3	5.6	64	-	0.5
21	5.3	9.9	43	2.1	5.3	65	-	0.3

Note: Reproduces Table III.1 in the WWC Procedures and Standards Handbook (Version 3.0): *Highest Differential Attrition for a Sample to Maintain Low Attrition, by Overall Attrition, Under Liberal and Conservative Assumptions*

Figure A-1: Letter from What Works Clearinghouse (WWC) Confirming Appropriateness of Reviewing the Report as Two Separate Studies

See the following pages.

What Works Clearinghouse WWC

A central and trusted source of scientific evidence for what works in education.

June 21, 2016

Dr. Carol O'Donnell
Smithsonian Science Education Center
Smithsonian Institution
901 D Street NW, Suite 704-B, MRC 952
Washington, DC 200024
odonnellc@si.edu

Dear Dr. O'Donnell,

Thank you for your email concerning the What Works Clearinghouse (WWC) review of M. Alberg's report entitled, "The LASER Model: A systemic and sustainable approach for achieving high standards in science education" (Released July 2015). In response to your inquiry, we conducted an in-depth, independent quality review to address the issues you raised.

The WWC quality review team responds to concerns raised about WWC reviews published on our website. When a quality review is conducted, a senior WWC researcher who was not involved in the initial review undertakes an independent assessment of the study in question. The senior reviewer also investigates the procedures used and decisions made during the original review of the study. These quality reviews are one of the tools used to ensure that the standards established by the Institute of Education Sciences are upheld on every review conducted by the WWC.

In your email dated June 4, 2016, you stated that the Alberg (2015) report should have been reviewed as two separate studies: one for the elementary school cohort and a second for the middle school cohort. You stated that these two cohorts were described and analyzed separately in the original report, and thus should be reviewed as two separate studies.

We investigated this issue in the quality review, revisited our review of the report, and came to the following conclusion: it is most appropriate to review this report as two separate studies.

The WWC Procedures and Standards Handbook (version 3.0) defines a study as:

> *"….the examination of the effect of an intervention on a particular sample (e.g., a set of students, schools, or districts) and set of outcomes. To be a separate study, the sampling errors must be independent. For randomized controlled trials, a study is defined by randomization. This definition excludes subgroups from being their own studies because they were randomized at the same time as the full sample and treats additional cohorts without rerandomization of the unit of assignment as a*

> *single study; however, if the same units were rerandomized to condition, then they are separate studies." (p. 7)*

The original WWC review of the Alberg report was completed in September of 2015. In March of 2016, the WWC published an additional document providing Reviewer Guidance for Use with the Procedures and Standards Handbook (available online at http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=260). This new Guidance document further clarifies that:

> *"…if a single research team reports pooled impacts combining distinct randomized samples, that will be considered one study. The reviewer should classify a sample (and corresponding analysis) as one study if the intervention and comparison groups were formed in either one randomization process or multiple processes reported as pooled findings by one team. The reviewer should classify a sample (and corresponding analysis) as* n *studies if the intervention and comparison groups were formed in* n *randomization processes and not combined…" (p. 4)*

In the reviewed study, the research team used a matched-pair randomization design where elementary schools were paired and randomized to conditions, and, through a separate process, middle schools were paired and randomized to conditions. Because the intervention and comparison groups were formed via different randomization processes for the elementary and middle schools, and given that the report presented findings separately for all elementary and middle school cohorts, the independent quality review confirmed that this report should be reviewed as two separate studies.

The WWC updates the reviews of all studies when the results of a QRT indicate a revision is necessary. As a result of this independent quality review, the WWC will update the review of the Alberg (2015) report to reflect that two studies are contained within the report.

I hope that this letter has addressed your concerns. If you have other concerns, please do not hesitate to contact the WWC through info@whatworks.ed.gov.

Sincerely,

Emily Tanner Smith

Emily E. Tanner-Smith, Ph.D.   Joshua R. Polanin, Ph.D.
What Works Clearinghouse at Development Services Group, Inc.

Figure A-2: What Works Clearinghouse (WWC) Email Chain Regarding Cluster-Level Inference

See the following pages.

Todd Alan Zoblotsky (tzbltsky)

From:	WhatWorks <What.Works@icfi.com>
Sent:	Tuesday, August 23, 2016 3:56 PM
To:	Todd Alan Zoblotsky (tzbltsky)
Subject:	RE: What Works Clearinghouse (WWC 6154 6157)

Dr. Zoblotsky,

Yes, that is correct. A cluster only contributes to cluster-level attrition if it met the requirements to be included in the subsample at the time of randomization.

Thank you,
What Works Clearinghouse

From: Todd Alan Zoblotsky (tzbltsky) [mailto:tzbltsky@memphis.edu]
Sent: Tuesday, August 23, 2016 11:10 AM
To: WhatWorks
Subject: RE: What Works Clearinghouse (WWC 6154 6157)

Thank you very much for your response. Just to clarify, in the case of attrition for a specific subgroup (e.g., ELL) at the cluster level, our starting point for attrition would only be clusters that had students in that particular subgroup, correct? In other words, we do not start with the full sample of clusters, and if a school does not have any students in a particular subgroup (e.g., ELL), count that school against our attrition for that subgroup, correct? My assumption is that for cluster-level attrition, the school is counted against our attrition only if all students in the particular subgroup (e.g., ELL) who had baseline data did not have outcome data. Am I understanding this correctly?

Thanks,

Todd

Todd A Zoblotsky, Ed.D.
Research Associate Professor
Center for Research in Educational Policy

From: WhatWorks [mailto:What.Works@icfi.com]
Sent: Tuesday, August 23, 2016 10:35 AM
To: Todd Alan Zoblotsky (tzbltsky) <tzbltsky@memphis.edu>
Subject: RE: What Works Clearinghouse (WWC 6154 6157)

Dear Dr. Zoblotsky,

Cluster-level attrition is assessed separately for each subgroup. However, if students in a subgroup are represented in all clusters at randomization and follow-up, attrition within the subgroup would be identical to attrition in the full sample.

Thank you,
What Works Clearinghouse

From: Todd Alan Zoblotsky (tzbltsky) [mailto:tzbltsky@memphis.edu]
Sent: Tuesday, August 23, 2016 6:54 AM
To: WhatWorks
Subject: RE: What Works Clearinghouse (WWC 6154 6157)

Thank you very much for your response.

If I could ask a follow up question: As the Reviewer Guidance documents states (p. 5): "If the study is making cluster-level inferences, the study can Meet WWC Group Design Standards without Reservations. The impact estimates reflect the combination of the effect on individuals who were in the clusters at the time of assignment and the effect on the composition of individuals within the clusters over time, which is noted in our reporting. For cluster-level inferences, attrition is assessed only at the cluster level."

Does this mean that attrition at the cluster level is only assessed for the overall sample (all students combined), or must the cluster level attrition also be assessed for each subgroup analysis reported (e.g., ELL, Students with a Disability)?

Thank You,

Todd

Todd A Zoblotsky, Ed.D.
Research Associate Professor
Center for Research in Educational Policy

From: WhatWorks [mailto:What.Works@icfi.com]
Sent: Monday, August 22, 2016 5:20 PM
To: Todd Alan Zoblotsky (tzbltsky) <tzbltsky@memphis.edu>
Subject: RE: What Works Clearinghouse (WWC 6154 6157)

Dear Dr. Zoblotsky,

The unit of analysis does not affect how the WWC treats joiners in reviews of cluster-level randomized control trials (RCTs). As clarified in the WWC guidance for the standards, the key issue is whether the study authors make claims about the effect of the intervention on outcomes for subcluster units, or only on cluster-level outcomes. Therefore, it is possible for an RCT with cluster-level assignment and that uses language consistent with cluster-level inferences to Meet WWC Group Design Standards Without Reservations regardless of whether the analysis was conducted using subcluster- or cluster-level units.

For example, if the authors of an RCT with school-level assignment make claims about the effect of the intervention on students, then the presence of students in the analytic sample who were not in schools at the time of random assignment jeopardizes the random assignment. However, if the authors only make claims about the effect of the intervention on schools then joiners are not considered a threat to inference. By definition, the effect of the intervention on school-level outcomes includes the possibility that the intervention affected the composition of the students within the schools. The unit of analysis has no relevance to these issues.

The unit of inference can affect the rating in studies that must demonstrate equivalence. In high-attrition cluster RCTs which make student-level inferences, baseline equivalence and statistical adjustment (if needed) must be performed using subcluster-level data. If the study makes cluster-level inferences, these could be done using either subcluster- or cluster-level data.

Finally, you also mentioned the relationship between the unit of analysis and statistical power. Please note that the WWC will apply a post hoc clustering correction to determine the statistical significance of findings from a subcluster-level analysis that did not appropriately adjust for clustering.

We hope this information is helpful.

Thank you,
What Works Clearinghouse

From: Todd Alan Zoblotsky (tzbltsky) [mailto:tzbltsky@memphis.edu]
Sent: Friday, August 19, 2016 3:23 PM
To: WhatWorks
Subject: RE: What Works Clearinghouse (WWC 6154 6157)

If I could add one more thing to my second question about cluster level analysis:

Reading the reviewer guidance document (p. 5), which I think would be the "final word" on this issue (?) they are clearly talking about how the analyses are discussed vs. analyzed:

"For an analysis of stayers and joiners, the rating depends on the type of inference described in the study, particularly in the discussion of findings. To understand the inferences the authors are making, read how they discuss the findings. If they are talking about improvement in school scores, then it is a cluster-level inference. If they are talking about increases in student scores, then it is an individual-level inference."

So this leads me to believe that for cluster level analysis/inference, it is how the analyses are discussed vs. conducted (i.e., it is acceptable to have individual level analysis with cluster level inference).

Thank You,

Todd

Todd A Zoblotsky, Ed.D.
Research Associate Professor
Center for Research in Educational Policy

From: WhatWorks [mailto:What.Works@icfi.com]
Sent: Friday, August 19, 2016 2:31 PM
To: Todd Alan Zoblotsky (tzbltsky) <tzbltsky@memphis.edu>
Subject: What Works Clearinghouse (WWC 6154 6157)

Hello,

Thank you for contacting the What Works Clearinghouse (WWC). We have received your two emails below. WWC staff are reviewing your request and we get back to you soon.

If you would like to be notified of new publications and updates to the WWC website, you can sign up for the News Flash email updates from the Institute of Education Sciences (IES). Please visit the link http://ies.ed.gov/newsflash/?url=%2Fncee%2Fwwc%2Findex%2Easp&site=What+Works+Clearinghouse, choose the National Center for Education Evaluation (NCEE) option (the second one), and select What Works Clearinghouse.

Thank you,

What Works Clearinghouse

The What Works Clearinghouse was established by the U.S. Department of Education's Institute of Education Sciences to provide educators, policymakers, researchers, and the public with a central and trusted source of scientific evidence of what works in education. For more information, please visit http://ies.ed.gov/ncee/wwc/.

-----Original Message-----
From: t.zoblotsky@memphis.edu [mailto:t.zoblotsky@memphis.edu]
Sent: Friday, August 19, 2016 1:39 PM
To: info@whatworks.ed.gov
Subject: IES WWC Website: Contact Us: Procedures and Standards Handbook, Reference ID Number: 1280259576

info@whatworks.ed.gov, this email was automatically sent through the Contact link on the WWC website.

From: t.zoblotsky@memphis.edu

Message: I need clarification on the following statement on page 10 of the current WWC handbook: &quot;Furthermore, a study with cluster-level assignment and cluster-level analysis may have changes in subcluster composition that are not subject to the attrition standard. A cluster-level analysis of stayers and joiners used to answer a cluster-level research question may Meet WWC Group Design Standards without Reservations. If the analysis is conducted at the individual level, any nonrandom movement or placement of individuals into the intervention or comparison groups after random assignment jeopardizes the random assignment design of the study. Individual-level studies of stayers or stayers plus joiners may Meet WWC Group Design Standards with Reservations if the study is able to demonstrate baseline equivalence of the analytic sample.&quot; Does this literally mean that for a cluster-level analysis that includes both stayers and joiner to Meet WWC Group Design Standards without Reservations, you must aggregate individual (student) data up to the cluster (school) level and do the analysis of school level means? This seems questionable due to the loss of power in moving from the student to the cluster level. In addition, how do control for covariates such as Special Education status or ELL status when you aggregate data from the individual to the cluster (school level)? Is it possible to have a study with cluster level assignment and cluster level research questions and inference with analysis at the individual level Meet WWC Group Design Standards without Reservations?

-----Original Message-----
From: t.zoblotsky@memphis.edu [mailto:t.zoblotsky@memphis.edu]
Sent: Thursday, August 18, 2016 4:22 PM
To: info@whatworks.ed.gov
Subject: IES WWC Website: Contact Us: Procedures and Standards Handbook, Reference ID Number: 1250124092

info@whatworks.ed.gov, this email was automatically sent through the Contact link on the WWC website.

From: t.zoblotsky@memphis.edu

Message: I have a question about how to apply the Benjamini-Hochberg correction for multiple comparisons in my situation. I have one outcome domain with three measures. For each of the three outcomes, I ran 9 separate statistical tests (All students combined, and 8 subgroups). Each analysis was run separately, so I have a total 0f 27 statistical tests that were done (3 outcomes x 9 groups). My question is: Should the total number of outcomes (M) I use for the correction be 27 (based on running 27 individual statistical tests) or should it be 3 (based on having 3 outcome measures)?

Figure A-3: What Works Clearinghouse (WWC) Email Chain Regarding Multiple Comparisons

See the following pages.

Todd Alan Zoblotsky (tzbltsky)

From: WhatWorks <What.Works@icfi.com>
Sent: Tuesday, August 23, 2016 10:34 AM
To: Todd Alan Zoblotsky (tzbltsky)
Subject: RE: What Works Clearinghouse (WWC 6154 6157)

Dear Dr. Zoblotsky,

Apologies for the unclear response. The WWC does not report on findings that describe the difference in effectiveness between two or more subgroups. Rather the WWC focuses on the main effects of the intervention for each subgroup of interest. For example, if a study examined the effectiveness of an intervention on boys and girls, but only reported how much more effective the intervention was for girls than for boys (often obtained from the coefficient on an interaction term), the WWC would ask the authors for measures of the effectiveness for each group separately along with standard errors and/or p-values from the test of the null hypothesis that the intervention had no effect on the subgroup. It is usually possible to obtain these subgroup-specific findings from an interacted regression, and the WWC would accept results from a variety of regression specifications.

From your description of your analysis, it sounds like the p-value you are reporting may reflect a null hypothesis about differences across subgroups in the effectiveness of the intervention, and not the effectiveness of the intervention for a specific subgroup as the WWC requires. If we are mistaken and the p-value you are reporting represents a null hypothesis about the effect of the intervention for the full sample (or some subgroup), then the WWC would review the finding, and if it were the only finding reported, the WWC would not apply a multiple comparisons adjustment (because one finding was reported for only a single sample). In other words, the regression specification used to report a set of findings does not affect how the WWC would adjust for multiple comparisons.

Finally, we should clarify that the WWC only applies the multiple comparisons adjustment to the findings that meet WWC design standards. Findings that do not meet standards are not included.

Thank you,
What Works Clearinghouse

From: Todd Alan Zoblotsky (tzbltsky) [mailto:tzbltsky@memphis.edu]
Sent: Monday, August 22, 2016 7:23 PM
To: WhatWorks
Subject: Re: What Works Clearinghouse (WWC 6154 6157)

Thank you for your reply. Sorry, but I am not quite clear. Are you saying that subgroups should be pulled out and run as separate analyses vs. including main and interaction effects?

Todd

Sent from my iPhone

On Aug 22, 2016, at 5:19 PM, "WhatWorks" <What.Works@icfi.com> wrote:

> Dear Dr. Zoblotsky,

At this time, WWC procedure is to use p-values from the statistical tests for each subgroup and adjust them for multiple comparison. If these p-values are not reported in the study, the WWC will evaluate whether these p-values can be generated based on available information by the WWC or send an author query. The WWC does not use p-values from interactions or other omnibus tests because the goal is to report on whether there is an effect in a specific subgroup (and not on whether effects are larger for one subgroup than another).

Thank you,
What Works Clearinghouse

From: Todd Alan Zoblotsky (tzbltsky) [mailto:tzbltsky@memphis.edu]
Sent: Monday, August 22, 2016 1:57 PM
To: WhatWorks
Subject: RE: What Works Clearinghouse (WWC 6154 6157)

If I could ask a follow up…would the same number of corrections (8) apply for secondary contrasts if they are run as cross-level interactions in HLM (which would also apply to interactions in OLS regression)? In the case of HLM, there is one test run with cross-level interactions for all 8 of the subgroup analyses. Would there still be a correction for 8 tests in that case, or since it was run as one analysis (with interactions), do the separate *p* values for the cross-level interactions need to be corrected at all (as opposed to running 8 separate ANCOVAs as in my original question, which was 8 separate statistical tests)?

Thank You,

Todd

Todd A Zoblotsky, Ed.D.
Research Associate Professor
Center for Research in Educational Policy

From: WhatWorks [mailto:What.Works@icfi.com]
Sent: Monday, August 22, 2016 1:39 PM
To: Todd Alan Zoblotsky (tzbltsky) <tzbltsky@memphis.edu>
Subject: RE: What Works Clearinghouse (WWC 6154 6157)

Hello,

Thank you for contacting the What Works Clearinghouse (WWC). The WWC conducts multiple comparison adjustments separately for primary and secondary contrasts. This is done to avoid penalizing main findings when the authors report results for multiple subgroups or multiple follow-up periods. In your example, the WWC would conduct multiple comparison adjustments for the primary contrasts based on 3 tests (all students sample, 3 outcome measures within the domain) and for the secondary contrasts based on 24 tests (8 subgroups by 3 outcomes measures within the domain). For more information, please see the WWC Reviewer Guidance, page 15 (available on our website here: http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_reviewer_guidance_030416.pdf).

We will respond to your other questions regarding cluster level analysis shortly.

Thank you,

What Works Clearinghouse

From: Todd Alan Zoblotsky (tzbltsky) [mailto:tzbltsky@memphis.edu]
Sent: Friday, August 19, 2016 3:23 PM
To: WhatWorks
Subject: RE: What Works Clearinghouse (WWC 6154 6157)

If I could add one more thing to my second question about cluster level analysis:

Reading the reviewer guidance document (p. 5), which I think would be the "final word" on this issue (?) they are clearly talking about how the analyses are discussed vs. analyzed:

"For an analysis of stayers and joiners, the rating depends on the type of inference described in the study, particularly in the discussion of findings. To understand the inferences the authors are making, read how they discuss the findings. If they are talking about improvement in school scores, then it is a cluster-level inference. If they are talking about increases in student scores, then it is an individual-level inference."

So this leads me to believe that for cluster level analysis/inference, it is how the analyses are discussed vs. conducted (i.e., it is acceptable to have individual level analysis with cluster level inference).

Thank You,

Todd

Todd A Zoblotsky, Ed.D.
Research Associate Professor
Center for Research in Educational Policy

From: WhatWorks [mailto:What.Works@icfi.com]
Sent: Friday, August 19, 2016 2:31 PM
To: Todd Alan Zoblotsky (tzbltsky) <tzbltsky@memphis.edu>
Subject: What Works Clearinghouse (WWC 6154 6157)

Hello,

Thank you for contacting the What Works Clearinghouse (WWC). We have received your two emails below. WWC staff are reviewing your request and we get back to you soon.

If you would like to be notified of new publications and updates to the WWC website, you can sign up for the News Flash email updates from the Institute of Education Sciences (IES). Please visit the link http://ies.ed.gov/newsflash/?url=%2Fncee%2Fwwc%2Findex%2Easp&site=What+Works+Clearinghouse, choose the National Center for Education Evaluation (NCEE) option (the second one), and select What Works Clearinghouse.

Thank you,

What Works Clearinghouse

The What Works Clearinghouse was established by the U.S. Department of Education's Institute of Education Sciences to provide educators, policymakers, researchers, and the public with a central and trusted source of scientific evidence of what works in education. For more information, please visit http://ies.ed.gov/ncee/wwc/.

-----Original Message-----
From: t.zoblotsky@memphis.edu [mailto:t.zoblotsky@memphis.edu]
Sent: Friday, August 19, 2016 1:39 PM
To: info@whatworks.ed.gov
Subject: IES WWC Website: Contact Us: Procedures and Standards Handbook, Reference ID Number: 1280259576

info@whatworks.ed.gov, this email was automatically sent through the Contact link on the WWC website.

From: t.zoblotsky@memphis.edu

Message: I need clarification on the following statement on page 10 of the current WWC handbook: &quot;Furthermore, a study with cluster-level assignment and cluster-level analysis may have changes in subcluster composition that are not subject to the attrition standard. A cluster-level analysis of stayers and joiners used to answer a cluster-level research question may Meet WWC Group Design Standards without Reservations. If the analysis is conducted at the individual level, any nonrandom movement or placement of individuals into the intervention or comparison groups after random assignment jeopardizes the random assignment design of the study. Individual-level studies of stayers or stayers plus joiners may Meet WWC Group Design Standards with Reservations if the study is able to demonstrate baseline equivalence of the analytic sample.&quot; Does this literally mean that for a cluster-level analysis that includes both stayers and joiner to Meet WWC Group Design Standards without Reservations, you must aggregate individual (student) data up to the cluster (school) level and do the analysis of school level means? This seems questionable due to the loss of power in moving from the student to the cluster level. In addition, how do control for covariates such as Special Education status or ELL status when you aggregate data from the individual to the cluster (school level)? Is it possible to have a study with cluster level assignment and cluster level research questions and inference with analysis at the individual level Meet WWC Group Design Standards without Reservations?

-----Original Message-----
From: t.zoblotsky@memphis.edu [mailto:t.zoblotsky@memphis.edu]
Sent: Thursday, August 18, 2016 4:22 PM
To: info@whatworks.ed.gov
Subject: IES WWC Website: Contact Us: Procedures and Standards Handbook, Reference ID Number: 1250124092

info@whatworks.ed.gov, this email was automatically sent through the Contact link on the WWC website.

From: t.zoblotsky@memphis.edu

Message: I have a question about how to apply the Benjamini-Hochberg correction for multiple comparisons in my situation. I have one outcome domain with three measures. For each of the three outcomes, I ran 9 separate statistical tests (All students combined, and 8 subgroups). Each analysis was run separately, so I have a total 0f 27 statistical tests that were done (3 outcomes x 9 groups). My question is: Should the total number of outcomes (M) I use for the correction be 27 (based on running 27 individual statistical tests) or should it be 3 (based on having 3 outcome measures)?