Empirical Education
EMPOWERING EDUCATORS FOR EVIDENCE-BASED DECISIONS

In School Settings, Are All RCTs Exploratory?

Denis Newman

Andrew P. Jaciw

Empirical Education Inc.

The motivation for this paper is our recent work on several randomized control trials in which we found the primary result, which averaged across subgroups or sites, to be moderated by demographic or site characteristics. We are led to examine a distinction that the Institute of Education Sciences (IES) makes between "confirmatory" and "exploratory" questions that are being addressed in an RCT. IES correctly wants to encourage a disciplined methodology in requiring that researchers "call their shots" by naming the small number of outcomes considered most important. All other questions are fine to look at but are in the category of exploratory work.

While we embrace the need for this kind of discipline, we want to guard against taking the notion of confirmation too far. Thus, we ask how much faith researchers, working in school settings, should put on the confirmatory results of RCTs? Is all rigorous school-based research exploratory at heart?

We suggest a positive answer to this based on concerns about generalizability and about the limits of external validity. We will cite two recent examples of our RCTs where the value of the confirmatory result is brought into question. In each case, we find moderator effects, which challenge the value for practitioners, developers, and policy makers of the overall confirmatory result.

What Confirmatory Means

Confirmatory questions are the small number that is taken as the primary outcomes.[1] The strict IES criteria were based on the principle that when a researcher is using tests of statistical significance, the probability of erroneously concluding that there is an impact when there isn't one increases with the frequency of the tests. The threshold

[1] We use primary to mean the main outcomes where the intervention is expected to have an impact. IES also distinguished "primary" and "secondary" confirmatory outcomes allowing adjustments for multiple comparisons to be done independently for the two sets of outcomes. Usage is also inconsistent among confirmatory "outcomes", "questions", or "analyses". In this paper, we use these terms as appropriate for the context.

for significance is made more stringent to keep the probability of falsely concluding that there was a difference for any of the outcomes at 5% (that is, $p < .05$). The adjustment of the threshold decreases the statistical power available to detect any one result so the researcher is cautious about naming too many. But more important, researchers must call their shots before the outcomes are evident. A similar issue applies to selecting a statistical model to be the basis for the answer before the results are known since it is inappropriate to run a large number of models and select the one that yields the lowest *p* value. These are all ways of preventing "fishing"—highlighting the best question or the best answer after the fact.

A paper by Schochet, published by IES, set out the basic idea behind exploratory analyses, as contrasted with confirmatory.

> *The purpose of the exploratory analysis is to examine relationships within the data to identify outcomes or subgroups for which impacts may exist. The goal of the exploratory analysis is to identify hypotheses that could be subject to more rigorous future examination, but cannot be examined in the present study because they were not identified ahead of time or statistical power was deemed insufficient. Results from post hoc analyses are not automatically invalid, but, irrespective of plausibility or statistical significance, they should be regarded as preliminary and unreliable unless they can be rigorously tested and replicated in future studies.* (Schochet, 2008, p. 4).

Consider the scenario in which the researcher has run an RCT and found no discernible difference on the confirmatory question: e.g., did the intervention have an impact on the state test score for math on average for the treatment group? Where the answer is no, it is then useful to see if there was an impact on any of the secondary outcomes or for some subgroup. But at this point, the researcher is inspecting the results to find where the intervention might have made a difference. The exploratory results may be valid but the experiment wasn't designed to confirm them. A new experiment in which the exploratory question becomes the confirmatory question is needed.

From the point of view of the internal logic of experiments, we agree with the need to stick with the principle of calling your shots. Selecting after the fact measures that happened to show an effect or subgroups for which the study measured a difference is susceptible to chance imbalance on unmeasured factors. Where we want to raise questions is about how evidence is used and whether by criteria of external validity do confirmatory results have greater value than exploratory results. Two examples will help put the issue in context.

Two RCTs as Case Studies

Alabama Math Science and Technology Initiative (AMSTI)

The Alabama Math, Science, and Technology Initiative (AMSTI) is a two-year intervention intended to better align classroom practices with national and statewide teaching standards—and ultimately to improve student achievement—by providing professional development, access to materials and technology, and in-school support for teachers. AMSTI, a schoolwide intervention, was introduced in a set of 20 schools in 2002. Each year since then, the state has rolled out the program to additional schools within its 11 regions. By 2009, about 40 percent of the state's 1,518 schools were designated as AMSTI schools. Funding for the program from the state legislature was $46 million in 2009. Given the policy relevance and level of investment in AMSTI, the Regional Educational Laboratory Southeast, with Empirical Education as the primary subcontractor, mounted a longitudinal, cluster randomized controlled trial to determine the effectiveness of AMSTI in grades 4–8, as implemented in five regions in the state. (Newman, Finney, Bell, Turner, Jaciw, Zacamy, & Feagans Gould, 2012).

Study design

This study is the first randomized controlled trial testing the effectiveness of AMSTI in improving mathematics problem solving and science achievement in upper-elementary and middle schools. AMSTI is an initiative specific to Alabama and was developed and supported through state resources.

In the cluster randomized trial, schools were randomized within matched pairs in which one school was randomly assigned to participate in AMSTI starting the first year and the second school was assigned to a control group the first year and to participate in AMSTI the second year. In all, 82 schools, 780 teachers, and 30,000 students participated in the study. The study's internal validity is based on a randomization procedure and is strengthened by the low rate (less than 5 percent) of attrition at all levels over the follow-up period.

Here we summarize the confirmatory and exploratory findings on the effect of AMSTI on the achievement of upper-elementary and middle school students and on classroom practices hypothesized to improve students' achievement. It also summarizes the effect of AMSTI on teacher content knowledge and student engagement and variations in effects on student achievement by specific subgroups after one year. The chapter concludes by identifying the study's strengths and limitations.

The statistically unbiased estimates of the effect of AMSTI were generated under authentic conditions for this program as implemented under ordinary conditions in

volunteer schools in Alabama. The study did not alter implementation specifically for the experiment but followed schools as they participated in the standard initiative.

Primary Confirmatory Findings

An important finding is the positive and statistically significant effect of AMSTI on mathematics achievement as measured by the SAT 10 mathematics problem solving assessment administered by the state to students in grades 4–8. After one year in the program, student mathematics scores were higher than those of a control group that did not receive AMSTI by 0.05 standard deviation, equivalent to 2 percentile points. (If the 50th percentile control student had been placed in an AMSTI school, the student would have scored in the 52nd percentile.) Nine of the 10 sensitivity analyses yielded effect estimates that were statistically significant at the .025 level, consistent with the main finding.

The effect is smaller than expected. Whether the statistically significant effect is important for education is open to interpretation. It might, however, be useful to convert the effect into the more policy-relevant metric of additional student progress measured in days of instruction. In these terms, the average effect of AMSTI can be translated into an estimated 28 days of additional student progress over students receiving conventional mathematics instruction. This value was obtained by dividing the estimate of the effect by the mean pretest to posttest difference on the SAT 10 mathematics problem solving assessment for the control group and assuming a 180-day school year.

The estimated effect of AMSTI on science achievement measured after one year was not statistically significant. Based on the SAT 10 science test administered by the state to students in grades 5 and 7, no difference between AMSTI and control schools could be discerned after one year.

Some Exploratory Findings: Reading and the Minority Interaction

AMSTI had a positive and statistically significant effect on reading achievement as measured by the SAT 10 test of reading. Reading scores of AMSTI students exceeded those of the control group by 0.06 standard deviation.

While interesting patterns of results emerged from the exploratory analysis of moderators, most did not meet reach the $p < .05$ threshold for statistical significance. In reading, AMSTI did have a statistically significant differential effect for minority and White students. This difference in estimated impact was 3.04 scale score points or an effect size of 0.08 standard deviations ($p < .001$) between the two groups. The effect of AMSTI on reading achievement for minority students was not statistically significant ($p = .294$); for White students, AMSTI had a positive and statistically significant effect on reading achievement ($p < .001$).

The moderator findings are detailed in Newman, et al (2012). Since exploratory analysis is meant to assist with hypothesis generation, it is useful to display the results graphically as shown in Figure 1. This represents the results reported in a table on p. 104 of the full report.

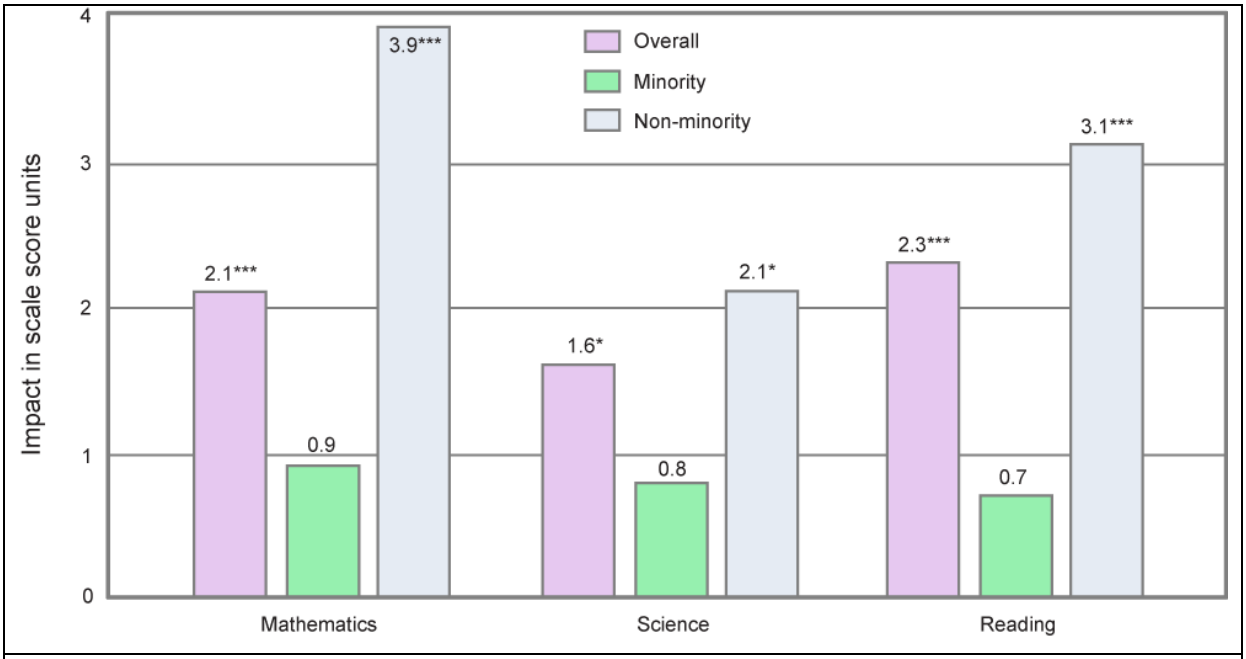

Figure 1. One Year Impacts Overall and by Minority Status

* Significant at $p < .10$ ** Significant at $p < .05$ *** Significant at $p < .01$

Here we can see that the same pattern is evident for math and science that was found statistically significant for reading. The same pattern is visually apparent also when free or reduced lunch is substituted for minority status. But again, those patterns did not reach the threshold for significance and not discussed as findings in the report.

HOUGHTON MIFFLIN HARCOURT FUSE: ALGEBRA 1

In spring 2010, Houghton Mifflin Harcourt (HMH) began planning a pilot of an application for the Apple iPad, *Houghton Mifflin Harcourt Fuse: Algebra 1 (HMH Fuse)*, which was then in development. The application was to be piloted in four California school districts during the 2010-2011 school year. HMH contracted with Empirical Education Inc. to conduct a one-year RCT aimed at producing evidence of the effectiveness of *HMH Fuse*. The full report is Toby, Ma, Lai, Lin, & Jaciw (2012).

HMH Fuse for the Apple iPad contains the content of the Holt McDougal Algebra 1 2011© text and includes interactive lessons, explanations, quizzes, and problem solving. In addition, *HMH Fuse* comes with the 300+ videos that are also available online to students using the traditional print version of the text. We compared classes

using HMH Fuse on the iPad with classes using the conventional text containing the same content.

For this RCT, we randomly assigned one algebra period for each of the 11 participating teachers to the program condition, in which they use *HMH Fuse*. Each teacher's remaining algebra sections formed the control group assigned to use the regular text version of the Holt McDougal Algebra 1 2011 program. Our primary or confirmatory outcome measure for algebra achievement was the California Standards Test (CST). In addition, we used the Riverside End of Course Assessment. Student attitudes were measured by means of a Student Attitude Questionnaire consisting of two pre-existing measures. We also gathered implementation data via student and teacher surveys to inform outcome results. Given the relatively small sample overall, our design aimed at establishing the average impact across the available units. An investigation of the differential impact across districts provided a strong indication that the average impact was misleading and has become the basis for the study's conclusion.

We found no impact of *HMH Fuse* on the primary measure of algebra achievement, the CST, on average across the four districts. There was no moderating effect of pretest on the outcome measure. Specifically, the impact of *HMH Fuse* was not different depending on the student's pretest scores on the CST.

One of the school districts initiated its own investigation of the data for the students that participated in both the *HMH Fuse* and the control classrooms. This work was conducted before Empirical Education had reported the overall results but found what appeared to be a strong impact (although an appropriate statistical test was not done). Because randomization was blocked by teacher, the evidence from that district alone constituted an RCT, albeit a very small one including only two teachers and nine sections of Algebra 1. We used the same statistical modeling approach to examine the subgroup impacts for the one district and for the other three. For the other three, and consistent with the overall results, there was no discernible difference between *HMH Fuse* and control. For the focal district, however, we found a substantial impact for which we can have strong confidence (p = .023). The adjusted effect size was 0.23, which is equivalent to a nine point increase in percentile standing. Our analysis does corroborate the results reported by the focal district for its participating teachers. It is also noteworthy that the teachers in that district reported more time instructing with HMH Fuse than reported by most of the other teachers in the study but that log data did not reflect more student usage.

After a one-year pilot implementation with *HMH Fuse,* we do not have evidence of a generalizable effect of the program on algebra achievement. We did find clear evidence that the effect was dependent on local conditions. For two teachers in one

school—selected for the study on the basis of experience with technology innovations—there was an impact. Many characteristics of these teachers, their students, school, or district that can be put forward explain the differential effectiveness of *HMH Fuse* in that district. The fact that the teachers reported using the application far more than other teachers is consistent with greater commitment to and experience with technology solutions. While we cannot generalize the results beyond these two teachers, the study is suggestive of approaches that may lead to success with applications such as *HMH Fuse*. It is notable that there is a positive effect on student attitudes toward math, and students with positive attitudes toward math achieve higher scores on the CST.

It is debatable whether the results for the focal district are exploratory or confirmatory. If we ask whether the focal district's special status was considered in the initial design, then the analysis can be considered post hoc. However, we planned the subgroup analysis before we knew the results for the overall average impact across the four districts so it was not planned as a check to see if we could find any subgroups where there was an effect. Certainly from the point of view of the district, the question was of central importance—the average impact was not as relevant. But for the sake of argument, we will imagine that the overall impact was the confirmatory result and that the results for the focal district were chanced upon and therefore exploratory.

Lessons from the Case Studies

Our goal in this paper is not to elevate exploratory analyses above the level of preliminary results that are useful in taking into the next research study or next practical application. But we believe it is worth raising questions about the special status afforded the confirmatory results as reliable final conclusions from the research.

In the case of *AMSTI*, the confirmatory results showed a substantial and, for the Alabama Department of Education, an educationally significant finding for math achievement (https://docs.alsde.edu/documents/55/NewsReleases2012/2-21-2012_AMSTI%20study%20results.pdf). In the conclusion of our report, we raise a number of limitations on this finding. In all cases, they address the generalizability of the finding: would it apply to schools that had not volunteered for AMSTI; would it apply to other states where different (possibly more rigorous) math programs are already in place. These are familiar cautions based limits to generalizability. The possibility that AMSTI may be ineffective for a subpopulation is not addressed among these limitations.

In the case of *HMH Fuse* without the additional findings from the one district, the published confirmatory conclusion would have been that the technology has no discernible effect. Instead, we report that there is at least one condition under which it is shown to have an impact.

Confirmatory findings from IES-sponsored RCT follow the same rules as employed by the What Works Clearinghouse (2008) in identifying the result from a study that is used in building the evidence available for an intervention. That is, the WWC usually makes use of only the primary finding rather than reporting also exploratory analyses that may be included in the report. If there is more than one primary finding, an adjustment for multiple comparisons is used. In all cases, the acceptability of a study's results is based on the study's internal validity—that is the validity of the logic behind the design and analysis. These considerations are very important and critiques and elimination of studies that do not meet this minimum criteria of design is appropriate.

But, we would argue, that IES does not go far enough in cautioning the reader concerning the unreliability of confirmatory findings. If we confirm an average impact but in exploratory analysis discover a plausible, policy-relevant, and statistically strong differential effects for subgroups, then some doubt about completeness may be cast on the value of the confirmatory finding. We may not be certain of the moderator effect but once it comes to light, the value of the average impact can also be doubted as incomplete or misleading. If it is necessary to conduct an additional experiment to verify a differential subgroup impact, the same experiment may verify that the average impact is not what practitioners, developers, and policy makers should be concerned with.

The two cases illustrate the point.

In the *AMSTI* experiment, including exploratory questions about the moderating effect of minority and economic status was not part of an arbitrarily long laundry list but rather called for by the subgroup accountability provisions of the No Child Left Behind Act. If the goal is to lift schools out of failing status in math and reading then a positive average effect that hides a differential by subgroup may fail because the subgroups that do not benefit will continue to make inadequate progress. This is not to claim that our experiment failed to confirm a positive average effect. But it is to say that it did not confirm that all important subgroups got close to the average.

In the *HMH Fuse* experiment, we confirmed that on average, there is no measurable impact of the iPad technology. Again, however, the result for a subgroup casts doubt on the value of the question that led to the average result. If the subgroup analysis, while plausible and statistically significant is considered exploratory and in need of confirmation in subsequent experiments, then the value of the average result for

practitioners, developers, and policy-makers also needs to be confirmed in further research.

Conclusion

Distinguishing between confirmatory and exploratory findings may tend to give practitioners, developers, and policy-makers greater confidence in the research results than are warranted. The criteria for conclusions about confirmatory questions are much more stringent and as a consequence some readers may conclude that the answers to such questions can be accepted as applicable to policy decisions more so than the answers to exploratory questions. But the stringent criteria are based on the internal logic of the experiment, not on its relevance to practical decisions.

We are proposing that any result from a school-based experiment should be treated as provisional by practitioners, developers, and policy-makers. The results of RCTs can be very useful but the challenges of generalizability of the results from even the most stringently designed experiment means that the results should be considered the basis for a hypothesis that the intervention may work under similar conditions.

For a developer considering how to improve an intervention, the specific conditions under which is appeared to work or not work is the critical information to have. For a school system decision-maker, insight into subpopulations that appear to benefit and the conditions that are favorable for implementation are the most useful pieces of information. For those concerned with educational policy, it is often the case that conditions and interventions change and develop more rapidly than research studies can be conducted. Using available evidence may mean digging through studies that have confirmatory results in contexts similar or different from their own and examining exploratory analyses that provide useful hints as to the most productive steps to take next. The practitioner in this case is in a similar position to the researcher considering the design of the next experiment. The practitioner also has to come to a hypothesis about how things work as the basis for action. In this context, the confirmatory analyses are no more or less useful than what has been indicated through exploratory research.

References

Newman, D., Finney, P.B., Bell, S., Turner, H., Jaciw, A.P., Zacamy, J.L., & Feagans Gould, L. (2012). *Evaluation of the Effectiveness of the Alabama Math, Science, and Technology Initiative (AMSTI)*. (NCEE 2012–4008). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

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.

Toby, M., Ma, B., Lai, G., Lin, L., & Jaciw, A. (2012, March). *Comparative Effectiveness of Houghton Mifflin Harcourt Fuse: Algebra 1: A Report of a Randomized Experiment in Four California School Districts.* (Empirical Education Rep. No. Empirical_HMH-6038-FR1-YR1_O.3). Palo Alto, CA: Empirical Education Inc.

What Works Clearinghouse. (2008). *Procedures and standards handbook* (version 2.1). Retrieved March 20, 2012, from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v2_1_standards_handbook.pdf

Citation for this paper: Newman, D., & Jaciw, A. (2012, April). In School Settings, Are All RCTs (Randomized Control Trials) Exploratory? In G. Stoker (Chair), *Current Studies in Program Evaluation to Improve Student Achievement Outcomes.* Roundtable session conducted at the annual meeting of the American Education Research Association, Vancouver, BC.

Contact:

Denis Newman
CEO
Empirical Education Inc.
Palo Alto, CA
dn@empiricaleducation.com
© 2012 Empirical Education Inc.