Making Sense of SCIENCE: The Challenges of Impacts and Assessments

Andrew P. Jaciw
Thanh Nguyen
Jenna Zacamy
Li Lin

Empirical Education Inc.

Making Sense of SCIENCE: The Challenges of Impacts and Assessments

Background.

With the introduction of the Next Generation Science Standards there is an urgency to evaluate impacts of programs aligned to the new Standards and to ensure that assessments validly reflect the core concepts and skills.

We investigate the impact of an NGSS-aligned program, Making Sense of SCIENCE (MSS) funded through the Investing in Innovation (i3) program. We report the main impact findings; however, our primary focus is on the challenge of assessment. That is, without a valid assessment, determination of “practical significance and meaningful effects”, both central to the theme of SREE 2020, cannot be achieved.

We faced the challenge that no assessments were available by the end of the study that adequately tapped the Depth Of Knowledge (DOK) and higher-order thinking skills that NGSS and MSS promote. Therefore, science content experts on the evaluation team, independently developed an assessment written to general specifications that drew on established previously operational items from multiple sources (e.g., NAEP and MOSART).

The test, rapidly developed, satisfied content needs, but presented challenges by introducing uncertainty in reliability, difficulty, and discriminability. At the same time, availability of item specific outcomes created a valuable opportunity to investigate the robustness of impact results to alternative approaches to scaling item responses. We focus on these aspects.

Research Questions

Confirmatory questions: What is the impact of MSS on student science achievement in 4th and 5th grade after two years of implementation compared to business as usual? What is the impact on students below the first tertile of incoming achievement?

Exploratory questions: Did impacts vary depending on state (Wisconsin versus California) and grade (4th vs 5th)?

Questions about the assessment: What were the test’s properties, including its difficulty, discriminability and reliability? How robust are impact findings to alternative methods to score scaling (e.g., use of Item Response Theory versus percent correct, and alternative approaches to handling missing responses).

Setting

The study was conducted in five districts in California and two in Wisconsin. The districts were a mixture of rural and urban.

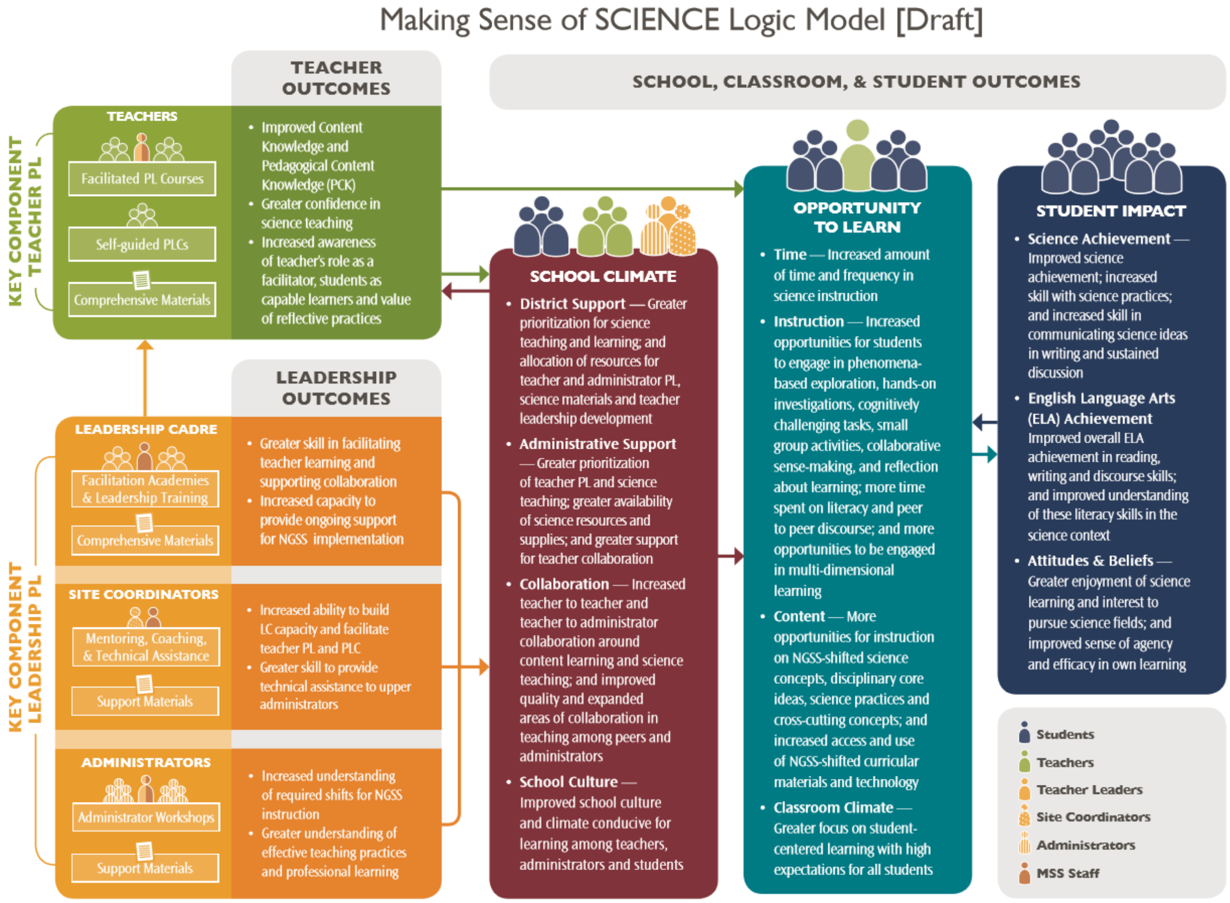
Participants

Impacts were assessed on 2,140 4th and 5th grade students in 55 schools. Student ethnicity was as follows: Asian (N=240, 11.2%), Black (N=318, 14.86%), Hispanic (N=881, 41.17%), White (N=448, 20.93%), Other and unknown (253, 11.80%).

Program

MSS is a PD program aimed at raising students’ science achievement through improving science instruction. It focuses on practices supporting the implementation of Next Generation Science Standards (NGSS). Professional learning activities for teachers include 30 hours of professional learning each summer for two summers and 12 hours of Professional Learning Communities (PLCs) each year. (See logic model in Figure 1, with key components at left).

Figure 1. Logic Model Making for Sense of SCIENCE


Making Sense of SCIENCE Logic Model [Draft]
TEACHER OUTCOMES
SCHOOL, CLASSROOM, & STUDENT OUTCOMES
KEY COMPONENT TEACHER PL
TEACHERS
Facilitated PL Courses
Self-guided PLCs
Comprehensive Materials
• Improved Content Knowledge and Pedagogical Content Knowledge (PCK)
• Greater confidence in science teaching
• Increased awareness of teacher's role as a facilitator, students as capable learners and value of reflective practices
LEADERSHIP OUTCOMES
KEY COMPONENT LEADERSHIP PL
LEADERSHIP CADRE
Facilitation Academies & Leadership Training
Comprehensive Materials
• Greater skill in facilitating teacher learning and supporting collaboration
• Increased capacity to provide ongoing support for NGSS implementation
SITE COORDINATORS
Mentoring, Coaching, & Technical Assistance
Support Materials
• Increased ability to build LC capacity and facilitate teacher PL and PLC
• Greater skill to provide technical assistance to upper administrators
ADMINISTRATORS
Administrator Workshops
Support Materials
• Increased understanding of required shifts for NGSS instruction
• Greater understanding of effective teaching practices and professional learning
SCHOOL CLIMATE
• District Support — Greater prioritization for science teaching and learning; and allocation of resources for teacher and administrator PL, science materials and teacher leadership development
• Administrative Support — Greater prioritization of teacher PL and science teaching; greater availability of science resources and supplies; and greater support for teacher collaboration
• Collaboration — Increased teacher to teacher and teacher to administrator collaboration around content learning and science teaching; and improved quality and expanded areas of collaboration in teaching among peers and administrators
• School Culture — Improved school culture and climate conducive for learning among teachers, administrators and students
OPPORTUNITY TO LEARN
• Time — Increased amount of time and frequency in science instruction
• Instruction — Increased opportunities for students to engage in phenomena-based exploration, hands-on investigations, cognitively challenging tasks, small group activities, collaborative sense-making, and reflection about learning; more time spent on literacy and peer to peer discourse; and more opportunities to be engaged in multi-dimensional learning
• Content — More opportunities for instruction on NGSS-shifted science concepts, disciplinary core ideas, science practices and cross-cutting concepts; and increased access and use of NGSS-shifted curricular materials and technology
• Classroom Climate — Greater focus on student-centered learning with high expectations for all students
STUDENT IMPACT
• Science Achievement — Improved science achievement; increased skill with science practices; and increased skill in communicating science ideas in writing and sustained discussion
• English Language Arts (ELA) Achievement Improved overall ELA achievement in reading, writing and discourse skills; and improved understanding of these literacy skills in the science context
• Attitudes & Beliefs — Greater enjoyment of science learning and interest to pursue science fields; and improved sense of agency and efficacy in own learning
Students
Teachers
Teacher Leaders
Site Coordinators
Administrators
MSS Staff

Research Design

The study randomized 60 schools to MSS or a two-year wait-list. Impacts on students on the main science assessment were assessed after two years. (Randomization was in winter 2016, with outcomes assessed in spring 2018.) Attrition of schools was low, with impact assessed on 2,140 students in classes of 174 teachers and 55 schools. Student and teacher joiners were allowed to enter the study to keep up the sample sizes, making this an RCT that will be reclassified as a QED for review under WWC 4.0.

Data Collection and Analysis

The assessments was administered online to students in Spring 2018. We estimated impact using 3-level Hierarchical Linear Models (HLMs) (student, cluster, randomized blocks), with individual student scores regressed against baseline covariates, a dummy variable indicating treatment assignment at the cluster level (MSS=1, control=0), and random effects at student, cluster, and block levels.

Results

Confirmatory Impact Findings

Impact of MSS on students is displayed in Table 1. We show results of progressively adding covariates to the model. The benchmark result (Model 9) shows no impact of MSS on student achievement after two years of implementation with standardized effects size .076 (*p*=.25). Baseline equivalence was achieved on both pretests: –.13 SD (*p*=.18) on the 3rd grade state ELA test, and –.07 SD (*p*=.48) on the math pretest. We also did not observe a significant effect on students below the first tertile, with effect size .085 (*p*=.158).

Table 1. Impact on Student Content Knowledge (n= 2,140, J= 55)

Effect	Model 1	Model 2	Model 3	Mode 4	Model 5	Model 6	Model 7	Model 8	**BENCHMARK Model 9**
Intercept	-0.059	0.028	-0.075	0.038	0.066	0.984**	-0.037	0.702**	**0.998****
In Treatment group	-0.010	0.053	-0.010	0.054	0.062	0.062	0.064	0.051	**0.074**
Effect size	-0.010	0.054	-0.010	0.055	0.064	0.064	0.066	0.052	**0.076**
Pretests		X		X	X	X	X	X	X
Grade 4					X				X
State CA			X	X	X	X	X	X	X
Other student covariates					X				X
Set 1 (teacher covariates)						X	X		X
Set 2 (teacher covariates)						X		X	X

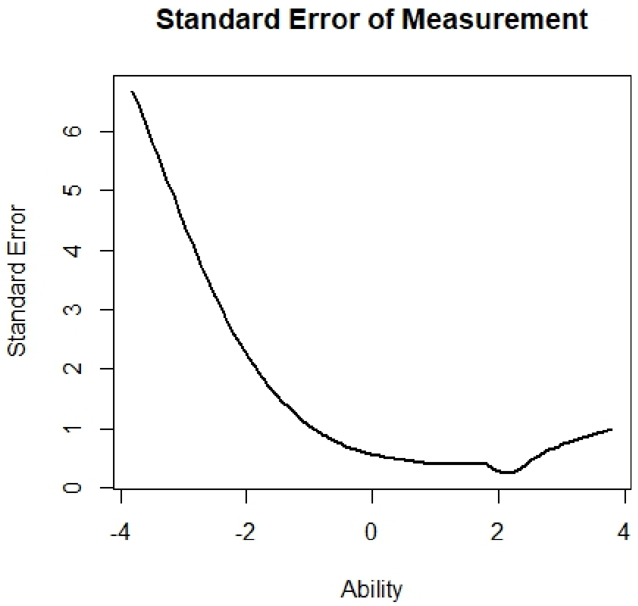
*p < .10 ; **p < .05 ; ***p < .01 ; ****p < .001; We do not show variance component estimates in this table.

Exploratory Questions. Fidelity Of Implementation (attendance at PD, PLC) was high, but just below required threshold. There was no difference in impact between WI and CA. We observed positive impact of the program in 5th grade (Effect size=.144, *p*=.045), but not in 4th grade (Effect size=-.032, p=.735). (Baseline equivalence on the pretest was achieved in each case.) We will report at SREE impacts depending on exposure, dosage and fidelity of implementation.

Questions about the Assessment.

We used a new assessment, well aligned to NGSS standards, and drawing on established items used in previous operational assessments. It was designed to allow students to demonstrate Depth Of Knowledge targeted by NGSS and MSS. The assessment was difficult (with averages of 42% and 38% in percent correct in 4th and 5th grades respectively) and with moderate reliability (Cronbach Alpha=.63). A test written to more challenging content standards was hard for lower achieving students, as supported by the test score standard error calculated across ability levels as displayed in Figure 2.

Figure 2. Standard error of the student science assessment

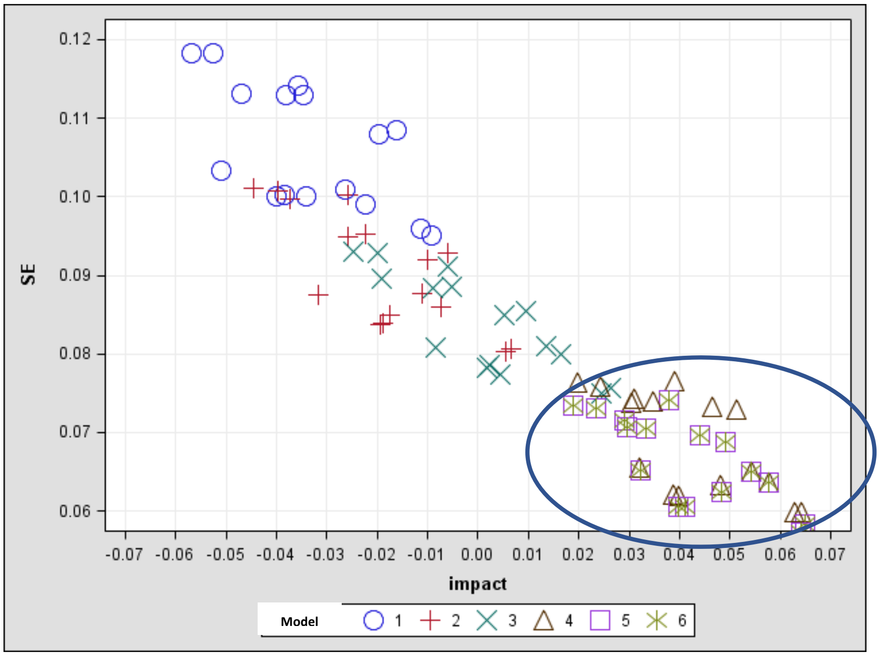


We addressed the general technical question of whether impact results depend on different approaches to scaling posttest scores. (This question is not normally asked, as testing companies typically provide ready scale scores.) Brevity of this abstract precludes full discussion, but for each of six families of impact models (Table 2) we examined impacts across 8 combinations of scaling factors (2 strategies for missing item responses × 2 ways to address items showing moderate Differential Item Functioning × 2 ways of scaling responses (3PL IRT calibration versus percent correct)), and 2 samples (including students in classes of teacher joiners or not). We focus on the oval in Figure 3, which shows that for each of Models 4, 5 and 6, **impacts ranged up to .05 standardized effect size units depending on how item responses were scaled**. This is an important finding given that impacts as small as .05 SD can be educationally meaningful (Newman et al., 2012).

Table 2. Characteristics of Six Basic Impact Models

Dimension	Description	Model					
		1	2	3	4	5	6
	Base: indicator of random assignment status, random cluster and student effects	X	X	X	X	X	X
Compositional	Including a pretest in analysis				X	X	X
	Including a block-level random effect		X	X		X	X
Structural	Using block random effects instead of cluster level random effects			X			X

Figure 3. Impacts and their standard errors with changes of structural, compositional and scaling factors of impact models



Conclusions

This work reminds us that impacts of programs cannot be separated from their contexts (Cronbach, 1975, 1982) including the assessment used. The attempt to evaluate deeper knowledge in this experiment led to a difficult test that may have restricted the range over which valid responses could be measured (Shadish, Cook and Campbell, 2002). Yet developing a NGSS-responsive assessment was critical. It is hard to establish practical significance of effects in a shifting educational landscape where science content, assessment of its learning, and educational policy driving it, are evolving.

References

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, 116-127.

Cronbach, L. J., (1982). *Designing evaluations of educational and social programs.* San Francisco, CA: Jossey-Bass.

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.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi experimental designs for generalized causal inference.* Boston: Houghton Mifflin.