Collaborative Research: Leveraging Comparison and Explanation of Multiple Strategies (CEMS) to Improve Algebra Learning

Jon R. Star, Bethany Rittle-Johnson, and Kelley Durkin

Advisory Board Project Update (February 28, 2020 & March 27, 2020)

Overview from Grant Proposal:

Productive learning of algebra is supported by reflection on multiple solution strategies through comparison and explanation of the reasons behind the strategies (***C**omparison and **E**xplanation of **M**ultiple **S**trategies:* **CEMS**). Existing theories of algebra learning focus on building conceptual knowledge and place less emphasis on how students gain expertise with symbolic strategies. Working with symbolic strategies is essential in algebra learning. Students need to develop procedural flexibility - knowing multiple strategies for solving a problem and selecting the most appropriate strategy for a given problem - and understand the conceptual rationale behind commonly used strategies. Knowledge of strategies (procedural knowledge) supports gains in both procedural flexibility and conceptual knowledge of algebra (Schneider, Star & Rittle-Johnson, 2011). In small-scale studies, redesigning lessons on equation solving to integrate a CEMS approach supported greater procedural knowledge, flexibility and/or conceptual knowledge than completing the lessons without a CEMS approach (Rittle-Johnson & Star, 2007, 2009; Rittle-Johnson, Star, & Durkin, 2009, 2012; Star & Rittle-Johnson, 2009). A preliminary set of supplemental materials to support a CEMS approach across the Algebra I curriculum was previously developed, with evidence that classroom teachers can implement the materials with good fidelity (Star, Pollack, et al., 2015). The current project seeks to build upon and improve these materials and the professional development opportunities that accompany them.

Feedback from Advisory Board:

- Focus on Year 3 results
    - Seeking your guidance on shaping a primary paper that reports on the overall study in Year 3, and additional papers we may consider.
        - Most important additional coding and analyses
        - How to package different findings together
    - Our results are complicated because the schools and students in treatment and control conditions were not equivalent at pretest on several dimensions.

Year 1, 2016-2017: Development and Piloting

- We worked with 3 teachers to refine our existing CEMS materials, to integrate the materials into their curriculum, and to validate outcome measures that assess multiple types of knowledge (e.g., procedural flexibility, conceptual knowledge, and procedural knowledge).

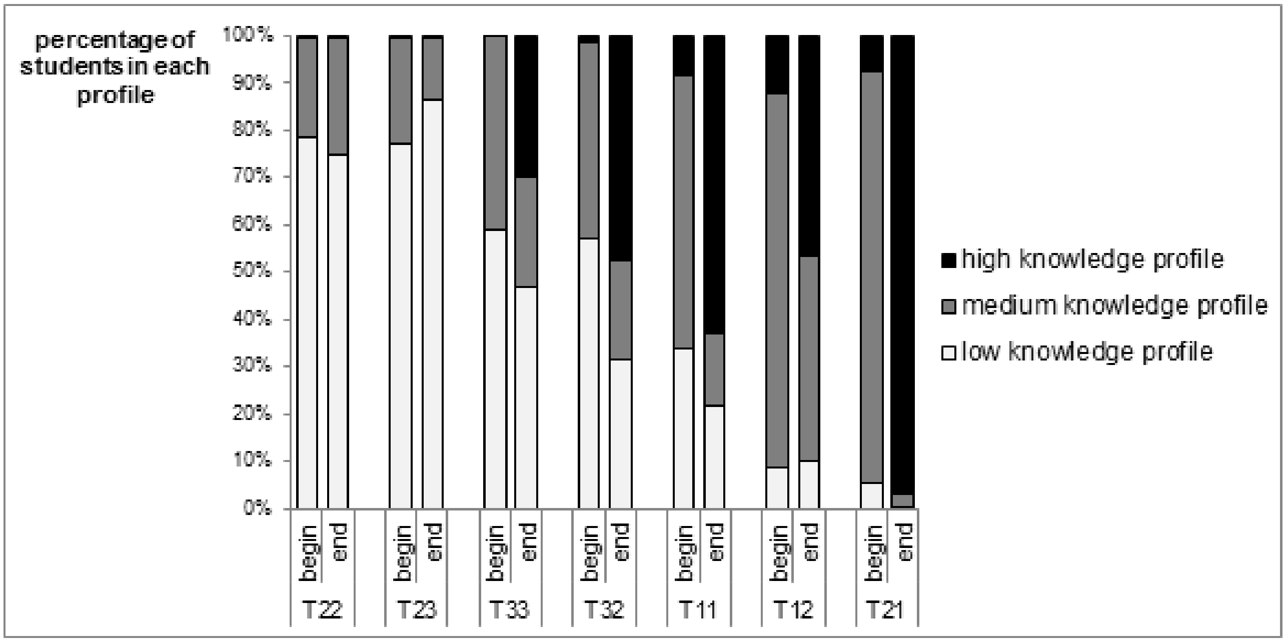
- We revised and expanded materials for 5 Algebra I topics that occur throughout the school year. *See example materials in the Appendix at the end of this document.*

Year 2, 2017-2018: First Implementation Year

- Nine teachers were supported in using our materials. Treatment teachers attended a week-long summer PD session and received "Just In Time" PD sessions before each topic to reflect on past work with the materials. Teachers were asked to use a CEMS approach within 8-9 lessons for each of the target units, although most teachers did not get to the final unit, and many teachers only began the 4th target unit.
- A total of 585 students (348 treatment, 237 control) participated. 9 treatment teachers across 3 school districts and 10 control teachers at 1 large school serving multiple communities participated. The data use agreement was delayed for control teachers, so overall pretest and Topic 1 & 2 data were not available for control classrooms, and we do not have a strong baseline knowledge measure because treatment and control students were from different states, so there is no common state assessment that can be used.
- As will be reported at AERA 2020: We explored variability in students' algebra learning and instructional features that predict learning when using our approach. We focused on the 7 ninth-grade teachers from 3 schools and their 315 students who used our CEMs approach as part of their instruction during a full-year Algebra I course (dropping the teacher who taught a remedial version of the course). Data included a researcher-developed algebra assessment, coding of videos of classroom lessons, and teacher logs. Latent transition analysis (LTA) was used to identify student knowledge profiles on the algebra assessment at the beginning of the school year and profile transition from the beginning to the end of the school year. Then, we explored variability between teachers in their students' initial knowledge profile and profile transitions and evaluated if 3 instructional features predicted this variability.

  Results. Three student knowledge profiles were identified in the LTA: students with a low, medium and high level of knowledge. There was large variability among teachers in their students' initial knowledge level as well as in the probability to transition to a higher knowledge profile on the end-of-year assessment (see Figure 1). We explored instructional features that could explain this variability. The higher teachers' use of our materials and the more teachers facilitated high-quality student interactions, the more likely their students were to have a higher knowledge profile at the beginning of the school year (Likelihood Ratio χ^2 (2) = 17.08, $p < .001$ and χ^2 (2) = 21.93, $p < .001$, respectively) and to transition to a higher-knowledge profile at the end of the school year (χ^2 (2) = 6.20, $p = .045$ and χ^2 (2) = 18.77, $p < .001$, respectively).

  Figure 1: Percentage of students in each knowledge profile at the beginning and at the end of the school year, by teacher



Note: T stands for Teacher, the first digit indicates the school number and the second digit indicates the teacher number at the school. Teachers are ordered by proportion of students in the low knowledge profile at the beginning of the school year.

Significance. Greater use of our *CEMS* approach was related to greater knowledge gains, providing preliminary support for the effectiveness of the approach, albeit with a small number of teachers. Greater support for high-quality student interaction was also associated with greater knowledge gains, highlighting the importance of students explaining ideas with classmates. However, some teachers struggled to implement our approach and some students did not learn much of our target content, especially in classrooms with many students with low initial knowledge, suggesting that our CEMS approach and teacher PD was not sufficiently powerful to aid learning by all students. The current findings highlight the potential of evidence-based instructional approaches for improving student learning, as well as persistent gaps in improving teaching quality and student learning broadly.

Year 3, 2018-2019: Treatment vs. Control Teachers with Baseline Scores

Participants

Students

- A total of 1082 students (573 treatment, 509 control) participated.
    - We continue to work on gathering demographic information at the student level from districts. However, we will not be able to get student- or class-level demographic data from several of the schools (treatment and control).

Teachers

- A total of 30 teachers participated.
    - The treatment group included 16 teachers from 4 schools, including 7 teachers who were treatment teachers in Year 2 of the study.
    - The control group was composed of 14 teachers from 6 schools.
    - Two teachers taught 8th grade (1 treatment and 1 control, with 20 and 23 students, respectively) and 28 teachers taught 9th grade, and their classes covered a wide range of student ability levels.

2018-2019 Teacher Characteristics

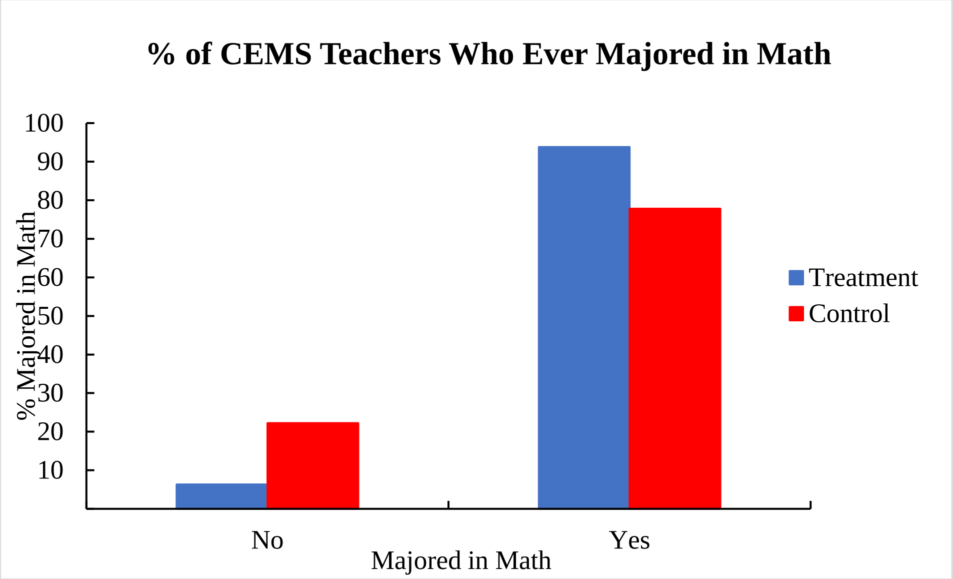
- As part of the summer PD sessions, teachers were asked to complete short background surveys with information about their education, teaching experience, and beliefs about and utilization of multiple strategies and discussion in their algebra classrooms.
    - *Note.* We are missing teacher survey data from 5 control teachers.

Teacher Experience

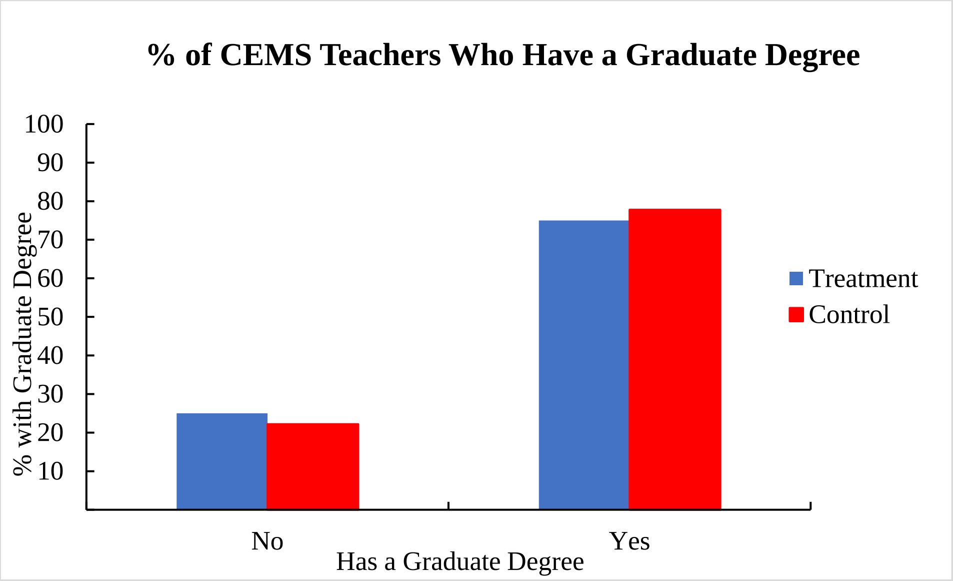
	Treatment (N = 16)				Control (N = 9)			
Teaching Experience	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Years Teaching	1.0	28.0	11.1	8.4	2.0	18.0	9.3	5.5
Years Teaching Middle and/or High School Math	1.0	28.0	10.7	8.3	2.0	18.0	9.3	5.5
Years Teaching Algebra	1.0	28.0	9.4	7.6	2.0	12.0	6.7	3.3

Analysis: Independent samples t-tests were performed to compare teachers' reported teaching experience by study condition. The test results suggest that the treatment and control teachers do not significantly differ in terms of their number of years teaching, *t*(23) = 0.6, p = .573, their experience teaching middle and/or high school math, *t*(23) = 0.4, p = .665, or the number of years they taught an algebra course, *t*(23) = 1.0, p = .352.
Findings: There were no significant differences between treatment and control teachers' reported teaching experience.

Teacher Education



Analysis: A chi-square test of independence was performed to examine the relationship between majoring in math and study condition. The test results suggest that the relationship between these two variables is not significant, $X^2(1, N = 25) = 1.4, p = .238$. *Findings: The proportion of teachers who reported ever majoring in math did not significantly differ by condition.*



Analysis: A chi-square test of independence was performed to examine the relationship between having a graduate degree and study condition. The test results suggest that the relationship between these two variables is not significant, $X^2(1, N = 25) = 0.0, p = .876$. *Findings: The proportion of teachers who reported having a graduate degree did not significantly differ by condition.*

Teacher Use of Multiple Strategies and Discussion During Algebra Lessons

- As part of the background survey, teachers answered several questions regarding their use of multiple strategies and discussion during their algebra lessons. Text responses were converted into the following ordinal scale:
    - 0 = Never
    - 1 = Less than once a month
    - 2 = 1-3 times per month
    - 3 = 1-2 times per week
    - 4 = 3-4 times per week
    - 5 = Every day

	Treatment (N = 16)				**Control (N = 9 of 14)**			
	Min	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
How often did students see multiple ways to solve the same math problem in your Algebra I class?	1.0	5.0	3.81	0.98	2.0	5.0	3.78	0.83
How often did students consider common errors or incorrect ways to solve problems in your Algebra I class?	2.0	5.0	3.31	1.01	2.0	5.0	3.78	1.09
How often did you engage your Algebra I students in a whole class mathematical discussion?	2.0	5.0	3.44	1.21	2.0	5.0	3.78	1.3
How often did you have students get in pairs or small groups to share their mathematical thinking with each other in your Algebra I class?	1.0	5.0	4.06	1.18	0.0	5.0	2.89	1.96

Analysis: Independent samples t-tests were performed to compare treatment and control teachers' reported use of multiple strategies and discussion during their Algebra lessons. Results suggest that treatment and control teachers did not significantly differ in terms of their reported use of multiple solution methods (*t*(23) = .01, p = .930) or common errors (*t*(23) = -1.1, p = .295) during their Algebra lessons. Likewise, treatment and control teachers did not differ significantly in their reported use of whole class mathematical discussions (*t*(23) = -0.7, p = .517) or their use of pairs or small group work (*t*(11.3) = 1.6, p =.130) during their Algebra lessons.

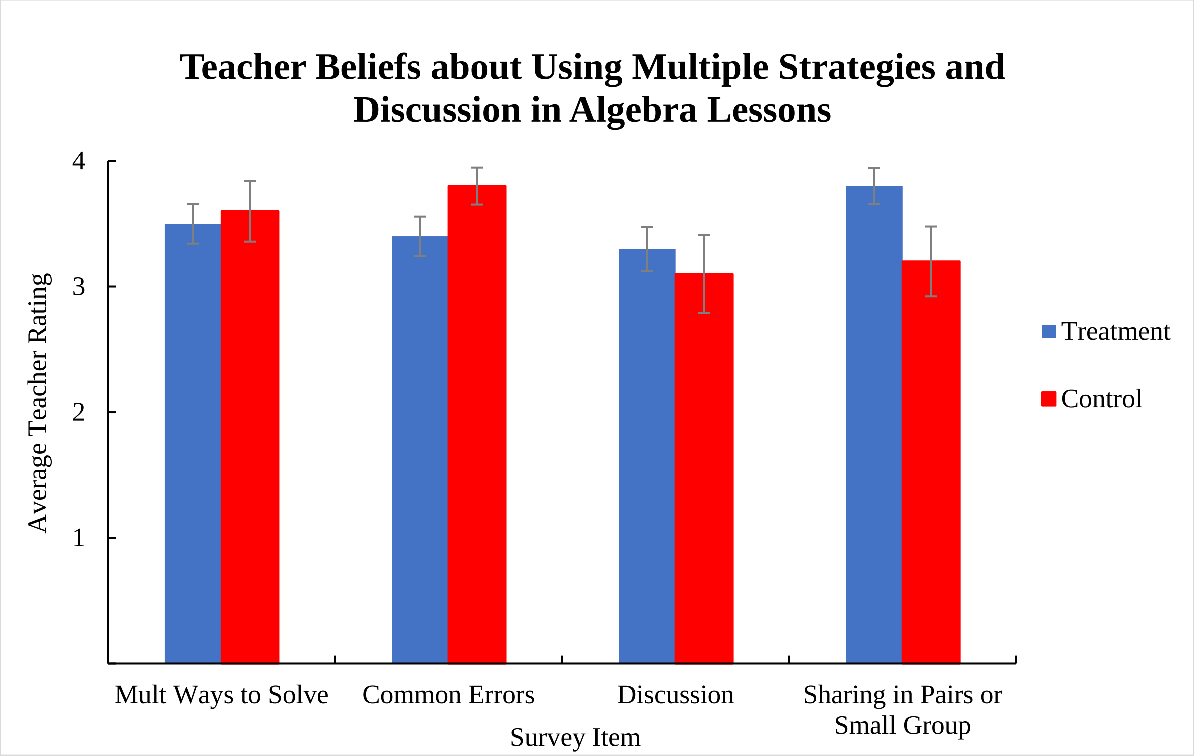
Findings: There were no significant differences by condition in teachers' reported usage of multiple strategies and discussion during Algebra lessons.

Note: Our sample of teachers did not differ much from a nationally representative sample of high school math teachers in a 2018 National Survey of Science and Mathematics Report (2018 NSSME+: Status of High School Mathematics) in terms of:

1. How many said they engaged in a whole class discussion at least once per week (76% of teachers in our sample vs. 84% in the NSSME report).
2. How many said they had students work in small groups (80% in our sample vs. 71% in the NSSME report).
3. In the NSSME report, 54% of teachers said they compared and contrasted different strategies in terms of their strengths and limitations at least once per week.

Teacher Beliefs about Multiple Strategies and Discussion

- As part of the background survey, teachers were asked to rate the following statements on a scale of 1 to 4, with 1 meaning ‘not very important at all’, and 4 meaning ‘very important’.
    - Q1: How important do you think having students see multiple ways to solve the same math problem is for their Algebra I learning?
    - Q2: How important do you think having students consider common errors or incorrect strategies is for their Algebra I learning?
    - Q3: How important do you think having students engage in whole class mathematical discussions is for their Algebra I learning?
    - Q4: How important do you think having students share their mathematical thinking with each other in pairs or small groups is for their Algebra I learning?



Teacher Beliefs about Using Multiple Strategies by Condition.
Note. Error bars represent standard error.

Analysis: Independent samples t-tests were performed to compare treatment and control teachers’ reported beliefs about the use of multiple strategies and discussion during their Algebra lessons. Results suggest that treatment and control teachers did not significantly differ in terms of how important they think it is to (1) present multiple solution methods (*t*(23) = -.2, p = .843), (2) have students consider common errors or incorrect strategies (*t*(21.7) = -1.6, p = .128), (3) have whole class mathematical discussions (*t*(23) = .6, p = .546), or (4) have students share thinking pairs or small groups (*t*(23) = 1.9, p = .074). *Findings: There were no significant differences by condition in teachers’ reported beliefs about the use of multiple strategies and discussion during Algebra lessons.*

Schools

- Participants were spread across 10 schools: 4 were in the treatment condition, and 6 were in the control group.
  - Participants in the treatment groups were enrolled in 3 schools in Massachusetts and 1 school in New Hampshire.
  - Control participants were in 3 schools in Massachusetts and 3 schools in New Hampshire.

2018-2019 School-Level Demographic Data

	Treatment Group (N = 4)			Control Group (N = 6)		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**
Attendance Rate	95.8	94.4	97.9	92.6	91.2	94.6
% ELL[1]	3.8	0.6	13.0	6.6	1.4	12.6
% Free/Reduced Lunch	17.2	5.8	39.1	34.5	9.6	46.6
Suspensions/Expulsions[2]						
% In-School Suspensions	0.7	0.0	2.1	4.0	0.7	10.0
% Out-of-School Suspensions	1.9	1.6	2.3	5.6	2.7	9.3
% Expulsions	0.0	0.0	0.0	0.0	0.0	0.1
SAT Scores[2]						
Reading/Writing SAT	562	505	601	537	501.5	610
Math SAT	574	520	609	532	503	602
Ethnicity						
% African American	5.4	0.9	16.3	5.7	3.5	13.6
% Asian	8.1	1.9	15.0	8.2	2.2	10.6
% Hispanic*	6.4	3.4	14.3	25.9	6.4	45.2
% White	77.1	50.4	89.7	57.4	31.5	74.6
% Native American	0.2	0.0	0.5	0.1	0.0	0.2
% Multi-Race, Non-Hispanic	2.8	2.0	3.7	2.7	1.1	3.8
% Native Hawaiian, Pacific Islander	0.0	0.0	0.1	0.1	0.0	0.2

* $p < .05$

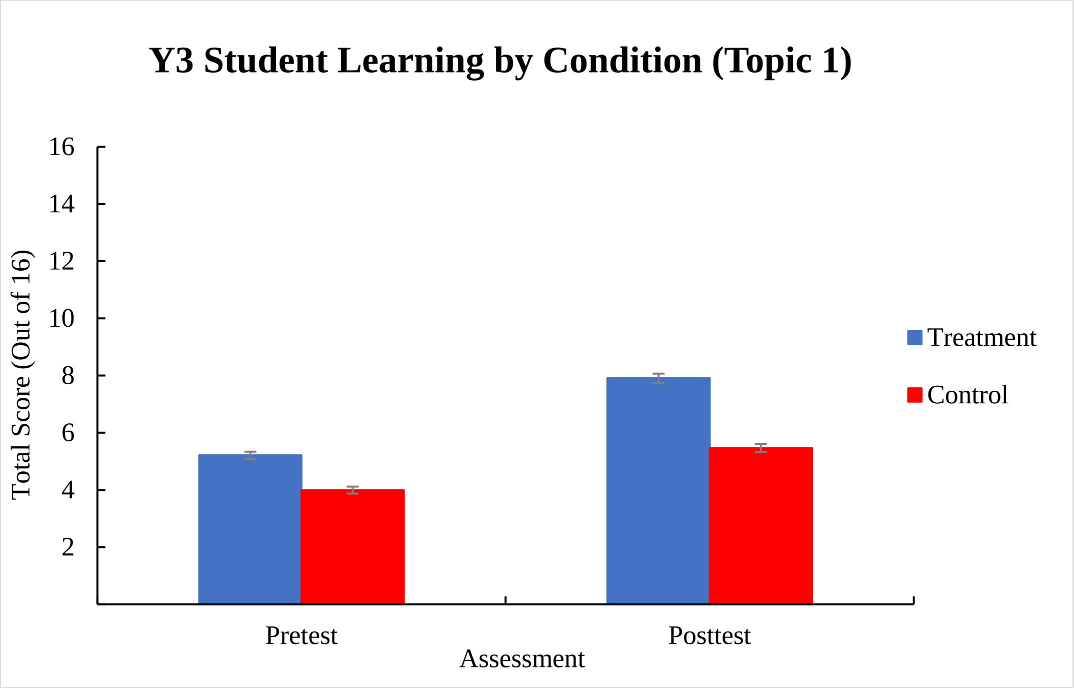
Note[1]. 3 control schools are not included in the %ELL calculations because they report an average number of ELL students instead of a percent.
Note[2]. 1 treatment school tracked school safety incidents broadly and did not break apart those incidents by in-school suspensions, expulsions, etc. Those data are excluded from the table.
Note[3]. 2 schools (1 treatment, 1 control) were middle schools and did not report SAT scores.

Analysis: Independent samples t-tests were performed to compare school-level demographics at treatment and control schools. Results suggest that there are not significant differences between treatment and control schools in terms of Free/Reduced Lunch status, $t(8) = -1.9$, p = .090, Reading/Writing SAT scores, $t(6) = 0.8$, p = .477, Math SAT scores, $t(6) = 1.3$, p = .237, or the percent of white students enrolled, $t(8) = 1.9$, p = .091. However, results suggest that there are between-group differences in terms of the percentage of Hispanic students enrolled, $t(8) = -2.9$, p = .019.

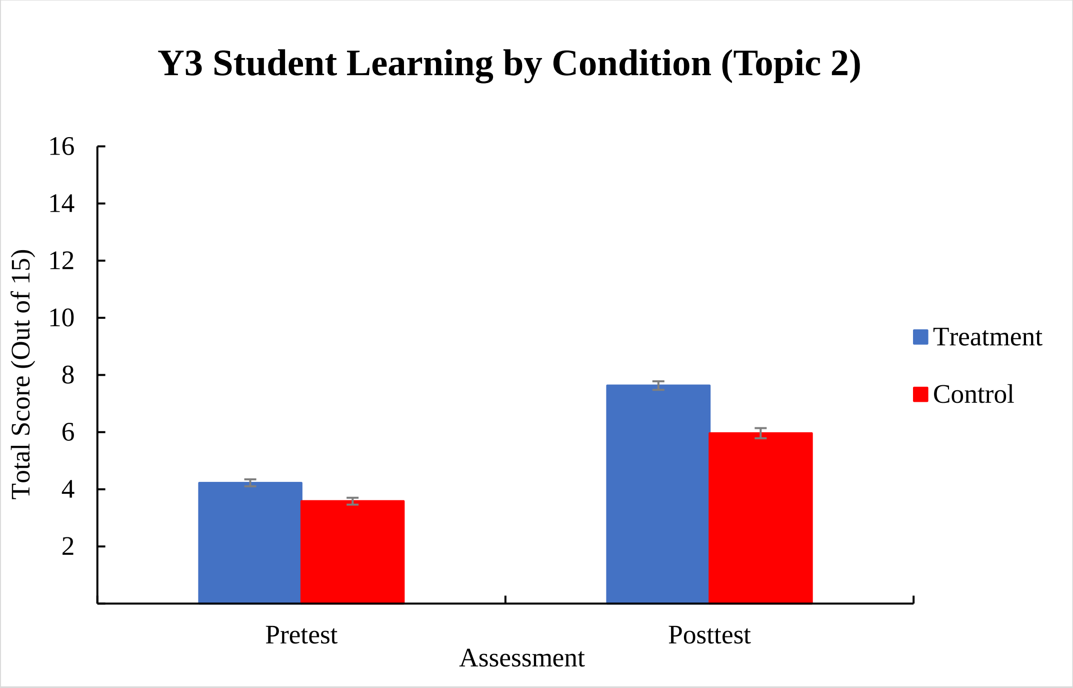
Findings: Treatment and control group schools have a significantly different percentage of Hispanic students enrolled, but do not differ significantly on the other demographic variables that were tested.

Assessment Data

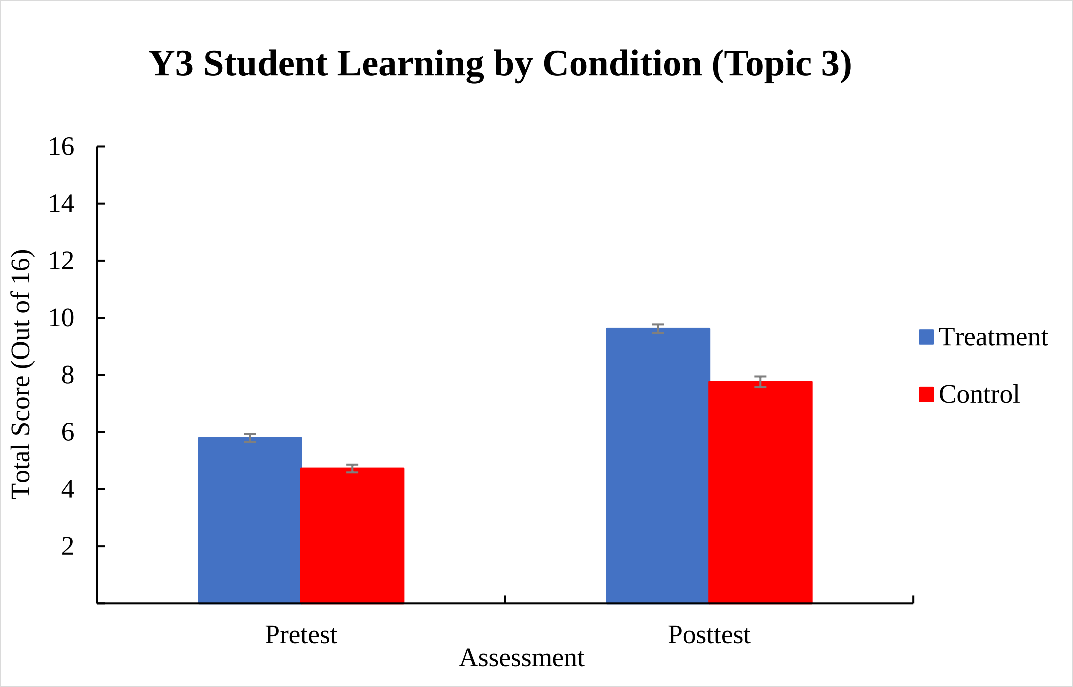
- We control for pretest scores in all analyses.
- Too few control teachers taught Topic 5 for it to be considered. Only 3 control teachers and 11 treatment teachers covered Topic 5.



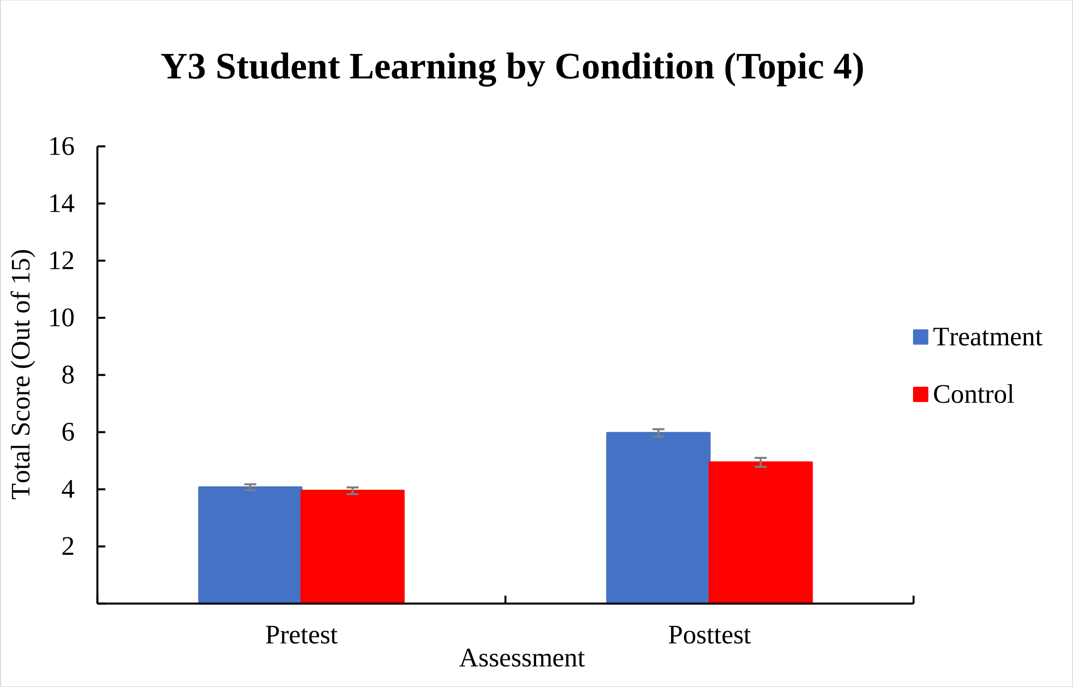
Mean Total Scores at Topic 1 Pretest and Posttest by Condition (2018-2019).
Note. Error bars represent standard error. Only students with complete data for Topic 1 (pre/post) were included (475 in the treatment group, and 359 in the control group).
Analysis: Multilevel models were run nesting students within section and within classroom and controlling for school-level demographics and pretest (when appropriate). There were no significant differences between treatment and control students at pretest, B = 0.09, p = .899. Treatment students outperformed control students at posttest, B = 1.29, p = .05.
Findings: Treatment students outperformed control students at posttest.



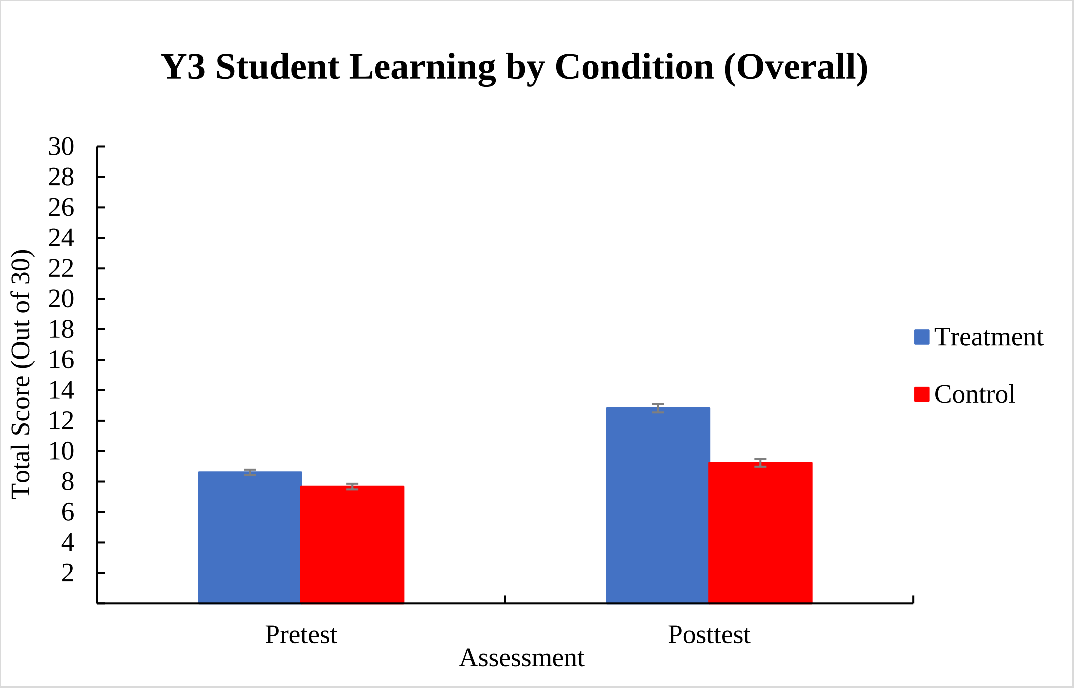
Mean Total Scores at Topic 2 Pretest and Posttest by Condition (2018-2019).
Note. Error bars represent standard error. Only students with complete data for Topic 2 (pre/post) were included (447 in the treatment group, and 321 in the control group).
Note. Item #9 was dropped from our analyses due to a printing error.
Analysis: Multilevel models were run nesting students within section and within classroom and controlling for school-level demographics and pretest (when appropriate). There were no significant differences between treatment and control students at pretest, B = 0.09, p = .899, or posttest, B = -1.09, p = .198.
Findings: There were no significant differences between conditions for Topic 2.



Mean Total Scores at Topic 3 Pretest and Posttest by Condition (2018-2019).
Note. Error bars represent standard error. Only students with complete data for Topic 3 (pre/post) were included (449 in the treatment group, and 325 in the control group).
Analysis: Multilevel models were run nesting students within section and within classroom and controlling for school-level demographics and pretest (when appropriate). There were no significant differences between treatment and control students at pretest, B = -0.10, p = .921, or posttest, B = 0.54, p = .619.
Findings: There were no significant differences between conditions for Topic 3.

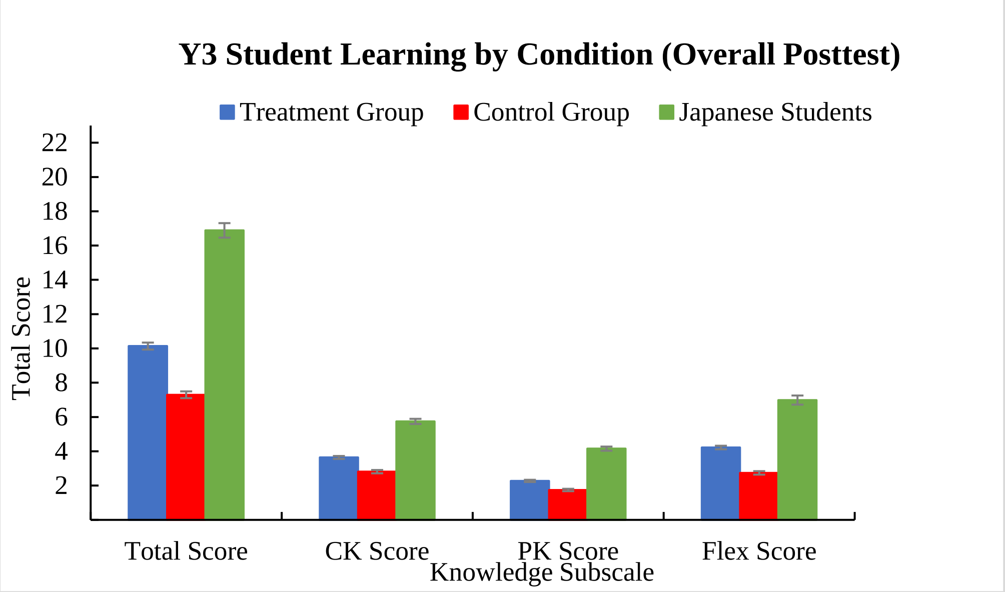


Mean Total Scores at Topic 4 Pretest and Posttest by Condition (2018-2019).
Note. Error bars represent standard error. Only students with complete data for Topic 4 were included (390 in the treatment group, and 310 in the control group).
Analysis: Multilevel models were run nesting students within section and within classroom and controlling for school-level demographics and pretest (when appropriate). There were no significant differences between treatment and control students at pretest, B = -0.13, p = .891, or posttest, B = 0.47, p = .722.
Findings: There were no significant differences between conditions for Topic 4.



Mean Total Scores at Overall Pretest and Posttest by Condition (2018-2019).
Note. Error bars represent standard error. Only students with complete data for the Overall assessment were included (431 in the treatment group, and 289 in the control group).
Analysis: Multilevel models were run nesting students within section and within classroom and controlling for school-level demographics and pretest (when appropriate). There were no significant differences between treatment and control students at pretest, B = -1.43, p = .371, or posttest, B = 1.65, p = .404.
Findings: There were no significant differences between conditions Overall.

Comparing CEMS Participants' Performance with Japanese Students



Note. The Total score is out of 23 possible points. The CK Score is out of 8 possible points, the PK Score is out of 5 possible points, and the Flex Score is out of 10 possible points.

Mean Total, Conceptual Knowledge, Procedural Knowledge, and Flexibility Scores at Overall Posttest by Condition (2018-2019).
Note. Error bars represent standard error. Only students with complete data were included (463 in the treatment group, 316 in the control group, and 78 Japanese students).
Analysis: Regression analyses were performed to compare the performance of CEMS' participants with the Japanese students. Results are presented in the table below.

	Standardized Coefficients	Unstandardized Coefficients			
Variable	**Beta (β)**	**B**	**SE**	**t**	***p***
Total Score					
(Constant)		16.89	0.46	36.97	0.000
Treatment Condition	-0.70	-6.75	0.49	-13.67	0.000
Control Condition	-0.96	-9.59	0.51	-18.80	0.000
Conceptual Knowledge Score					
(Constant)		5.74	0.20	28.78	0.000
Treatment Condition	-0.54	-2.10	0.22	-9.74	0.000
Control Condition	-0.73	-2.93	0.22	-13.15	0.000
Procedural Knowledge Score					
(Constant)		4.15	0.15	28.29	0.000
Treatment Condition	-0.65	-1.88	0.16	-11.84	0.000
Control Condition	-0.80	-2.41	0.16	-14.70	0.000
Flexibility Knowledge Score					
(Constant)		6.99	0.24	29.40	0.000
Treatment Condition	-0.57	-2.77	0.26	-10.77	0.000
Control Condition	-0.85	-4.25	0.27	-16.00	0.000

Findings: Japanese students outperformed students in the treatment and control groups on the Total score and on all knowledge subscales (Conceptual, Procedural, and Flexibility).

Alpha Reliability for Y3 Student Assessment Data

Assessment	N of Items	Student N	Cronbach's α
Overall Pretest	30	975	.595
Overall Posttest	30	779	.806
Topic 1 Pretest	16	891	.597
Topic 1 Posttest	16	908	.744
Topic 2 Pretest	15	898	.544
Topic 2 Posttest	15	849	.721
Topic 3 Pretest	16	852	.629
Topic 3 Posttest	16	829	.747
Topic 4 Pretest	15	788	.390
Topic 4 Posttest	15	743	.609
Topic 5 Pretest	12	382	.396
Topic 5 Posttest	12	317	.612

Measures of Instructional Practices

We recorded around 3 lessons for each teacher for each topic they covered. For all teachers, we used a General Fidelity Coding Scheme on a subsample of their videos to determine whether they exposed students to multiple strategies, compared strategies, engaged students in partner/small group work, or had whole-class discussions. For treatment teachers, we also used a coding scheme that measured the quality of their instruction when using our materials.

CEMS Y3 General Fidelity Coding Summary:

	Condition	Topic	GF_MS1a	GF_MS1b	GF_MS1c	GF_SG2	GF_Dis3
Topic 1	Treatment	1	100%	100%	97%	90%	83%
	Control	1	8%	4%	0%	42%	12%
Topic 2	Treatment	2	100%	100%	84%	81%	81%
	Control	2	29%	25%	4%	25%	0%
Topic 3	Treatment	3	100%	100%	84%	77%	81%
	Control	3	24%	16%	0%	24%	8%
Topic 4	Treatment	4	100%	100%	85%	69%	92%
	Control	4	25%	25%	0%	17%	8%
Total	Treatment	Total	100%	100%	87%	81%	82%
	Control	Total	20%	15%	1%	29%	7%

Description of General Fidelity Codes:

1a. Were students exposed to multiple strategies?

1b. If students were exposed to multiple strategies, were the strategies presented side-by-side?

1c. If students were exposed to multiple strategies, did the teacher or students compare the multiple strategies for at least a 1.5-minute continuous block?

2. Did all students engage in partner or small group work focused on math content for at least a 1-minute continuous block?

3. Was there a whole-class discussion for at least a 1.5-minute continuous block?

- Discussion included the following: (a) teacher is asking conceptual or open-ended questions and more than one student is responding to the questions (multiple students do not have to answer the same question) and/or (b) teacher is redirecting conversation by following up on a student's response to ask another student to respond to the same question or to the previous student's idea.

Findings: Treatment teachers were much more likely to engage in all of these practices. Engaging students in a whole-class discussion for at least a 1.5-minute block positively predicted students' posttest scores, even after controlling for pretest score and school-level demographics, B = 3.92, p = .038.

CEMS Y3 Treatment Coding Summary (ratings on a scale from 1 to 4, 4 being highest):

	Making Sense of Procedures	***Supporting Procedural Flexibility***	***Teacher Questioning***	***Student Responses***	***Opportunities for Interaction***
Topic 1 Average	2.5	3.1	3.1	2.8	1.5
Topic 2 Average	3.1	1.5	2.9	2.5	1.7
Topic 3 Average	2.3	2.5	3.0	2.8	1.7
Topic 4 Average	2.8	1.8	2.9	2.8	2.0
Topic 5 Average	2.8	2.8	3.0	2.7	1.6
Overall Average	**2.7**	**2.3**	**3.0**	**2.7**	**1.7**

Description of Treatment Codes:

Making Sense of Procedures: intended to capture the extent that the teacher's explanations and/or questions are intended to push students toward making sense of procedures and strategies in the WEP portion of the lesson and refers to deliberate actions that the teacher takes

Supporting Procedural Flexibility: intended to capture the extent to which teachers present procedures and strategies such that students had the opportunity to develop procedural flexibility, particularly focusing on multiple strategies and working with students to consider which strategies to use on certain problems, and this code focuses on the actions that the teacher takes in support of procedural flexibility

Teacher Questioning: intended to capture the extent that the teacher (via questioning) creates an opportunity for students to engage in deep and sustained mathematical thinking

Student Responses: intended to capture the extent that the classroom environment created by the teacher is one where students feel comfortable expressing themselves and *that a variety of students do so* – that students are inspired to contribute in response to mathematical questions from the teacher

Opportunities for Student Interaction: intended to assess the degree to which the teacher creates a classroom environment where students begin engaging in mathematical talk with each other and not only with the teacher

Findings: When using our materials, teachers generally had higher levels of making sense of procedures, teacher questioning, and student responses. Supporting procedural flexibility was most supported by our Which-is-better comparison type and understandably wasn't always seen in other comparison types with a different goal. It was difficult to raise teachers' opportunities for student interaction. Supporting procedural flexibility marginally positively predicted students' posttest scores, B = 3.28, p = .085. Higher student responses positively predicted students' posttest scores, B = 5.47, p = .04.

WEP Usage Data

The following table reports the number of WEPs used by treatment teachers per topic. For teachers with multiple sections (11, 12, 13, 21, 23, 41, 48), an average across their sections is reported.

Teacher ID	# Topic 1 WEPs (9)	# Topic 2 WEPs (8)	# Topic 3 WEPs (9)	# Topic 4 WEPs (9)	# Topic 5 WEPs (7)	# Total WEPs (42)	Avg WEP Duration (minutes)
11	6	8	8	8	0	**29**	**14.1**
12	8	6	9	8	0	**31**	**15.9**
13	6	5	0	0	0	**11**	**33.9**
21	9	8	9	9	1	**36**	**14.5**
22	9	8	9	2	0	**28**	**18.7**
23	5	2	1	0	0	**8**	**15.2**
31	8	8	9	8	7	**40**	**21.2**
41	8	7	5	4	4	**28**	**17.5**
42	8	7	6	4	3	**28**	**15.9**
43	8	6	4	3	3	**24**	**24.6**
44	7	7	6	3	5	**28**	**14.0**
45	7	7	5	4	5	**28**	**16.6**
46	8	7	6	4	5	**30**	**18.8**
48	8	6	5	4	5	**28**	**17.2**
410	8	7	5	4	4	**28**	**16.1**
411	8	7	3	4	4	**26**	**15.7**
Average*	**7**	**6**	**6**	**5**	**3**	**28**	**16.7**

*T13 only intended to cover Topics 1 and 2 in her course, as it is designed for struggling students. Teacher is excluded from topic 3-5 and total numbers.

The following table reports the number of WEPs of each type used by treatment teachers. For teachers with multiple sections (11, 12, 13, 21, 23, 41, 48), an average across their sections is reported.

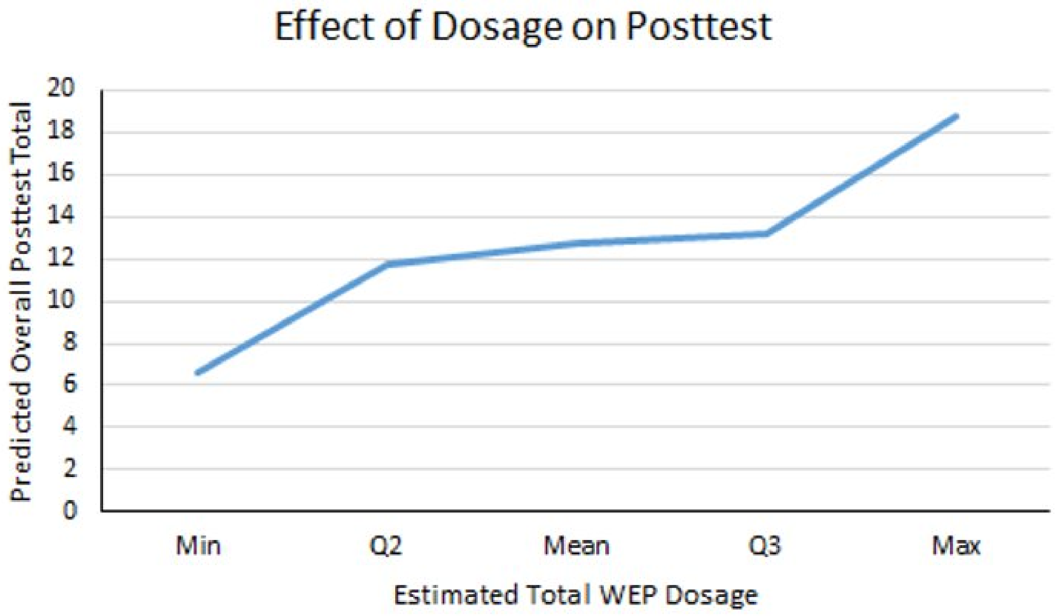
Teacher ID	# Why does it work? (13)	# Which is better? (16)	# Which is correct? (10)	# How do they differ? (3)	Total # WEPs (42)
11	10	8	8	3	**29**
12	10	11	8	3	**31**
13	3	2	5	1	**11**
21	11	13	9	3	**36**
22	10	10	6	2	**28**
23	4	2	1	1	**8**
31	13	14	10	3	**40**
41	7	12	8	1	**28**
42	9	12	6	1	**28**
43	5	12	6	1	**24**
44	7	15	5	1	**28**
45	7	13	6	2	**28**
46	8	13	8	1	**30**
48	7	13	6	2	**28**
410	8	12	6	2	**28**
411	7	11	6	2	**26**
Average*	**8**	**11**	**7**	**2**	**28**

*T13 only intended to cover Topics 1 and 2 in her course, as it is designed for struggling students. Teacher is excluded from this average.

The following table reports the correlations between Overall Assessment gains and the number of WEPs that students were exposed to. The gains scores only include students who took both the Overall pretest and posttest. The 'Number' variables reflect the total number of WEPs used per WEP type. The 'Proportion' variables reflect what percentage of the total WEPs used each WEP type made up.

Correlations

	Corrected_CK_Gain	Corrected_PK_Gain	Corrected_Flex_Gain	Corrected_Overall_Gain
Total_WEP_Number	.194**	.289**	.361**	.401**
Why_Does_It_Work_WEP_Number	.241**	.345**	.275**	.403**
Which_Is_Better_Number	-0.002	0.042	.330**	.188**
Which_Is_Correct_Number	.197**	.254**	.240**	.325**
How_Do_They_Differ_Number	.247**	.351**	.108*	.322**
Why_Does_It_Work_Proportion	0.062	0.038	-.128**	-0.022
Which_Is_Better_Proportion	-.149**	-.151**	.129**	-0.067
Which_Is_Correct_Proportion	.115*	.138**	0.047	.138**
How_Do_They_Differ_Proportion	.140**	.173**	-.146**	0.063



We examined whether estimated dosage of WEP materials (number of minutes across the school year) predicted overall posttest scores for treatment students.
Analysis: Multilevel models were run nesting students within section and within classroom and controlling for school-level demographics and pretest. Dosage positively predicted posttest scores, B = 0.02, p = .032.
Findings: More exposure to our materials predicted higher posttest scores.
Note: We tried using instrumental variable estimation models, as in our previous work, but due to the lower variability in dosage with our new implementation framework, this model did not work.

Interview Data

Prior to being introduced to the curriculum materials (pre-PD), teachers were interviewed about their beliefs. Interviewers used a structured protocol including the questions: 1) “When you discuss multiple strategies for solving a math problem, do you think that it is important to tell students that one strategy is better than another for certain problems? Why or why not?” and 2) “Do you think it’s valuable to ask your students a correct way to solve a problem and an incorrect way to solve the same problem? Why or why not?” At the end of the year, teachers participated in exit interviews using the same questions.

Transcripts of the pre-PD and exit interviews were analyzed using an open-coding process in order to identify common themes within the answers. Teacher responses to the two questions mentioned above were grouped into “Yes”, “No”, or “To some extent” categories. “Yes” responses included those where teachers expressed support for the specific type of comparison in all or most situations. An answer was coded as “No” if the respondent expressed no support for the specific type of comparison. “To some extent” answers include those where the respondent specifically said they sometimes support the strategy or where the respondent listed conditions that must be met for them to support the comparison.

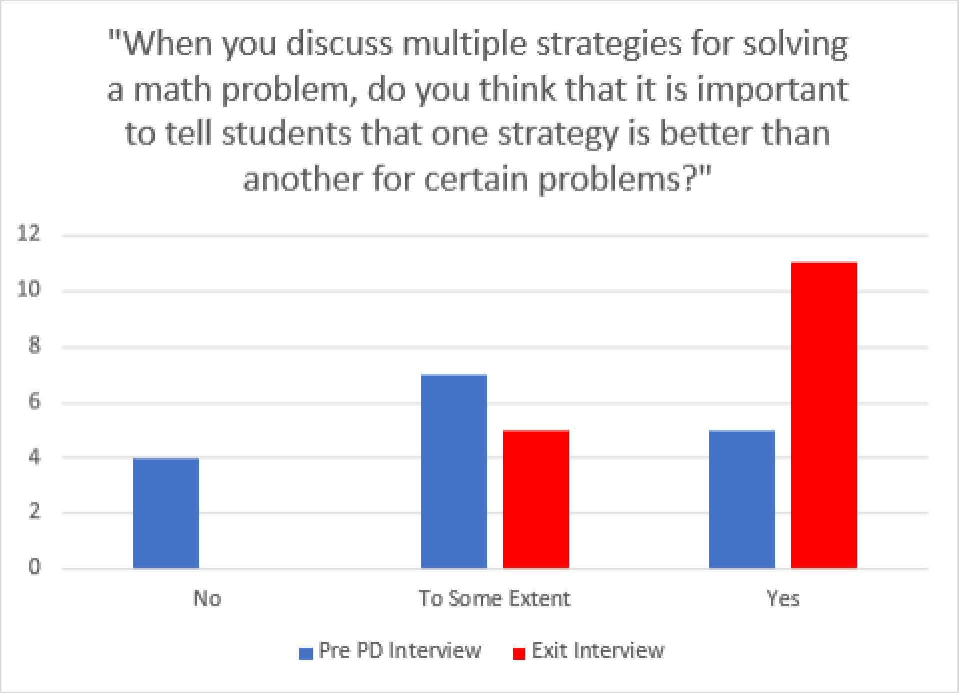
Beliefs about Comparison to More Efficient Strategies

When asked at the beginning of the project whether it is important to tell students that one strategy is better than another for solving certain problems, there was a fairly even distribution of teachers across response categories (Table 1).

Table 1: Frequency of teachers' beliefs about comparison to more efficient strategies across the year

		Exit Interview		
		No	To Some Extent	Yes
Pre-PD Interview	No	0	2	2
	To Some Extent	0	2	5
	Yes	0	1	4

Figure 1



Teachers who supported this type of comparison spoke most commonly about the importance of students learning to operate efficiently both within mathematics as well as other areas of life. Teachers who did not support language such as "better" when comparing methods emphasized encouraging students to solve problems in any way they felt comfortable, even if the chosen method is less efficient than another. Those who said it is important to some extent to discuss when a method is better than another saw value in at least pointing out efficiency to students but ultimately allowing students to solve problems in ways with which they felt comfortable. There was variability in how teachers thought about pushing their students to recognize the efficiency of certain strategies, even though evidence suggests it is a practice that benefits students' procedural knowledge and flexibility.

Many teachers' beliefs regarding the importance of efficiency-focused comparisons changed after a year of using the supplemental curriculum that emphasized this type of comparison. Overall, teachers found more value after using the curriculum throughout the school year (Table 1). All teachers who did not support this type of comparison at the pre-PD interview supported it more after a year of using the curriculum: 2 expressed conditional support and 2 expressed full support at the exit interview. Five of the 7 teachers who showed conditional support at the pre-PD interview moved to fully supporting the strategy at the exit interview. Conversely, one teacher who showed full support at the pre-PD interview expressed conditional support at the exit interview. In general, exit interviews indicated that although some teachers remained concerned with student comfort in problem-solving, they found more value in telling students that one method is better than another in certain cases and this belief changed after using the curriculum. A quote from one of the exit interviews exemplifies a common shift in belief that was observed:

> So, at the very beginning I felt that, no. I think that if kids find one method to be more comfortable than another, then they should run with that...But, there's of course, always room for the suggestion of, hey, let's look at maybe a more efficient way of doing this. And that, I think, it's a valuable lesson to learn. It's not always safe to take the scenic route. Although it's always nicer to look at, sometimes you've got to get your destination. So, providing them a highway or a quicker way of solving things is always beneficial.

The teacher that moved from fully supporting this comparison type to only conditionally supporting it stated that during the discussions that occurred in his class, students often brought up interesting ideas that they may not have if he attempted to convince them that one method was better than another. Overall, these results are promising that teachers' beliefs about using comparison to more efficient strategies can be changed if teachers have a curriculum supporting them to do so.

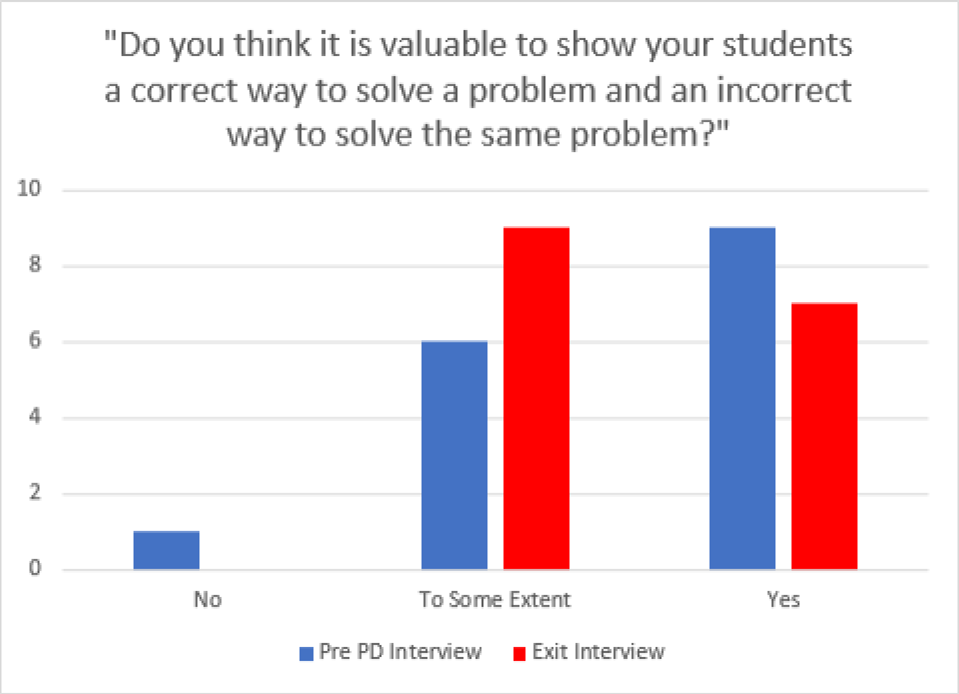
Beliefs about Comparison to Incorrect Strategies

When asked at the beginning of the project whether it is valuable to compare a correct way to solve a problem with an incorrect way, teachers generally saw value in this kind of comparison (Table 2).

Table 2: Frequency of teachers' beliefs about comparison to incorrect strategies across the year

		Exit Interview		
		No	**To Some Extent**	**Yes**
Pre-PD Interview	**No**	0	0	1
	To Some Extent	0	4	2
	Yes	0	5	4

Figure 2



Most of the teachers who found value in this type of comparison said that by exposing students to common mistakes, they are more likely to prevent students from making those mistakes in their own work. The respondent who indicated that it is not valuable expressed concern that if students are exposed to an incorrect method, they may make that mistake in their own work. Overall, the teachers who said these comparisons are valuable to some extent saw value in error analysis problems but acknowledged that caution must be used in the timing, framing, and presentation of the problem. They mentioned that it is difficult to effectively use this type of strategy, and if done poorly, it can lead to students “latching onto” the wrong method.

Table 2 also illustrates how teachers’ beliefs changed after a year of using the supplemental curriculum. The teacher who saw no value in this comparison type at the pre-PD interview expressed full support in the exit interview. At the pre-PD interview, this teacher stated that students often discover common mistakes during their own attempts at solving problems and would not benefit by these mistakes being shown to them by teachers. However, after using the supplemental curriculum, the teacher said this comparison type allowed students to understand their common mistakes on a deeper level and often led to productive mathematical class discussions.

This shift in beliefs was echoed by 2 teachers who moved from ‘To some extent’ at the pre-PD interview to ‘Yes’ at the exit interview. One of these teachers who answered ‘To some extent’ during the pre-PD interviews said that the value was conditional upon skill level of the class. She stated that these comparisons can be helpful for higher performing students but may be too difficult for lower level classes to benefit. During the exit interview, this teacher said these comparisons were very helpful in helping her students distinguish between shortcuts and errors and did not mention skill level.

There were 5 teachers who answered ‘Yes’ in the pre-PD interview but ‘To some extent’ in the exit interview. These teachers became more aware of and sensitive to some of the difficulties in utilizing this comparison type when using the curriculum. Though all of these teachers expressed some value in comparing correct and incorrect solutions, they mentioned some caveats that they felt must be met for this type of comparison to have the desired effect. Common caveats were: 1) framing the comparison so that it ends with the correct method being reinforced rather than discussing the incorrect method, and 2) only using this type of comparison to discuss common mistakes. The following quote highlights some of the perceived benefits and risks:

> I think error analysis is definitely useful...I think the timing of it is really important. If I’m just introducing material, the last thing I want to show them is how not to do it, because they’re going to look at that and they’re going to remember that and that puts almost like a negative spin on their understanding...So it’s very useful, but you have to time it well.

Teachers’ beliefs about comparison to incorrect strategies were not changed in the same way beliefs about comparison to more efficient strategies were. Over 30% of teachers became more cautious about using comparison to incorrect strategies over the year, even though research suggests that such comparison can be useful when introducing material as well (e.g., Durkin & Rittle-Johnson, 2012). This indicates that more support is needed for teachers to feel comfortable using comparison to incorrect strategies effectively at varying points during a lesson. Further exploration is needed to better understand differences in how teachers were impacted by the use of the supplemental curriculum.

Person-Presentation Study

Research Question

Does person-presentation harm generalization when used with effective learning techniques in a classroom context?

Participants

Teachers

- Five 9th grade Algebra I teachers were randomly assigned to condition using a matched randomization method
    - Person-presentation condition: 2 teachers
    - Strategy-label condition: 3 teachers

Students

- 168 students enrolled in the participating math teachers' Algebra I classes
    - Person-presentation condition: 76 students
    - Strategy-label condition: 92 students

Schools

- 2 schools in suburban Massachusetts

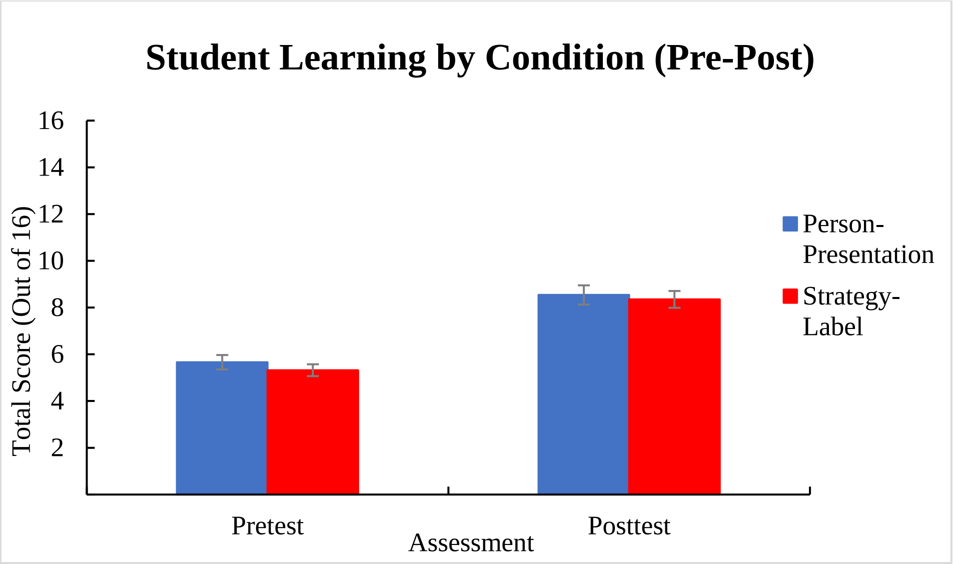
	School 1 (N = 116 students)	**School 2 (N = 52 students)**
% Free/Reduced Lunch	11	6
% White	79	89
% Asian	13	3
% Hispanic	3	4
% African American	3	1

Method

- Teachers used a supplemental curriculum with 9 worked example pairs
    - Teachers in the person-presentation condition used an average of 7.2 WEPs (range 6-8)
    - Teachers in the strategy-label condition used an average of 7.9 WEPs (range 6-9).
- After comparing and explaining each WEP, students individually rated the generalizability of each of the two strategies on a worksheet.
- Students also completed a 16-item pre/post assessment which measured their Conceptual Knowledge, Procedural Knowledge, and Procedural Flexibility.

Results

- No negative (or positive) effects of person-presentation on learning.
- No negative (or positive) effects of person-presentation on evaluations of generalizability of strategies.

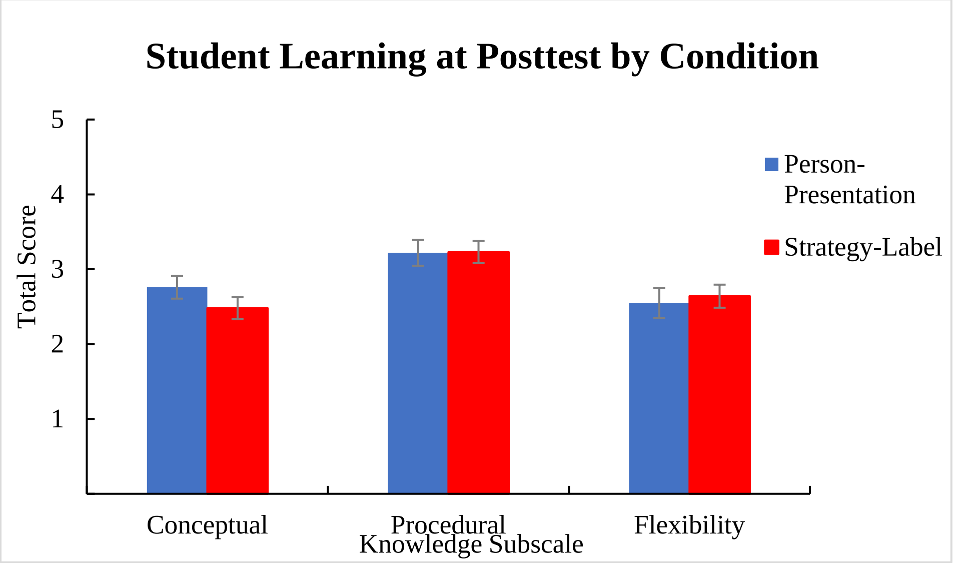


Mean Total Scores at Pretest and Posttest by Condition.

Note. Error bars represent standard error.

Analysis: An ANCOVA revealed no significant effect of condition on posttest scores controlling for pretest scores, F(1, 165) = .003, p = .96, η2p< .001.

Findings: There were no significant differences on students' posttest scores by condition.

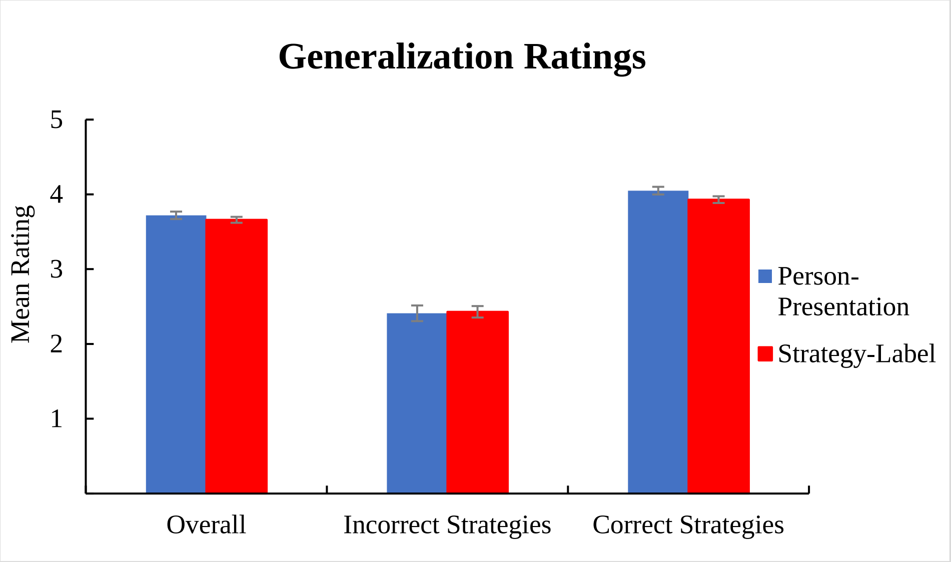


Mean Knowledge Subscale Scores at Posttest by Condition.

Note. Error bars represent standard error.

Analysis: ANCOVA models revealed no significant effect of condition on conceptual knowledge sub-scores, F(1, 165) = 1.21, p = .27, η2p= .007, procedural knowledge sub-scores, F(1, 165) = .11, p = .74, η2p= .001, or procedural flexibility sub-scores, F(1, 165) = .56, p = .46, η2p= .003, controlling for pretest total scores.

Findings: There were no significant differences in students' scores by condition.



Mean Generalization Ratings by Condition.
Note. Error bars represent standard error.
Analysis: T-tests revealed that students' ratings of the generality of the strategies in the person-presentation condition did not differ from students' ratings in the strategy-label condition, t(159) = -1.00, p = .32, Cohen's d = .40. Students rated the generality of correct strategies significantly higher than incorrect strategies, t(156) = 22.14, p < .001, Cohen's d = 3.89. However, there was no significant difference between the two conditions in generalization ratings for correct strategies, t(159) = 1.64, p = .10, Cohen's d = .61, nor incorrect strategies, t(155) = .22, p = .83, Cohen's d = .04.
Findings: There were no significant differences between the two conditions in generalization ratings.

SUMMARY OF FINDINGS

Year 3 findings are a little more promising. Unfortunately, there are demographic differences in our treatment and control groups, and teachers were not randomly assigned to condition. Most important, on Topic 1, students in treatment classrooms have higher scores at posttest, even after controlling for pretest scores. Overall, increased dosage of our materials is related to better assessment scores, and our materials do increase the use of important instructional practices, like comparison, in classrooms. Further, an exploratory study indicates that our use of characters and character names is not reducing student learning from a CEMS approach.

Publications & Presentations

Book Chapters

Rittle-Johnson, B., Star, J., Durkin, K. & Loehr, A. (in press). Compare and discuss to promote deep learning. In E. Manalo (Ed.), *Deeper Learning, Communicative Competence, and Critical Thinking: Innovative, Research-Based Strategies for Development in 21st Century Classrooms* (pp. 48-64). New York, NY. Routledge.

Rittle-Johnson, B., Star, J. R., & Durkin, K. (2017). The power of comparison in mathematics instruction: Experimental evidence from classrooms. In D. Geary, D. Berch, R. Ochsendorf, & K. Mann Koepke (Eds.), *Acquisition of complex arithmetic skills and higher-order mathematics concepts* (pp. 273-295). Cambridge, MA: Academic Press.

Journal Articles

Durkin, Rittle-Johnson, B., Star, J. R., & Loehr. A. M. (in prep). Effects of comparing and discussing multiple strategies on students' algebra learning.

Durkin, K., Star, J. R., & Rittle-Johnson, B. (2017). Using comparison of multiple strategies in the mathematics classroom: lessons learned and next steps. *ZDM*, *49*(4), 585-597.

Loehr, A., Rittle-Johnson, B., Durkin, K. & Star, J. R. (in press). Does calling it 'Morgan's way' reduce student learning? Evaluating the effect of person-presentation during comparison and discussion of worked-examples in mathematics classrooms. Applied Cognitive Psychology.

Rittle-Johnson, B., Star, J. R., & Durkin, K. (accepted pending revisions). How can cognitive science research help improve education? The case of comparing multiple strategies to improve mathematics learning and teaching. Current Directions in Psychological Science.

Conference Papers & Presentations

Durkin, K., Loehr, A. M., Rittle-Johnson, B., & Star, J. (2018, April). Effects of encouraging comparison and explanation of multiple strategies on instructional practices in algebra classrooms. Roundtable presentation at the annual meeting of the American Educational Research Association (AERA). New York City, NY.

Loehr, A. M., Rittle-Johnson, B., Star, J. R., Kang, J. M., & Durkin, K. (2017, October). Assessing conceptual understanding of algebra. Poster presented at the Cognitive Development Society (CDS), Portland, OR.

Loehr, A. M., Rittle-Johnson, B., Star, J. R., & Desharnais, C. (2018, April). Developing a more comprehensive measure of formal algebra knowledge. Poster presented at the annual meeting of the American Educational Research Association (AERA). New York City, NY.

Loehr, A. M., Durkin, K., Rittle-Johnson, B., & Star, J. R. (2019, April). Impact of comparison and explanation of multiple strategies on learning and flexibility in algebra. Paper presented at the American Educational Research Association (AERA) annual meeting, Toronto, Canada.

Loehr, A. M., Bethany Rittle-Johnson, Kelley Durkin, & Jon R. Star (2019, October). Does calling it ‘Morgan’s way’ reduce adoption and generalization of the strategy? Paper presented at the Cognitive Development Society (CDS), Louisville, KY.

Rittle-Johnson, B. Star, J., Durkin, K. & Loehr, A. (2018, May). Comparing solution strategies to promote algebra learning and flexibility. In Hsieh, F. & Kaur, B. (Eds) *Proceedings of the 8th ICMI-East Asia Regional Conference on Mathematics Education*, Volume 1. Taipei, Taiwan: National Taiwan Normal University.

Rittle-Johnson, B. (2019, October). Compare and Discuss to deepen learning. Talk presented at the National Council of Teachers of Mathematics Regional Conference, Nashville, TN.

Rittle-Johnson, B., Hickendorff, M., Star, J., Durkin, K. & Loehr, A. (2020, April). Comparing and Explaining Examples of Multiple Strategies to Promote Algebra Learning: Instructional Features that Predict Learning. Paper presented at the American Educational Research Association Annual Conference, San Francisco, CA.

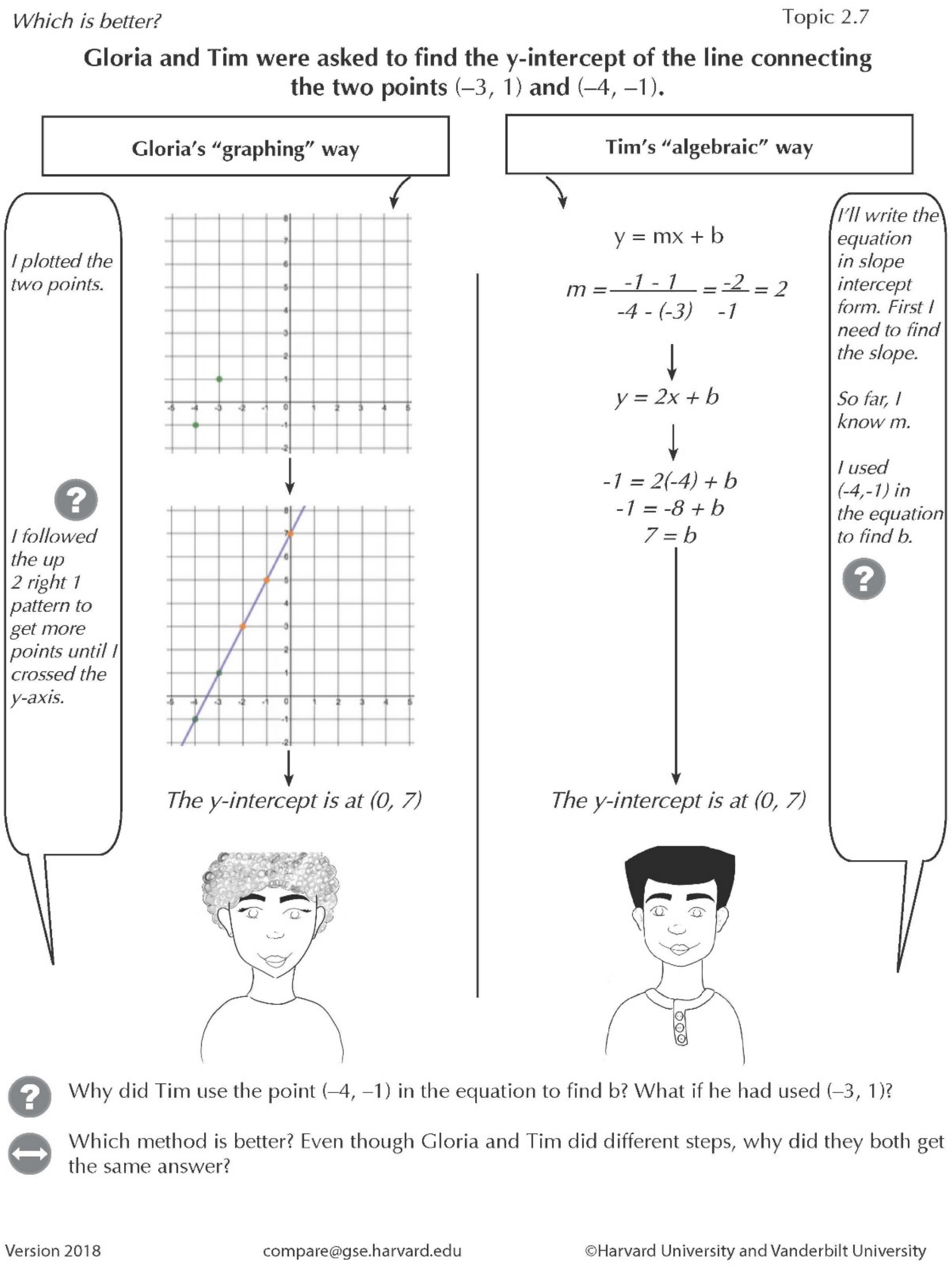
Shero, M., Durkin, K., Rittle-Johnson, B., & Star, J. R. (2020, January). Teacher beliefs surrounding comparison in algebra instruction. Poster presented at the Tennessee STEM Education Research Conference, Cookeville, TN.

Star, J. R., Rittle-Johnson, B., Durkin, K., Shero, M., & Sommer, J. (2020, June). Teaching for improved procedural flexibility in mathematics. Paper presented at the International Conference of the Learning Sciences (ICLS), Nashville, TN.

Zhang, Y., Fine, S., Loehr, A., Star, J., & Rittle-Johnson, B. (2018, May). Procedural flexibility for algebra: Assessment development. Poster presented at the 8th East Asia Regional Conference on Mathematics Education. Taipei, Taiwan.

Appendix

WEP with student worksheet



Which is better? Topic 2.7

Gloria and Tim were asked to find the y-intercept of the line connecting the two points (–3, 1) and (–4, –1).

Gloria's "graphing" way

I plotted the two points.

I followed the up 2 right 1 pattern to get more points until I crossed the y-axis.

The y-intercept is at (0, 7)

Tim's "algebraic" way

$y = mx + b$

$m = \frac{-1 - 1}{-4 - (-3)} = \frac{-2}{-1} = 2$

$y = 2x + b$

$-1 = 2(-4) + b$

$-1 = -8 + b$

$7 = b$

The y-intercept is at (0, 7)

I'll write the equation in slope intercept form. First I need to find the slope.

So far, I know m.

I used (-4,-1) in the equation to find b.

? Why did Tim use the point (–4, –1) in the equation to find b? What if he had used (–3, 1)?

⟷ Which method is better? Even though Gloria and Tim did different steps, why did they both get the same answer?

Version 2018 compare@gse.harvard.edu ©Harvard University and Vanderbilt University

Student Name: ____________________

Student ID: ____________________

Discuss Connections

If the points were changed to (3, -4) **and** (4, 2) **find the y-intercept. Did you use Gloria's "graphing" way or Tim's "algebraic" way, and which is better?**

Think, Pair. First, think about the question(s) above independently. Then, get with a partner and and discuss your answers. After talking with your partner, what is your answer?

Think	Pair

Share. After reviewing the worksheet as a class, summarize the answer(s) your class agrees on. Was this different from your original response?

Big Idea. When your teacher tells you to do so, write what you think is the big idea of this example, in your own words.