Abstract Title Page

Title: The Impact of a Comparison Curriculum in Algebra I: A Randomized Experiment

Authors and Affiliations:

Jon R. Star, Harvard University

Bethany Rittle-Johnson, Vanderbilt University

Kelley Durkin, Harvard University

Kristie Newton, Temple University

Courtney Pollack, Harvard University

Kathleen Lynch, Harvard University

Claire Gogolen, Harvard University

Abstract Body

Background / Context:

Comparison is a powerful tool that has been shown to improve learning in a variety of domains. In both laboratory studies (e.g., Namy & Gentner, 2002; Gentner, Loewenstein, & Thompson, 2003) and small-scale classroom studies (e.g., Star & Rittle-Johnson, 2009), having learners compare and contrast worked examples has been shown to reliably lead to gains in students' knowledge. Comparison is also integral to "best practices" in mathematics education. Having students share solution procedures for a particular problem and then discuss the similarities and differences in the different procedures lies at the core of reform pedagogy in many countries throughout the world (e.g., Australian Education Ministers, 2006; Brophy, 1999; Kultusministerkonferenz, 2004; National Council of Teachers of Mathematics, 2000; Singapore Ministry of Education, 2006; Treffers, 1991).

Our past empirical work has illustrated the potential benefits of comparison for students' learning of mathematics (e.g., Rittle-Johnson & Star, 2007; Rittle-Johnson, Star, & Durkin, 2012). Students who were shown two worked examples side-by-side and given the opportunity to compare and discuss similarities and differences between problems, solutions, and strategies achieved greater gains in conceptual knowledge, procedural knowledge, and flexibility, as compared to control students. The current intervention sought to build upon past research by scaling up materials to encourage comparison in the classroom throughout the academic year using a randomized control trial design.

Purpose / Objective / Research Question / Focus of Study:

Here we report the results of a year-long experiment examining the impact of researcher-designed supplemental curriculum materials that 'infused' comparison into the learning and teaching of Algebra I. Our research question is as follows: What is the effect of the supplemental comparison curriculum on Algebra I students' knowledge?

Setting:

Data were collected from teachers and students in 57 public schools across the state of Massachusetts during the 2010-2011 school year. Suburban, urban, and rural schools were represented. Teachers in the treatment condition were asked to implement the intervention materials in their classrooms at least twice a week. Teachers in the control condition followed business-as-usual practices in their classrooms.

Population / Participants / Subjects:

Seventy-seven teachers participated. Teacher age ranged from 23-66, with an average age of about 43 years. Thirty-one percent of teachers had a mathematics undergraduate degree and 81% of teachers had a graduate degree. The majority of teachers, 88%, were female. Teacher experience ranged from 1-38 years, with a mean of 10 years. All teachers taught a first-year algebra class during the 2010-2011 school year. Most of the teachers taught an 8th or 9th grade class, while a few teachers taught a 7th grade class or a class with mostly high school sophomores and above.

There were 1,661 students who participated in the study. Student age ranged from 12-19, with an average age of about 14 years. Fifty-two percent of the students were female. The majority of students were white (80%); the remaining 20% of students were approximately 4% African American, 6.5% Asian, and 6% Hispanic, with a small percentage of students classified as Native American or multi-racial. Twenty-two percent of students qualified for free or reduced lunch.

Intervention:

Teachers implemented a supplemental Algebra I curriculum designed by the research team, which included both teachers and researchers. The intervention was intended to integrate with teachers' existing algebra materials. We focused on learning in algebra classrooms because many students struggle with algebra, partially because they often memorize rules and do not learn flexible and meaningful ways to solve equations (Kieran, 1992).

The supplemental materials were a set of *worked-example pairs*, a presentation of two solved problems, placed side-by-side (see Figure 1). Approximately 150 worked example pairs were provided to choose from, spanning topics commonly found in first-year algebra courses (e.g., order of operations, equation solving, quadratics, rational expressions). Each worked example pair was classified into one of four categories with a different instructional aim (e.g., to compare two different solution methods for the same problem). Each worked example pair had a corresponding set of discussion questions, a page that displayed the worked example pair's instructional aim, and a student worksheet.

During a one-week (35 hour) professional development, treatment teachers learned about and practiced using the intervention materials. Treatment teachers were instructed to use the materials in a target class at least two times per week for the 2010-2011 school year. Teachers were not required to use all of the worked example pairs; rather, they were able to select the pairs that worked best with their course content. Class time spent on a worked example pair could vary from a small number of minutes to the majority of the class period.

Research Design:

The current study involved an experimental, randomized control trial design. Teachers were randomly assigned to condition.

Data Collection:

Data was collected in both treatment and control classrooms throughout the academic year. Student learning was assessed via two achievement measures. First, a standardized and commercial algebra readiness test (Acuity™) was given to students at the beginning and at the end of the academic year. Second, a researcher-designed assessment (the Contrasting Cases [CC] assessment, adapted from our prior small-scale studies) was administered at the beginning, middle, and end of the academic year. The researcher-designed measure included questions that tapped conceptual knowledge, procedural knowledge, and flexibility. Students were provided 40 minutes for each test administration.

Demographic information was collected for both teachers and students. Teacher background measures included age, years of teaching experience, level and type of education (i.e., undergraduate degree in mathematics, graduate degree in any field), and gender. For each student, teachers provided gender, ethnicity, prior achievement (score on the most recent state standardized test in mathematics - the 6th grade Massachusetts Comprehensive Assessment System [MCAS] score), grade level (i.e., middle school or high school), and free or reduced lunch status.

Implementation fidelity was assessed in two ways. First, treatment teachers were asked to complete an on-line implementation log after every use of the supplemental curriculum materials. Teachers were asked to provide yes/no answers to four questions relating to features of the desired implementation that were a priori deemed critical. Second, fidelity was assessed through lesson videos that were collected and submitted by teachers. Treatment teachers were asked to collect one video per month of a lesson where the supplemental materials were used, and an additional one video per month of a lesson where the supplemental materials were not used. Control teachers were asked to submit one video per month. A coding rubric was developed to assess the extent that all teachers used comparison in their classrooms. In addition, a more detailed coding rubric was also developed for use with only the treatment teachers' videos to assess whether the supplemental materials were used as intended.

Analysis / Findings / Results:

Analysis of Treatment Effects. We began by comparing the treatment and control groups on the two pretest measures and on the MCAS (see Table 1). We used a two-level hierarchical linear model, with students nested within classrooms, to estimate math achievement differences at pretest for each measure separately. There was no effect of condition on any of the three measures of prior knowledge, t's < .95, p's > .35.

To estimate the impact of treatment on student outcomes, we used a two-level hierarchical linear model, with students nested within classrooms, for our two outcomes - Acuity and our researcher-designed CC measure – in separate models. Teachers only had one class period that participated in the research. We did not include a third, school-level in the model because 75% of schools had only one teacher who participated, and 18% of schools had two teachers who participated, one in the control condition and one in the treatment condition, so only 7% of schools had more than 2 participating teachers. Demographic characteristics of the students, teachers and classrooms were included as covariates. Restricted maximum-likelihood estimation and an unstructured covariance structure were specified for estimating all models using the "proc mixed" procedure in SAS version 9.3.

Results indicated that treatment condition had a negligible impact on both outcomes. See Table 1 for means by condition and Table 2 for parameter estimates from the models. Model estimates indicated that students in the treatment condition scored about 2 percentage points higher on the CC measure and about 5 points higher on the Acuity measure, on average, than students in the control condition (see Table 2). Prior knowledge measures were strong predictors of both outcomes. A few other variables were predictive of outcomes on the CC measure, although not on the Acuity measure, including student gender, average class achievement level and the teacher having a graduate degree.

Analysis of Fidelity Effects. The lack of significant main effects may have been due to treatment teachers' failure to use our materials as frequently as we had intended. Teachers reported using our materials during an average of 19 class periods (Range 4 – 56), indicating that teachers, on average, were using our materials less than once a week.

We had requested that teachers use the materials at least twice a week. As a result, we conducted additional analyses to determine whether variation in dosage and variation in quality of implementation were linked to student outcomes. We quantified dosage as the estimated total time (in minutes) teachers spent using our materials. This was calculated by multiplying the average time spent using our materials per treatment video by the number of times teachers reported using our materials. On average, treatment teachers used our materials for an estimated total of about 269 minutes ($SD = 177$), indicating that average dosage was about 4 ½ hours over the entire school year. We quantified quality of implementation as the mean score teachers received on our treatment video coding measure of important lesson features (out of 6 possible features). On average, treatment teachers used 4.55 important features in their lessons ($SD = 0.94$).

We again used a two-level hierarchical linear model, with treatment students nested within classrooms, for our two outcomes - Acuity and our CC measure – in separate models. We included dosage and quality as Level 2 predictors in separate models. For parsimony, we included only the 3 Level 1 predictors from the original model that were significant predictors of outcomes for students in treatment classrooms: Acuity pretest score, CC pretest score, and MCAS score.

Dosage significantly predicted performance on our researcher-designed CC measure but not on the Acuity measure (see Table 3 for parameter estimates). We also examined whether dosage predicted performance on each of the subcomponents of the CC assessment. For these multilevel models, procedural, conceptual, and flexibility knowledge were used as outcome measures. Dosage did not significantly predict procedural knowledge ($\beta = 0.02$, $t = 1.44$, $p = 0.159$), but it did predict conceptual and flexibility knowledge ($\beta = 0.03$, $t = 2.82$, $p = 0.008$ and $\beta = 0.03$, $t = 3.17$, $p = 0.003$, respectively). Thus, spending additional time with our materials seemed particularly helpful for improving students' conceptual and flexibility knowledge.

Quality also marginally predicted performance on the CC measure but not on the Acuity measure (see Table 4 for parameter estimates). We also examined whether quality predicted performance on each of the subcomponents of our researcher-designed assessment. Quality did not significantly predict procedural or flexibility knowledge ($\beta = 4.88$, $t = 1.81$, $p = 0.078$ and $\beta = 3.34$, $t = 1.77$, $p = 0.085$, respectively), but it marginally predicted conceptual knowledge ($\beta = 3.51$, $t = 1.97$, $p = 0.057$). Consequently, including important instructional features may have been particularly helpful for improving students' conceptual knowledge.

Conclusions:

On the whole, results suggest that, when implemented with sufficient dosage and instructional quality, use of the supplemental curriculum that 'infused' comparison into the learning and teaching of Algebra I, improved students' learning of mathematics. Future research will determine whether encouraging sufficient dosage and instructional quality of this supplemental curriculum in classrooms might lead to better student outcomes than business-as-usual classrooms. Comparison is an important instructional tool, and supporting comparison in classrooms remains a promising method for improving student outcomes.

Appendix A. References

Australian Education Ministers. (2006). *Statements of learning for mathematics.* Carlton South Vic, Australia: Curriculum Corporations.

Brophy, J. (1999). Teaching. *Education Practices Series No. 1, International Bureau of Education.* Retrieved from http://www.ibe.unesco.org

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology, 95*(2), 393-405.

Kieran, C. (1992). The learning and teaching of school algebra. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 390-419). New York: Simon & Schuster.

Kultusministerkonferenz. (2004). *Bildungsstandards im fach mathematik für den primarbereich [educational standards in mathematics for primary schools].* Luchterhand: Munchen-Neuwied.

Namy, L. L., & Gentner, D. (2002). Making a silk purse out of two sow's ears: Young children's use of comparison in category learning. *Journal of Experimental Psychology: General, 131*(1), 5-15.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: NCTM.

Rittle-Johnson, B, & Star, J.R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology*, *99*(3), 561-574.

Rittle-Johnson, B., Star, J.R., & Durkin, K. (2012). Developing procedural flexibility: When should multiple procedures be introduced? *British Journal of Educational Psychology*, 82, 436-455.

Singapore Ministry of Education. (2006). *Secondary mathematics syllabuses.*

Treffers, A. (1991). Didactical background of a mathematics program for primary education. In L. Streefland (Ed.), *Realistic mathematics education in primary school* (pp. 21–56). Utrecht, The Netherlands: Freudenthal Institute.

Appendix B. Tables and Figures

Table 1.
Comparison of Treatment and Control on Knowledge Measures at Pretest and Posttest.

	Treatment		*Control*	
Variable	M	(SD)	M	(SD)
6th grade MCAS score	250.67	(17.03)	252.87	(15.69)
CC Pretest	41.03	(16.74)	42.76	(15.69)
Acuity Pretest	687.73	(58.50)	693.62	(55.11)
CC Posttest	66.28	(20.78)	67.09	(20.83)
Acuity Posttest	733.50	(62.88)	738.00	(69.33)

Table 2.
Parameter Estimates for Multilevel Models for Condition Effects.

	CC			Acuity		
Fixed Effects	Coefficient	SE	t	Coefficient	SE	t
Intercept	57.30	3.70	15.49***	717.26	11.91	60.22***
Student-level						
Acuity Pretest	0.09	0.01	9.26***	0.36	0.03	11.64***
CC Pretest	0.23	0.03	7.78***	0.42	0.09	4.47***
MCAS	0.28	0.04	7.60***	0.94	0.12	7.99***
Gender (Girl)	0.36	0.18	2.06*	0.48	0.71	0.68
Minority student	0.06	0.41	0.16	1.34	1.35	0.99
Free/reduced lunch	0.02	0.22	0.09	-0.54	0.71	-0.76
Class-level						
Condition	1.67	2.14	0.78	5.19	7.36	0.71
Class-achieve	0.35	0.14	2.41*	0.67	0.48	1.39
Degree in math	-0.11	2.57	-0.04	4.27	8.89	0.48
Graduate degree	6.26	2.82	2.22*	15.99	9.19	1.74~
Years teaching	0.33	0.17	1.95~	0.41	0.58	0.71
A high school	-2.18	4.24	-0.51	-16.78	14.41	-1.16
% free/reduced lunch in class	0.13	0.07	1.98~	0.16	0.22	0.72
% minority in class	0.04	0.07	0.65	0.11	0.23	0.45
Random Effects	Estimate	SE	Z-value	Estimate	SE	Z-value
Level-1 residual variance	142.46	5.67	25.14***	1359.28	55.87	24.33***
Level-2 residual variance	57.40	12.18	4.71***	597.21	131.98	4.52***

Note. Unstandardized coefficients are shown. All continuous predictor variables were grand mean centered.

~ $p < .1$, *$p < .05$, **$p < .01$, ***$p < .001$

Table 3.
Parameter Estimates for Multilevel Models Including Dosage.

	CC			Acuity		
Fixed Effects	Coefficient	SE	t	Coefficient	SE	t
Intercept	64.75	1.87	34.70***	728.87	5.19	140.25***
Student-level						
Acuity Pretest	0.09	.01	6.65***	0.41	.04	9.15***
CC Pretest	0.22	.04	5.45***	0.45	.13	3.47***
Achievement	0.24	.05	4.98***	0.89	.15	5.90***
Class-level						
Dosage	0.03	.01	2.48*	0.04	.03	1.37
Random Effects						
Level-1 residual variance	137.87	7.90	***	1373.24	80.40	***
Level-2 residual variance	114.72	31.52	***	838.80	248.76	***

Note. Unstandardized coefficients are shown. Pretest and achievement measures were grand mean centered.

**p* < .05, ***p* < .01, ****p* < .001

Table 4.
Parameter Estimates for Multilevel Models Including Quality.

	CC			Acuity		
Fixed Effects	Coefficient	SE	t	Coefficient	SE	t
Intercept	65.13	1.91	34.03***	729.45	5.19	140.45***
Student-level						
Acuity Pretest	0.09	.01	6.48***	0.41	.04	9.06***
CC Pretest	0.22	.04	5.42***	0.44	.13	3.45***
Achievement	0.24	.05	5.08***	0.90	.15	6.00***
Class-level						
Quality	3.95	1.98	1.99τ	6.76	5.33	1.27
Random Effects						
Level-1 residual variance	137.84	7.90	***	1373.03	80.37	***
Level-2 residual variance	122.39	33.35	***	848.40	250.72	***

Note. Unstandardized coefficients are shown. Pretest and achievement measures were grand mean centered.

τ $p < .06$, *$p < .05$, **$p < .01$, ***$p < .001$

Figure 1.
Worked example pair excerpt from the intervention materials.