Learning from Comparison in Algebra

Running header: LEARNING THROUGH COMPARISON

Learning from Comparison in Algebra

Jon R. Star
Courtney Pollack
Harvard University

Kelley Durkin
University of Louisville

Bethany Rittle-Johnson
Vanderbilt University

Kathleen Lynch
Harvard University

Kristie Newton
Temple University

Claire Gogolen
Harvard University

April 17, 2014

Authors' Note

Correspondence should be addressed to Jon R. Star, 442 Gutman Library, 6 Appian Way, Harvard Graduate School of Education, Cambridge, MA, 02138; 617-496-2511 (voice), 617-496-3095 (fax), jon_star@harvard.edu.

We would like to thank Alejandro Ganimian and Richard Murnane for their thoughtful comments on earlier versions of the manuscript, and John Willett for his guidance during the data analysis. Thanks as well to the many research assistants who worked on this project, including Nataliia Perova, Erin O'Connell, and Christine Size, and to the Algebra I teachers and students throughout Massachusetts who participated in the study.

This research is supported by a grant to Jon R. Star, Bethany Rittle-Johnson, and Kristie Newton from the National Science Foundation (DRL0814571); the ideas in this paper are those of the authors and do not represent official positions of the National Science Foundation.

Abstract

Mastery of algebra is an important yet difficult milestone for students, suggesting the need for more effective teaching strategies in the algebra classroom. Learning by comparing worked-out examples of algebra problems may be one such strategy. Comparison is a powerful learning tool from cognitive science that has shown promising results in prior small-scale studies in mathematics classrooms. This study reports on a yearlong randomized controlled trial testing the effect of an Algebra I supplemental comparison curriculum on students' mathematical knowledge. 141 Algebra I teachers were randomly assigned to either implement the comparison curriculum as a supplement to their regular curriculum or to be a 'business as usual' control. Use of the supplemental curriculum was much less frequent than requested for many teachers, and there was no main effect of condition on student achievement. However, greater use of the supplemental curriculum was associated with greater procedural student knowledge. These findings suggest a role for comparison in the algebra classroom but also the challenges of supporting teacher integration of new materials into the curriculum.

Keywords: Comparison; Algebra; Procedural knowledge; Supplemental curriculum; Flexibility; Conceptual knowledge

Learning from Comparison in Algebra

Mastery of algebra is an important milestone for students. Algebra serves as a gatekeeper for citizenship (Education Commission of the States, 1998) and also provides students with the ability to harness new technologies and take advantage of the job opportunities resulting from them (Moses & Cobb, 2001). Furthermore, success in algebra is necessary for access to higher mathematics and is correlated with positive life outcomes such as college graduation (Adelman, 2006; National Mathematics Advisory Panel, 2008).

Unfortunately, many students struggle with algebra. In particular, students often find the transition from arithmetic to algebra difficult (e.g., National Research Council, 2001). Algebra is the first time in mathematics where students engage in prolonged abstraction and symbolization (Kieran, 1992), for example, by frequently working with symbols that have an abstract meaning, such as variables (e.g., ‘x’). National and international assessments have drawn attention to pervasive student difficulties in algebra (e.g., Beaton et al., 1996; Blume & Heckman, 1997; Lindquist, 1989; Schmidt et al., 1999). For example, eighth-grade NAEP data show that students continue to struggle on very straightforward algebra problems: Only 59% of 8th graders were able to find an equation that is equivalent to $n + 18 = 23$, and only 31% of 8th graders were able to find an equation of a line that passes through a given point and with a negative slope (National Assessment of Educational Progress, 2011).

Improving students’ mastery of algebra requires finding effective strategies for teaching and learning algebraic topics. To this end, we report the results of a yearlong intervention based on the application of a promising approach from cognitive science to the teaching and learning of mathematics – namely, contrasting and comparing examples. In particular, we tested 1) whether the use of a supplemental comparison curriculum increased students’ knowledge in algebra, and

2) if greater use of the curriculum materials was associated with greater student knowledge in algebra. In the following section, we motivate the present study by discussing prior research on learning through comparison, both in the laboratory and the classroom.

Background on Comparison

Comparison is a powerful tool that has been shown to improve learning in a variety of domains. In both laboratory studies (e.g., Kurtz, Miao, & Genter, 2001; Namy & Gentner, 2002; Gentner, Loewenstein, & Thompson, 2003) and small-scale classroom studies (e.g., Rittle-Johnson & Star, 2007), having learners compare and contrast worked examples has been shown to reliably lead to gains in students' knowledge (see Rittle-Johnson & Star, 2011, for a review). For example, infants can learn to distinguish between cats and dogs by comparing a picture of a cat and a picture of a dog, side-by-side (Oakes & Ribar, 2005). By comparing the two pictures (or several sets of dog-cat pairs), infants may better understand important features that distinguish cats from dogs and thus learn to distinguish between the animals more easily than if the infants learned about each animal separately. Generally, comparing side-by-side examples can help individuals understand important features of a problem, which in turn may aid with novel problem solving (Gentner, Loewenstein, & Thompson, 2003) as well as future learning (Schwartz & Bransford, 1998).

In addition, comparison is 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), including the Common Core Standards (Common

Core State Standards Initiative, 2010) in mathematics in the US. At all grade levels, teachers are encouraged to create an environment where students can engage in thinking and communicating deeply about mathematics—in discussing, collaborating, justifying, conjecturing, experimenting, and responding to the ideas of their peers (National Council of Teachers of Mathematics, 2000, p. 18). The development of mathematical understanding is believed to be enhanced by classroom discussions, where students share procedures and evaluate the procedures of others (Lampert, 1990; Silver, Ghousseini, Gosen, Charalambous, & Strawhun, 2005), including informal and non-standard algorithms (Carroll, 2000; Mack, 1990). Expert teachers in the U.S. as well as teachers from high-performing countries have students compare different ways to solve the same math problem (e.g., Ball, 1993; Lampert, 1990; Richland, Zur, & Holyoak, 2007). A recent Practice Guide from the US Department of Education (Woodward et al., 2012) identified comparison as one of five recommendations for improving mathematical problem solving in the middle grades.

This practice guide recommendation is largely based on a number of small-scale experimental classroom studies documenting the benefits of comparison to students' learning of mathematics (Guo & Pang, 2011; Rittle-Johnson & Star, 2007, 2009; Star & Rittle-Johnson, 2008, 2009; Rittle-Johnson, Star, & Durkin, 2009, 2012). Each of these studies include two common features of experimental research on the benefits of comparison – the use of worked examples and prompts for explanation – both of which have been shown to improve learning (e.g., Atkinson, Derry, Renkl, & Wortham, 2000; Pashler, Bain, Bottge, Graesser, Koedinger, McDaniel, & Metcalfe, 2007). A general finding from prior small-scale studies on comparison is that students who are shown two worked examples side-by-side and given the opportunity to compare and discuss similarities and differences between problems, solutions, and strategies significantly outperform control students on a variety of outcome measures.

For example, in a one-week-long experimental study, Rittle-Johnson and Star (2007) randomly assigned 70 7th grade student pairs to learn multistep linear equation solving by either comparing two worked-out examples presented side-by-side, or by studying isomorphic worked-out examples sequentially. Using a pretest-intervention-posttest design, the authors found that students in the comparison condition demonstrated greater procedural and flexibility knowledge than students in the sequential condition. In a similar one-week-long study involving 157 5th and 6th grade students, Star and Rittle-Johnson (2009) extended the benefit of comparison to a second domain of mathematics: computational estimation. Using the same research design, the authors showed that students in the comparison condition demonstrated greater flexibility knowledge than students in the sequential condition. Comparison has also been shown to improve fourth graders' learning about the altitude of a triangle (Gao & Pang, 2011).

Thus, there is emerging evidence from laboratory studies in cognitive science and from short-duration, researcher-led classroom studies on the benefits of comparison; the above-mentioned Practice Guide noted that there is moderate evidence in support of this practice (Woodward et al., 2012). Yet research is lacking on the potential for comparison to improve long-term learning in classrooms. In a study examining feasibility of classroom implementation, Newton, Star, & Lynch (2010) demonstrated that, through comparison of worked examples, struggling 9th, 10th, and 11th grade students were able to learn and appreciate multiple strategies for solving problems during a three-week, researcher-led algebra course. However, no prior study has examined the impact of comparison on students' learning of mathematics in authentic (e.g., teacher-enacted, full-year-long) classroom environments. The present study seeks to fill this gap, by evaluating the impact of teachers' use of a supplemental Algebra I comparison curriculum. We adapted and expanded intervention materials from prior studies on comparison (e.g., Rittle-Johnson & Star, 2007) to create a full-year-long supplemental curriculum; we provided an intensive professional development to teachers on how to implement the curriculum; and we evaluated fidelity of teachers' implementation of the curriculum as well as their students' learning gains.

Supplemental Comparison Curriculum

Analyses of large-scale efforts to reform mathematics instruction suggests that when an innovation closely resembles current practices and is easy to implement, teachers are more likely to adopt the innovation (Cohen & Hill, 2001). As a result, we chose to supplement rather than replace teachers' current Algebra I curriculum with easy-to-implement materials designed to infuse comparison into teachers' regular practice. Specifically, a team of mathematics education experts, including researchers, mathematicians, and Algebra I teachers, developed the materials by going through a typical Algebra I course syllabus, identifying common student difficulties and misconceptions, and then creating materials to attempt to address them. Semi-structured interviews with a small group of teachers confirmed that comparison was indeed a reasonable adaptation of their current practice (e.g., many of the teachers introduced multiple strategies for at least some problem types, but they did not explicitly compare the strategies). These teachers piloted the supplemental comparison curriculum for a full-year prior to the present study.

At the core of the supplemental Algebra I curriculum were 141 worked example pairs (WEPs). Each WEP showed the mathematical work and dialogue of two hypothetical students, Alex and Morgan, as they attempted to solve one or more algebra problems. The curriculum contained four types of WEPs, with the types varying in what is being compared and the instructional goal of the comparison (see Figure 1). Two of the WEP types were very similar in content and format to the intervention materials used in prior work. First, *Which is better?* WEPs show the same problem solved in two different, correct methods, with the goal of understanding when and why one method is more efficient or easier than another method for a given problem (e.g., Rittle-Johnson & Star, 2007). Second, *Which is correct?* WEPs show the same problem solved with a correct and incorrect method, with the goal of understanding and avoiding common errors (e.g., Durkin & Rittle-Johnson, 2012). In addition, the curriculum also included two new types of WEPs, which first emerged during the classroom study by Newton et al. (2010) and were further developed by mathematics educators on the research team during the curriculum development process. *Why does it work?* WEPs showed the same problem solved with two different correct methods, but with the goal of illuminating the conceptual rationale in one method that is less apparent in the other method. Finally, *How do they differ?* WEPs showed two different problems solved in two different ways, but with an interest in illustrating what the relationship between problems and answers of the two problems revealed about an underlying mathematical concept. (See Rittle-Johnson & Star, 2011, for a more in-depth discussion of past research used when developing the curriculum materials.) The emergence of two new types of comparison that had not previously been identified in past research illustrates the potential impact of classroom-motivated research on more basic research.

Prompts for explanation accompany each WEP. Each WEP includes three types of prompts, which were designed to scaffold appropriate discussions and also to build consistency across examples. First, *Understand* prompts, such as, “How did Morgan solve the equation?” were intended to provide students the opportunity to understand each worked example individually, prior to comparing them. Second, *Compare* prompts, such as, “What are some similarities and differences between Alex’s and Morgan’s ways?” were meant to encourage comparison of the two worked examples. *Understand* and *Compare* prompts were very similar across WEP types and were intended to prepare students to engage in productive reflection on the final, *Make Connections* prompts, such as “On a timed test, would you rather use Alex’s way or Morgan’s way? Why?” and “In thinking about the similarities and differences between Alex’s and Morgan’s ways, what conclusions can you draw about how to solve this type of problem?”

Our pilot work revealed that sometimes teachers skipped or inadequately addressed the *Make Connection* prompts (often due to time constraints) (Newton & Star, 2013), so we supplemented each WEP with an additional, “take-away” page. On the take-away page, the fictitious students Alex and Morgan identify the learning goal for that WEP (see Figure 1). Our intent was that the teacher would use the take-away page to provide an explicit summary statement of the instructional goal of the WEP. Prior research suggests that direct instruction is needed to supplement student-generated comparisons (Schwartz & Bransford, 1998), and a feature of high-quality instruction is that teachers summarize the instructional goals of a lesson (Brophy & Good, 1986).

In order to evaluate the impact of the supplemental Algebra I comparison curriculum on students’ learning of mathematics, and consistent with prior work on comparison, we assessed students’ procedural knowledge, conceptual knowledge, and flexibility (e.g., Rittle-Johnson & Star, 2007). Procedural knowledge refers to having knowledge of action sequences for solving a problem (e.g., an algorithm for solving linear equations), while conceptual knowledge refers to having knowledge of concepts in a domain (e.g., understanding what the equal sign means) and the interconnection of the concepts or ideas in that domain (Rittle-Johnson & Alibali, 1999). Flexibility refers to the ability to solve mathematics problems in multiple ways and to know when it is most appropriate to apply particular solution methods for a given problem (e.g., Star, 2005, 2007; Star & Seifert, 2006; Krutetskii, 1976; Verschaffel, Luwel, Torbeyns, & Van Dooren, 2007). Schneider, Rittle-Johnson, and Star (2011) provided evidence for the differentiation of these three types of mathematical knowledge in algebra.

Because all four WEP types showed worked examples of algebra problems, it seems plausible that growth in procedural knowledge was potentially supported by all WEPs. However, each WEP was designed to foreground the development of certain types of knowledge. *Which is better?* WEPs aimed to foster flexibility, given that these materials illustrated multiple methods for solving problems and asked students to consider which ways were better. *Which is correct?* WEPs targeted procedural knowledge (and, to a lesser extent, conceptual knowledge), in that these examples focused on eliminating common student errors. Both *Why does it work?* and *How do they differ?* WEPs explicitly targeted conceptual knowledge.

The Current Study

Understanding the impact of innovative strategies in the classroom is a challenging and complex endeavor (Guskey, 2000). Teachers must first be adequately trained to use the new strategies and then they must actually use them in the classroom. This *use* of new strategies must be understood from two perspectives: quality and degree of implementation. In the current study, we considered both perspectives prior to our analyses, which focused on the impact of comparison on students' knowledge of algebra.

Specifically, we conducted a yearlong study in which teachers were randomly assigned to implement our Algebra I comparison intervention. First, we asked: Does the offer of a supplemental comparison curriculum increase Algebra I students' knowledge of algebra? Second, and recognizing that variation in implementation is to be expected among participating teachers, we also asked: Does increased use of the supplemental comparison curriculum lead to greater student knowledge in algebra? Based on prior work, we hypothesize that the teachers' use of a supplemental comparison curriculum would lead to improvements in their students' knowledge. Furthermore, we expected to find a dosage effect, such that students whose teachers implemented the supplemental curriculum the most would have the largest gains.

To explore these two questions, we first estimated the effects of the offer of the comparison curriculum on students' algebra knowledge; as we discuss below, such an estimate was problematic due to attrition from our sample. We then used instrumental variables estimation (IVE) to estimate the dose-response relationship between teachers' use of the intervention materials and students' knowledge for those teachers who actively participated in the study, to determine if increased student exposure to the comparison approach led to greater knowledge.

Method

Participants

Data come from 8th and 9th grade Algebra I teachers and a target class of their students from across Massachusetts. We initially selected 141 volunteer public school teachers to participate in the study during the 2010-2011 school year. To qualify for the study, teachers had to be teaching middle- or high-school Algebra I. Additionally, teachers agreed to attend a one-week professional development during the summer to learn how to implement the intervention materials. Of the 141 volunteer teachers, 64 of them had a colleague in their school who also volunteered to participate, while 77 were the only teachers in their school to volunteer. These 141 teachers came from 85 schools and 66 school districts.

There was considerable attrition between random assignment and the beginning of the school year. Attrition occurred due to professional reasons (e.g., a teacher was no longer teaching Algebra I), personal reasons (e.g., extenuating family or life circumstances), or reasons that were not shared with the research team (e.g., teachers discontinued contact). Teachers were considered attriters if they did not return pretest scores for their target class. Based on initial teacher interview data, teachers included in the analysis and teachers who attrited did not statistically significantly differ in age, years of teaching experience, the proportion with an undergraduate degree in math, or the proportion with a graduate degree. However, teachers who attrited were slightly more likely to be in the control condition than the treatment condition, $\chi^2(1) = 3.75$, $p = .053$. We believe this occurred because we were in more frequent contact with the treatment teachers from the moment of random assignment, particularly in the logistical planning for the summer professional development institute. In contrast, we had very limited contact with the control teachers from random assignment until they were asked to administer the pretest several months later, and thus their motivation to follow through on their initial agreement to participate in the study may have been reduced.

After attrition, 76 teachers across 56 schools remained in the study: 44 treatment teachers and their students (n = 945) and 32 control teachers and their students (n = 698). Finally, 8 teachers' classes were not included because the teacher did not administer the researcher-designed posttest. In the final sample, there were 68 teachers across 51 schools: 39 treatment teachers and their students (n = 781) and 29 controls and their students (n = 586). See Table 1 for teacher and student demographics. Forty-one schools had one participating teacher and 10 schools had more than one participating teacher.

Design

As an incentive to participate, we used a waitlist control design, such that teachers were randomly assigned to use the supplemental intervention curriculum in Year 1 or Year 2, with teachers assigned to Year 2 serving as the control, business-as-usual condition in Year 1. Here we report on Year 1. Random assignment occurred within schools such that if there were two participating teachers in a school, one was randomly assigned to treatment and the other to control. (We used a similar procedure when there were more than two participating teachers in the same school.) We asked all teachers to identify one target Algebra I class to participate in the study. Within their target class, treatment teachers incorporated the supplemental comparison materials into their existing curriculum.

Measures

Prior to random assignment, a member of the research team interviewed all 141 teachers to collect information on teacher demographics, including age, years of teaching experience, and education (e.g., undergraduate major in mathematics; possession of a graduate degree; level and type of teaching certification credentials).

There were two student assessments. First, teachers administered a standardized algebra readiness test, the Acuity™ Algebra Diagnostic Readiness Exam (CTB/McGraw Hill, 2007), at the beginning and end of the academic year. The Acuity test is intended to test students' readiness for the Algebra I course; it was used primarily as a measure of students' prior algebraic knowledge. The Acuity test consists of 40 multiple-choice questions that survey a range of mathematics content from fraction and integer arithmetic to equation solving, word problems, and basic probability. Each question is worth one point; raw scores are converted to a scaled score (510 – 930). In our sample, internal consistency on the exam was high (a = .90).

Second, teachers administered a researcher-designed algebra assessment at the beginning and end of the academic year. This assessment consisted of 36 multiple-choice questions testing students' knowledge of Algebra I. Portions of the assessment had been used in several prior studies of comparison learning in algebra (Rittle-Johnson & Star, 2007, 2009; Rittle-Johnson, Star, & Durkin, 2009, 2012; Star & Rittle-Johnson, 2009), while other assessment questions were taken from state and national standardized assessments. Twelve questions tested procedural knowledge (e.g., how to solve a linear equation), 13 questions tested conceptual knowledge (e.g., finding an equivalent expression or like term), and 11 questions tested flexibility (e.g., selecting the best first step in a solution). Each question on the assessment was worth one point and contributed an equal amount to the overall score, which was expressed as percent correct. The tests were machine scored. Internal consistency for the overall assessment was high (a = .89) and was satisfactory for the three sub-tests (a = .77, a = .77, and a = .76, for the conceptual, procedural, and flexibility knowledge items, respectively).

Student demographic information was collected, including sex, ethnicity, age, and free or reduced lunch status. In addition, as a general measure of prior mathematics knowledge, students' sixth grade scores on the state standardized mathematics test, the Massachusetts Comprehensive Assessment System (MCAS), were collected (see Table 1).[1]

Professional Development

All treatment teachers attended a one-week (35 hours) summer professional development institute, designed and delivered by the research team, in order to become familiar with the supplemental curriculum materials and the desired implementation model (see Newton & Star, 2013, for additional details). During the summer institute, teachers were given the opportunity to read through the supplemental curriculum materials, view videotaped exemplars of other teachers using the supplemental curriculum, and plan and teach sample lessons using the materials to their peers. In addition, teachers were given detailed guidance on the desired implementation model for the curriculum materials. Furthermore, teachers evaluated their own and their peers' sample lessons for adherence to the desired implementation model, using the instrument designed to assess implementation fidelity.

The desired implementation model included four features that were deemed *a priori* to be particularly critical to the successful use of each WEP. First, it was expected that treatment teachers would ask questions from *all three* of the different types of reflection prompts (*Understand*, *Compare*, *Make Connections*). Second, these three types of reflection prompts would be covered in the presented order (*Understand*, then *Compare*, then *Make Connections*). Third, treatment teachers would provide opportunities for students to engage in a whole-class discussion around the *Make Connections* prompts. Finally, teachers were expected to display and read to the class the take-away page for the WEP, which describes the learning objective for that WEP.

Curriculum Implementation and Fidelity

Treatment teachers were asked to use the supplemental comparison curriculum at least once per week, adhering as closely as possible to the implementation model that was discussed at the summer professional development. However, teachers had considerable flexibility in selecting which supplemental curriculum materials to use and how to integrate the supplemental materials with their regular curriculum. Each time a teacher used the supplemental materials, they were asked to submit an online implementation log. This log gathered information about which curriculum materials were used, whether the lesson contained the four critical features of the implementation model via four yes/no questions, and an open-ended comments prompt.

Based on these logs, many treatment teachers were using the supplemental materials much less frequently than intended (range from 0 to 56 days, *M* = 19.46, *SD* = 12.68). In fact, 18% of participating treatment teachers did not report using the materials even once; 30% of them reported using it 5 times or fewer.

In addition, all participating teachers were also asked to submit one video per month of their target class. Compliance varied across teachers. Control teachers submitted an average of 5.2 videos (range from 0 to 10) and treatment teachers submitted an average of 4.7 videos using our materials (range from 0 to 11). We provided a videocamera to each teacher, and all teachers selected which lessons to videotape and submit.

Fidelity measures. Fidelity was assessed by coding the classroom videos submitted by participating teachers. For all teachers, we coded videos for seven features: four items indicated whether teachers used specific instructional practices integral to the intervention (e.g., whether teachers introduced students to multiple strategies) and three items assessed whether teachers used general instructional practices that were related to the intervention (e.g., the presence of a whole-class discussion). For treatment teachers, we coded for four addition features to capture whether teachers adhered to the desired implementation model (the four critical features described above) (see Table 2). All submitted videos were analyzed by members of the research team for fidelity. Each item was scored using a dichotomous (yes/no) rating. To assess interrater reliability at least 30% of each rater's videos were double-coded. Average percent agreement on all double-coded fidelity items was 88.6%.

Analysis of fidelity. Table 2 shows the results of our analysis of teachers' fidelity. First, treatment teachers largely adhered to the desired implementation model, with mean fidelity scores ranging from .83 to .96 on the four critical features. Second, control teachers did not frequently use specific instructional practices that were integral to the intervention. Although control teachers exposed students to multiple strategies in about one-third of videos, they presented multiple strategies side-by-side only 17% of the time on average, explicitly compared multiple strategies only 14% of the time on average, or used multiple strategies in ways that resembled the intent of the intervention materials only 19% of the time on average. Third, with respect to general instructional practices that were related to the intervention, both treatment and control teachers frequently engaged students in mathematical discussions that (at least in part) involved a discussion of one or more solution methods. However, control teachers did not regularly provide a concluding summary of the major points of the discussion (unlike treatment teachers, for whom offer a concluding summary was one of the four critical features of the desired implementation model). Overall, treatment teachers were often using the supplemental materials much less often than intended (i.e., low degree of implementation), but when they did use the materials, they used them with high fidelity (i.e., high quality of implementation).

Missing Data

Of the original 1,643 students, approximately one-third of the students (*n* = 546) across 24 teachers had missing data for at least one variable used in the analysis. To manage missing data, we employed multiple imputation by chained equations (*m* = 5) for all exogenous variables (e.g., variables that were unaffected by the intervention such as demographics and pretest scores). We chose this method over multivariate normal imputation since the chained equations method does not assume a multivariate normal distribution, and so allowed us to impute values for categorical (e.g., race) and binary (e.g., gender) variables. The imputed sample consisted of 68 teachers across 51 schools: 39 treatment teachers and their students (*n* = 781) and 29 control teachers and their students (*n* = 586). Forty-one schools had one participating teacher and 10 schools had more than one participating teacher. Note that even with imputation, 8 teachers' classes were not included because the teacher did not administer the researcher-designed posttest.

We present results using multiply imputed samples. The complete case data show the same pattern of findings. Estimates from the multiply imputed sample help address the limitations of missing data. For the dosage analysis, integrating multiple imputation and instrumental variable estimation (IVE) requires us to manually calculate parameter estimates and standard errors across each set of five imputations using Rubin's Rules (Rubin, 1987).

Results

We organize the results around our two research questions – the first relating whether the offer of the intervention impacted students' knowledge and the second relating to the dosage (whether increased use of the intervention impacted students' knowledge). We then present an additional exploratory analysis that examines the relationship between usage of each WEP type and student knowledge outcomes.

Does the Offer of the Intervention Affect Students' Algebra Knowledge?

Model. First, we examined the effect of the offer of the intervention on our researcher-designed test, as well as for the three subscales of the researcher-designed test (conceptual knowledge, procedural knowledge, and flexibility). The predictor of interest was random assignment to the intervention or control group (*TRT*). To increase the precision of the estimates, we included a matrix of teacher characteristics (***X***) that contained teacher age (*AGE*) as of summer 2010 (23-65 years), the amount of teaching experience (*YRSEXP*) as of summer 2010 (9-35 years), and two dichotomous predictors in which '1' described if the teacher had an undergraduate degree in mathematics (*UDEG*) and if the teacher had a graduate degree (*GDEG*).

Similarly, we also included a matrix of student covariates (***V***) that contained demographic and prior knowledge characteristics. For demographics, we included dichotomous predictors for race, whether the student qualified for free or reduced lunch (*FR_LUNCH*), and whether the student was female. For prior knowledge characteristics, we included pretest percent correct scores on the researcher-designed test (*CC_PRE*); pretest scaled scores on the algebra acuity test (*ACUITY_PRE*); and scaled 6th grade MCAS mathematics scores (*MCAS6*). The two groups were balanced on almost all characteristics after adjusting for multiple comparisons; there were a greater proportion of White students in the intervention group and a greater proportion of Black students in the control group. Lastly, we included in the models a matrix of school fixed effects (***I***) that represents the school each student attended.

We began by estimating the effect of the intervention offer on students' algebra knowledge. Equation (1) describes a two-level multi-level model with school fixed-effects:

$$Y_{ics} = a_0 + a_1 TRT_{cs} + a_2 \boldsymbol{X}_{cs} + a_3 \boldsymbol{V}_{ics} + a_4 \boldsymbol{I}_s + z_{ics} \qquad (1)$$

in which students are nested within classrooms, which are nested within schools. In this equation, Y_{ics} represents each of the outcomes for student *i* in classroom *c* in school *s*, TRT_{cs} denotes the offer of the intervention at the class (or teacher) level, $\boldsymbol{X}_{cs}$ and $\boldsymbol{V}_{ics}$ represent teacher and student covariates respectively, and $\boldsymbol{I}_s$ denotes the fixed effects of schools. To fit these models, we used robust standard errors adjusted to account for the clustering of students within schools. We used robust standard errors to account for potential bias in the standard errors that may be caused by heteroskedasticity and non-independence within clusters, and to prevent underestimation of standard errors. The pattern of results is similar when using conventional standard errors (with the effects being slightly stronger with conventional standard errors; Cameron & Trivedi, 2005).

The parameter of interest here is a_1, which represents the effect of the offer on students' knowledge posttest scores. All tests were conducted at the .05 level.

Findings. Means for the pre and posttest measures are shown in Table 3. Results for the above model are shown in Table 4. There were no significant relationships between the offer of the intervention and students' overall, procedural, conceptual, and flexibility knowledge. The offer of the intervention was associated with a 0.97 percentage point increase in overall knowledge scores, a 2.54 percentage point increase in procedural knowledge scores, a 0.88 percentage point increase in conceptual knowledge scores, and a 0.63 percentage point decrease in flexibility knowledge scores, on average. The large standard errors associated with these estimates (see Table 4) suggested that we may have been underpowered to detect small effects that were present.

These results indicated that the offer of the intervention did not significantly improve student outcomes. However, the estimates reported above were problematic due to teacher attrition from the study, which prohibits a true intent-to-treat estimate. In addition, treatment teachers often reported using the intervention materials much less often than intended. Thus, a dose-response analysis provides a more informative look at the value of the intervention.

Does Increased Use of the Intervention Lead to Increased Algebra Knowledge?

Model. In estimating the dose-response relationship, we introduced dosage, a potentially endogenous predictor that captured teachers' self-reported decisions about how often and for how long to implement the materials. We calculated dosage for each treatment teacher based on the usage they reported through implementation logs and videos. The online implementation logs provided information about how many days teachers used the supplemental curriculum materials. The number of days that treatment teachers used the materials ranged from 0 to 56 days (*M* =

19.46, *SD* = 12.68). Submitted videos provided information about how long (in minutes) teachers used the comparison curriculum materials in a given lesson. Recall that teachers were given wide latitude in determining how to integrate the supplemental materials into the existing curriculum, as long as the desired implementation model was followed. Using the videos submitted by each treatment teacher where our supplemental materials were used, we determined how long (in minutes, on average) a teacher spent using the comparison materials in a single lesson. By multiplying a teacher's average number of minutes per lesson (from the videos) by the total number of days the teacher used the materials (from the implementation log), we arrived at an estimate of the total dosage of the materials provided by the teacher during the Algebra I course for treatment teachers. Dosage ranged from 0 to 864 minutes (*M* = 140, *SD* = 184).

The dose-response analysis addresses whether increased use of the intervention led to increased algebra knowledge in treatment classrooms. However, the variation in dosage across treatment teachers is the result of both exogenous and endogenous factors. In part, the frequency with which teachers used the intervention materials (i.e., dosage) resulted from endogenous decisions, such as how students had previously reacted to the intervention, how the teacher felt about the intervention, diligence in using the intervention, and the time or content demands on the teacher, among other reasons. Therefore, simply including dosage as an additional predictor in a multi-level model would create bias in the estimate of the causal impact of the intervention on students' algebra knowledge (Murnane & Willett, 2011).

To account for the potential endogeneity of dosage, we used an instrumental variables estimation (IVE) strategy to examine the causal impact of students' increased exposure to the intervention materials on their algebra knowledge. Instrumental variables estimation is an analytic technique that allows us to separate out the exogenous variation in dosage and use only that variation to recover the causal impact of increased use of the intervention on students' algebra knowledge. This is done using two-stage least-squares regression (2SLS). During the first stage, dosage is regressed on one or more predictors that act as a source of exogenous variation (i.e., the "instrument(s)"). This effectively carves out the exogenous variation in dosage that is subsequently used as a predictor of students' algebra knowledge in the second stage regression. Thus, the second stage regression provides an unbiased causal dose-response relationship of the effect of receiving more exposure to the curriculum materials on students' algebra knowledge. In our analysis, this can be interpreted as a treatment-on-the-treated estimate.

We used random assignment to the offer of the intervention (*TRT*) as the principal instrument. In order for *TRT* to be a viable instrument, two assumptions must hold. First, the offer of the intervention must be related to the endogenous question predictor (Murnane & Willett, 2011). In other words, *TRT* must be related to dosage. We argue this condition is met in that *TRT* is clearly related to teachers' use of the intervention materials. Only teachers who were assigned to the treatment condition had access to and used the intervention materials. Eighty-two percent of treatment teachers who participated in the study used the materials at least once. Second, a viable instrument can only be related to the outcome through the endogenous predictor (Murnane & Willett, 2011). That is, assignment to the treatment group or the control group (*TRT*) must impact students' algebra knowledge only through dosage. We argue this assumption is reasonably met; the only way that random assignment to treatment or control would impact students' algebra knowledge is likely through the use of the intervention materials.

To meet the linearity assumption of 2SLS, we adjusted the distribution of dosage by taking its square root (*S_DOSE*). The first-stage model is given in equation (2) by:

$$S\_DOSE_{ics} = g_0 + g_1 TRT_{cs} + g_2 (TRT_{cs}\ \acute{}\ \boldsymbol{M}_s) + g_3 \boldsymbol{X}_{cs} + g_4 \boldsymbol{V}_{ics} + g_5 \boldsymbol{I}_s + n_{ics} \qquad (2)$$

in which S_DOSE_{ics} represents the square root of the estimated total number of minutes in the Algebra I course in which teachers used the intervention materials and TRT_{cs} is the primary instrument. We also included a set of additional instruments, TRT_{cs} x $\boldsymbol{M}_s$, that represent the interaction of the offer with those schools in which more than one teacher participated in the study. The purpose of including multiple instruments is to carve out more exogenous variation in dosage, resulting in stronger instrumentation (Murnane & Willett, 2011). Specifically, including this set of additional instruments allowed us to use within school variation to obtain additional exogenous variation in dosage. The remaining variables in the model are as defined for equation (1). As discussed above, the purpose of the first stage is to carve out exogenous variation in dosage in order to obtain predicted values of *S_DOSE* for each student. As part of the 2SLS procedure, these predicted values are automatically used as a predictor in the second stage model to estimate the causal effect of dosage on algebra knowledge. The second stage equation is given by equation (3):

$$Y_{ics} = b_0 + b_1 S\_\hat{D}OSE_{ics} + b_2 \boldsymbol{I}_s + b_3 \boldsymbol{X}_{cs} + b_4 \boldsymbol{V}_{ics} + e_{ics} \tag{3}$$

in which $S\_\hat{D}OSE_{ics}$ represents the predicted values from equation (2); all other predictors are as defined in equation (1). The parameter of interest is b_1, which provides the estimate of the causal effect of teachers' increased use of the intervention.

Findings. We hypothesized that increased use of the supplemental comparison curriculum would lead to increases in students' algebra knowledge. To estimate the dose-response relationship, we fit equation (2) in which we use *TRT* and its interaction with schools where two or more teachers participated in the intervention (***M***) to obtain predicted values of *S_DOSE*.

The second-stage models from equation (3) estimated the relationship between the square root of dosage and students' knowledge scores on all outcome measures, using the imputed sample (see Table 5). The square root of dosage had a marginally significant positive effect on procedural knowledge. For each one-unit increase in the square root of dosage (minutes of use), students increased their procedural knowledge scores by 0.19 percentage points (*p* = .081). Figure 2 displays the relationship between percent correct and dosage for procedural knowledge, conceptual knowledge, and flexibility. On average, the difference between using the curriculum materials for the minimum (Dosage = 0) and maximum (Dosage = 864) number of minutes is a gain of 5.6 percentage points. Dosage did not have a statistically significant effect on the two other student outcomes (see Table 5).

To further understand these findings, we examined differences in gain scores between the different quartiles of dosage students received. As a point of reference, students on average gained 24 points on the research-designed measure overall, 31 points on procedural knowledge, 18 points on conceptual knowledge, and 23 points on flexibility knowledge,. In the first quartile of dosage, students gained an average of 21 points on the research-designed measure overall, 30 points on procedural knowledge, 15 points on conceptual knowledge, and 18 points on flexibility knowledge,. In the fourth quartile of dosage, students gained an average of 27 points on the researcher-designed measure overall, 34 points on procedural knowledge, 21 points on conceptual knowledge, and 28 points on flexibility knowledge,. Overall, these results indicate that the average gains in this study were not very large, but students who received higher dosage did generally have higher gain scores.

Correlations between WEP Usage and Knowledge Outcomes

As an additional exploratory analysis, we investigated the relationship between teachers' usage of particular WEP types and student knowledge outcomes. Recall that each WEP type was designed to facilitate the development of a specific type of knowledge. All examples targeted procedural knowledge because they included worked out solutions, but *Which is better?* WEPs targeted flexibility, *Which is correct?* foregrounded procedural knowledge, while both *Why does it work?* and *How do they differ?* targeted conceptual knowledge.

Table 6 shows the mean usage of each WEP type. *Which is better?* WEPs were used most frequently by treatment teachers, with an average of 6.2 WEPs used during the full-year Algebra I course. *Which is correct?* was used in similar frequency (M = 6.0 WEPs used), followed by *Why does it work?* (M = 5.1). The *How do they differ?* WEP types was used relatively infrequently (M = 2.9), in part because this WEP type was the least common in the instructional materials.

To explore the relations to student outcomes, we calculated the average classroom gain score, subtracting the average classroom score on a measure at pretest from the average classroom score at posttest. Only the frequency of use of *Why does it work?* WEPs was significantly correlated with the class average of students' overall score on the researcher-design test ($r = .35$, $p < .05$). There was not a significant relationship between the frequency with which a teacher used each WEP type and particular knowledge types. However, limited variability in frequency of use of each type makes these findings very tentative.

Summary

Overall, the offer of the intervention did not significantly affect students' knowledge. However, the treatment materials were used infrequently by many teachers, and a dose-response analysis indicated that increased use of the intervention had a marginally significant positive effect on procedural knowledge, although not on other student knowledge outcomes. Finally, only frequency of use of the *Why does it work?* WEP type significantly correlated with classroom gains on the researcher-designed test.

Discussion

Volunteer teachers were randomly assigned to use a comparison-based supplemental algebra curriculum with their Algebra I class or to continue with business as usual. We found no effect of the offer of the intervention on students' algebra knowledge. However, the offer of the use of intervention materials is unlikely to affect student learning if the intervention is not taken-up (Guskey, 2000). Although efforts to ensure fidelity in the current study seemed effective, many treatment teachers used the intervention materials much less often than intended. This underscores the importance of the second analysis, the dose-response analysis, which estimates the treatment effect for treatment teachers based on amount of use of the materials.

In estimating the dose-response relationship, we found that students' increased exposure to the intervention materials had only a marginally significant positive effect on students' procedural knowledge scores. This finding aligns with but is considerably weaker than what has been shown in prior work, where comparison has had a consistent positive impact on procedural knowledge (Rittle-Johnson & Star, 2007; Rittle-Johnson & Star, 2009). Nevertheless, the present results suggest that comparison may be effective at improving procedural knowledge in authentic classroom settings, with teachers who use a variety of different types of district-mandated curricula. The impact on procedural knowledge may reflect the fact that all four comparison types focus on comparing solution procedures, and thus should support procedural knowledge.

The dose-response analysis does not provide evidence of an effect on students' conceptual or flexibility knowledge.

There are a number of possible explanations for why the results of this study were not as strong as have been found in prior smaller, more controlled studies. First, implementation of the intervention in the smaller studies was much more uniform, both in terms of which WEPs were used (the shorter-duration studies used a much smaller set of WEPs, and all students experienced the same set of WEPs) and also in the instruction (researchers served as instructors in the shorter studies). Second and related, it may be the case that the smaller set of WEPs used in the shorter studies were more carefully designed. Despite our best efforts, it is possible that the expansion of the curriculum from a few days to a full year, which required over a ten-fold increase in the number of WEPs, led to greater variation in, and possibly lower, quality in the materials. Finally, our inability to replicate the small-scale results may result from the expansion of instruction time without comparable expansion of testing time, which may be an unavoidable consequence of transition from the 'lab' to the classroom (Davison, Fehr, & Seipel, 2011). The posttest used to assess learning in the shorter-duration studies (which focused exclusively on procedures and concepts related to linear equation solving) was approximately the same length as the posttest for the present study, despite the fact that substantially greater material was covered in the yearlong study. Thus the instrument used to assess learning in the present study may have been less sensitive for detecting changes in student learning, as compared to the shorter studies.

Despite the challenges in reproducing lab-tested results in the classroom, these findings suggest the possibility that teachers can use carefully designed comparisons to strengthen students' procedural knowledge in algebra. This work lends preliminary experimental evidence to best practices in mathematics teaching about sharing and comparing strategies for solving problems. The important role of classroom discussions where students share and evaluate the procedures of others is well established in mathematics education (Lampert, 1990; Silver, Ghousseini, Gosen, Charalambous, & Strawhun, 2005). Having students compare different ways to solve the same problem is an oft-noted feature of expert teachers in the US as well as in other high-performing countries around the world (e.g., Ball, 1993; Lampert, 1990; Richland, Zur, & Holyoak, 2007). However, effective comparison of solutions is often not supported by average U.S. teachers (Richland, Zur, & Holyoak, 2007). The current materials are a potentially promising approach to supporting effective comparison of solution methods by typical classroom teachers.

At the same time, there are a number of possible explanations for our failure to find effects for conceptual knowledge and flexibility. Of particular interest are two related explanations pertaining to (a) teachers' choices about which of our materials to use, and (b) teachers' implementation of the supplementary comparison materials. Both of these explanations offer suggestions for improvements to the comparison curriculum as well as for future research.

First, our inability to find significant effects for conceptual knowledge and flexibility may be due to teachers' choices about which of the comparison materials to use. Recall that teachers were given freedom to choose when to use the supplemental materials and which WEPs to use and also that the WEPs were designed to foreground particular knowledge outcomes. It may be the case that teachers did not use specific WEP types with sufficient frequency to lead to the desired outcomes. Had teachers used more *Which is better?* WEPs, for example, we might have seen greater gains in flexibility, or that increased use of *Why does it work?* WEPs might have led to increased gains in conceptual knowledge.

However, although WEPs were designed to target specific knowledge outcomes, these linkages have not been empirically validated. As noted above (see Rittle-Johnson & Star, 2011, for more detail), prior research focused only on the *Which is correct?* and *Which is better?* WEPs; Rittle-Johnson and colleagues have shown how use of these two WEP types can lead to gains in conceptual, procedural, and flexibility knowledge (e.g., Durkin & Rittle-Johnson, 2012; Rittle-Johnson & Star, 2007, 2009; Rittle-Johnson, Star, & Durkin, 2009, 2012). But despite our intuitions, we do not have empirical evidence that (for example) *Which is better?* WEPs are particularly useful for improving flexibility more than any of the other WEP types. In addition, there is no empirical evidence for the impact of *Why does it work?* or *How do they differ?* comparison types on student knowledge.

A complementary explanation for the present findings may relate to teachers' implementation of the supplemental curriculum. Recall that we developed a desired implementation model that included four critical features (see Table 2) and familiarized teachers with this model during the summer professional development. Our failure to find the desired effects of the materials could be related to (a) deficiencies in how we measured teachers' implementation of four critical features, and/or (b) the identification of features that were not the most critical for student learning. An example of the former is that presence of a high quality mathematical discussion was core to our vision of how the WEPs would be implemented, which made it imperative that we could identify when such a discussion occurred in a mathematics classroom. Yet doing so proved to be challenging and may not have been adequately captured by our fidelity measure. An example of the later is that anecdotal reports from coders of treatment teachers' videos identified several instances where teachers failed to engage students in conversations about the affordances and constraints of multiple methods, which we consider important for supporting flexibility. It appears that some teachers were not comfortable identifying some methods as better than others, and our implementation model and fidelity measure did not clearly specify the need to do this. Students who struggle may be especially hesitant to move away from methods that work in all cases (Newton et al., 2010; Lynch & Star, 2014); such a tendency would be exacerbated by a teacher who encouraged students to use whatever method was most comfortable for them (Lynch & Star, in press).

Overall, the largest limitation of the current study was that teachers used our materials much less often than intended. Almost a third of the sample reported using our materials on 5 or fewer occasions across the entire school year, and teachers used our materials for an average of 19 class periods for an average of 140 minutes across the entire school year. We intentionally chose to give teachers a great deal of choice in which materials to use and when to use them, but this freedom seemed to lead a substantial number of teachers to use the materials infrequently, if at all. Interest, training and carefully designed materials were not enough for many teachers to adopt our materials with much regularity.

Reflections on Classroom Research

The current study reflects the challenges and rewards of implementing interventions with classroom teachers over a sustained period of time. One reward was that, although the present study came about due to our desire to move research from the small-scale laboratory and classroom studies to more authentic school environments, our results raise many new and interesting questions that might best be explored in more basic research. In other words, not only did the lab research inform the development of these materials, but our work in classrooms has now generated additional questions that we hope to explore in more controlled settings, such as isolating the impact of *Why does it work?* and *How do they differ?* comparisons on particular student outcomes.

Several other potentially useful features of our study design are worth highlighting for other researchers interested in transitioning to the evaluation of classroom-based interventions. First, we found that the wait-list control design used here appeared to help improve recruitment since all teachers eventually had access to the intervention. Second, an additional potential aid in recruiting was the fact that teachers had a great deal of freedom in choosing how they wanted to implement our supplemental curriculum materials. Third, our approach specified an *a priori* desired implementation model and we created materials to explicitly support it (e.g., including a take-away page to highlight the summary of the big idea of each WEP), both of which appeared to lead to high fidelity of use when treatment teachers implemented the materials. Fourth, having teachers videotape themselves generally produced useful videos for minimal cost that we could effectively code for fidelity of implementation. Finally, making implementation logs available to complete on the web helped collect dosage information with minimal cost.

Several additional but valuable suggestions have also emerged based on unexpected challenges. First, our experiences suggest the delaying of random assignment as long as possible in order to try to reduce attrition. In the current study, 46% of teachers who were randomly assigned to condition left the study prior to providing any student data. Consequently, we were unable to recover or impute these data and the teachers were excluded from the analysis. However, this suggestion (while sound) must also be balanced against the advance notice that teachers need in order to be able to attend professional development before the study begins. Second, methods for improving the *quality* of implementation (i.e., fidelity) are important but inadequate; methods are also needed to encourage a high *degree* of implementation (i.e., dosage) of new strategies or of materials. For example, creating online communities of participating teachers, providing teachers with ongoing professional development, and frequent personal or phone check-ins with teachers are all suggestions that have the potential to increase dosage (but that may bring additional challenges of their own).

Conclusion

This study provides evidence that comparison may help ameliorate some of students' difficulties in Algebra I. It suggests that teachers can use carefully designed comparisons to strengthen students' procedural knowledge in algebra. Additionally, this work lends experimental evidence to best practices in mathematics teaching about sharing and comparing strategies for solving problems. Future research should focus on how to increase adoption of comparison and how to better support students' conceptual and flexibility knowledge. Such research will further nuance our collective understanding of when comparison is a successful learning tool and what types of comparison support particular types of knowledge.

Notes

[1] The mathematics MCAS is administered each year to students in grades 3 to 8. While most students in participating teachers' classes were in the 8th or 9th grade, a small number of students were in the 7th grade or in 10th-12th grades. We chose to collect 6th grade MCAS scores for two reasons. First, although all students in grade 8 and older had completed the MCAS in 7th grade, these scores were not yet released at the beginning of the study. And second, the few 7th grade students in the study had not yet taken the 7th grade MCAS. Thus 6th grade MCAS was the most recent MCAS math score that was available for all participating students in the study. However, because the 6th grade test results may have been somewhat dated for many students, we also administered the Acuity readiness test and used it as an additional measure of prior knowledge.

References

Adelman, C. (2006). *The toolbox revisited: Paths to degree completion from high school though college*. Washington, DC: United States Department of Education.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, 70(2), 181-214.

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

Ball, D. L. (1993). With an eye on the mathematical horizon: Dilemmas of teaching elementary school mathematics. *Elementary School Journal*, 93(4), 373-397.

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzales, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics achievement in the middle years: IEA's third international mathematics and science study*. Boston: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Blume, G. W., & Heckman, D. S. (1997). What do students know about algebra and functions? In P. A. Kenney & E. A. Silver (Eds.), *Results from the sixth mathematics assessment* (pp. 225-277). Reston, VA: National Council of Teachers of Mathematics.

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

Brophy, J. E., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). New York: Macmillan.

Cameron, A.C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. Cambridge: Cambridge University Press.

Carroll, W. M. (2000). Invented computational procedures of students in a standards-based curriculum. *Journal of Mathematical Behavior,* 18(2), 111-121.

CBT/McGraw Hill (2007). Acuity^{TM} Algebra. http://www.acuityforschools.com

Cohen, D .K., & Hill, H. (2001). *Learning policy: When state education reform works.* New Haven, CT: Yale University Press.

Common Core State Standards Initiative. (2010). *Common Core State Standards for Mathematics*. Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

Davison, M., Fehr, C., and Seipel, B. (2011, September). *From the lab to the classroom: Expanding and scaling up the curriculum domain*. Paper presented at the fall conference of the Society for Research on Educational Effectiveness, Washington, DC.

Durkin, K., & Rittle-Johnson, B. (2012). The effectiveness of using incorrect examples to support learning about decimal magnitude. *Learning and Instruction,* 22, 206-214.

Education Commission of the States (1998). *Equity 2000*. Denver, CO. Retrieved from http://www.ecs.org/html/Document.asp?chouseid=1510

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

Guo, J. P., & Pang, M. F. (2011). Learning a mathematical concept from comparing examples: The importance of variation and prior knowledge. *European Journal of Psychology of Education*, 26, 495-525.

Guskey, T. R. (2000). *Evaluating professional development.* Thousand Oaks, CA: Corwin Press.

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.

Krutetskii, V. A. (1976). *The psychology of mathematical abilities in school children*. Chicago: University of Chicago Press.

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

Kurtz, K., Miao, C. H., & Gentner, D. (2001). Learning by analogical bootstrapping. *The Journal of the Learning Sciences*, 10, 417–446.

Lampert, M. (1990). When the problem is not the question and the solution is not the answer: Mathematical knowing and teaching. *American Educational Research Journal*, 27(1), 29-63.

Lindquist, M. M. (Ed.). (1989). *Results from the fourth mathematics assessment of the National Assessment of Educational Progress*. Reston, VA: National Council of Teachers of Mathematics.

Lynch, K., & Star, J.R. (2014). Views of struggling students on instruction incorporating multiple strategies in Algebra I: An exploratory study. *Journal for Research in Mathematics Education*, 45(1), 6-18.

Lynch, K., & Star, J.R. (in press). Teachers' views about multiple strategies in middle and high school mathematics: Perceived advantages, disadvantages, and reported instructional practices. *Mathematical Thinking and Learning*.

Mack, N. (1990). Learning fractions with understanding: Building on informal knowledge. *Journal for Research in Mathematics Education*, 21, 16–32. Retrieved from: http://www.jstor.org/stable/749454

Moses, R. P. & Cobb, C. E., Jr. (2001). Rad*ical equations: Civil rights from Mississippi to the Algebra Project*. Beacon Press: Boston.

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. New York: Oxford University Press.

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, 5-15.

National Association of Education Progress, Question Tool. (2011). U.S. Department of Education. Retrieved from http://nces.ed.gov/nationsreportcard/itmrlsx/search.aspx?subject=mathematics

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

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education. Retrieved February 17, 2010, from http://www.ed.gov/about/bdscomm/list/mathpanel/report/final-report.pdf

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press.

Newton, K.J., & Star, J.R. (2013). Exploring the nature and impact of model teaching with worked example pairs. *Mathematics Teacher Educator*, 2(1), 86-102.

Newton, K., Star, J.R., & Lynch, K. (2010). Exploring the development of flexibility in struggling algebra students. *Mathematical Thinking and Learning*, 12(4), 282-305.

Oakes, L. M., & Ribar, R. J. (2005). A comparison of infants' categorization in paired and successive presentation familiarization tasks. *Infancy*, 7, 85–98.

Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., and Metcalfe, J. (2007) *Organizing Instruction and Study to Improve Student Learning* (NCER 2007-2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ncer.ed.gov.

Richland, L. E., Zur, O., & Holyoak, K. J. (2007). Cognitive supports for analogies in the mathematics classroom. *Science*, 316, 1128-1129.

Rittle-Johnson, B. & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology*, 91, 1-16.

Rittle-Johnson, B., & Star, J.R. (2011). The power of comparison in learning and instruction: Learning outcomes supported by different types of comparisons. In B. Ross & J. Mestre (Eds.), *Psychology of Learning and Motivation: Cognition in Education* (Vol. 55, pp. 199-226). San Diego: Elsevier.

Rittle-Johnson, B. & Star, J.R. (2009). Compared with what? The effects of different comparisons on conceptual knowledge and procedural flexibility for equation solving. *Journal of Educational Psychology*, 101(3), 529-544.

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. (2009). The importance of prior knowledge when comparing examples: Influences on conceptual and procedural knowledge of equation solving. *Journal of Educational Psychology*, 101(4), 836-852.

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.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. New York: J. Wiley & Sons.

Schmidt, W. H., McKnight, C. C., Cogan, L. S., Jakwerth, P. M., & Houang, R. T. (1999). *Facing the consequences: Using TIMMS for a closer look at U.S. mathematics and science education*. Dordrecht: Kluwer.

Schneider, M., Rittle-Johnson, B., & Star, J. R. (2011). Relations between conceptual knowledge, procedural knowledge, and procedural flexibility in two samples differing in prior knowledge. *Developmental Psychology*, 47(6), 1525-1538.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction*, 16(4), 475-522.

Silver, E. A., Ghousseini, H., Gosen, D., Charalambous, C., & Strawhum, B. (2005). Moving from rhetoric to praxis: Issues faced by teachers in having students consider multiple solutions for problems in the mathematics classroom. *Journal of Mathematical Behavior*, 24, 287-301.

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

Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education*, 36(5), 404-411.

Star, J. R. (2007). Foregrounding procedural knowledge. *Journal for Research in Mathematics Education*, 38(2), 132-135.

Star, J. R., & Rittle-Johnson, B. (2008). Flexibility in problem solving: The case of equation solving. *Learning and Instruction*, 18, 565-579.

Star, J. R., & Rittle-Johnson, B. (2009). It pays to compare: An experimental study on computational estimation. *Journal of Experimental Child Psychology*, 102, 408-426.

Star, J. R., & Seifert, C. (2006). The development of flexibility in equation solving. *Contemporary Educational Psychology*, 31, 280-300.

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.

Verschaffel, L., Luwel, K., Torbeyns, J., & Van Dooren, W. (2007). *Developing adaptive expertise: A feasible and valuable goal for (elementary) mathematics education*? Ciencias Psicologicas, 2007(1), 27–35.

Woodward, J., Beckmann, S., Driscoll, M., Franke, M., Herzig, P., Jitendra, A., Koedinger, K. R., & Ogbuehi, P. (2012). *Improving mathematical problem solving in grades 4 through 8: A practice guide* (NCEE 2012-4055). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http:// ies.ed.gov/ncee/wwc/publications_reviews.aspx#pubsearch/.

Figure Captions

Figure 1. Sample materials from the intervention curriculum.

Figure 2. Fitted dose-response relationship showing the effects of increased use of the intervention materials on procedural, conceptual, and flexibility knowledge, holding all teacher and student covariates at their sample means. Dosage values displayed on the horizontal axis represent standard deviation increments. Dosage does not have a statistically significant effect, on average, on conceptual knowledge or flexibility, and has a marginally statistically significant effect on procedural knowledge (*a* = .05).

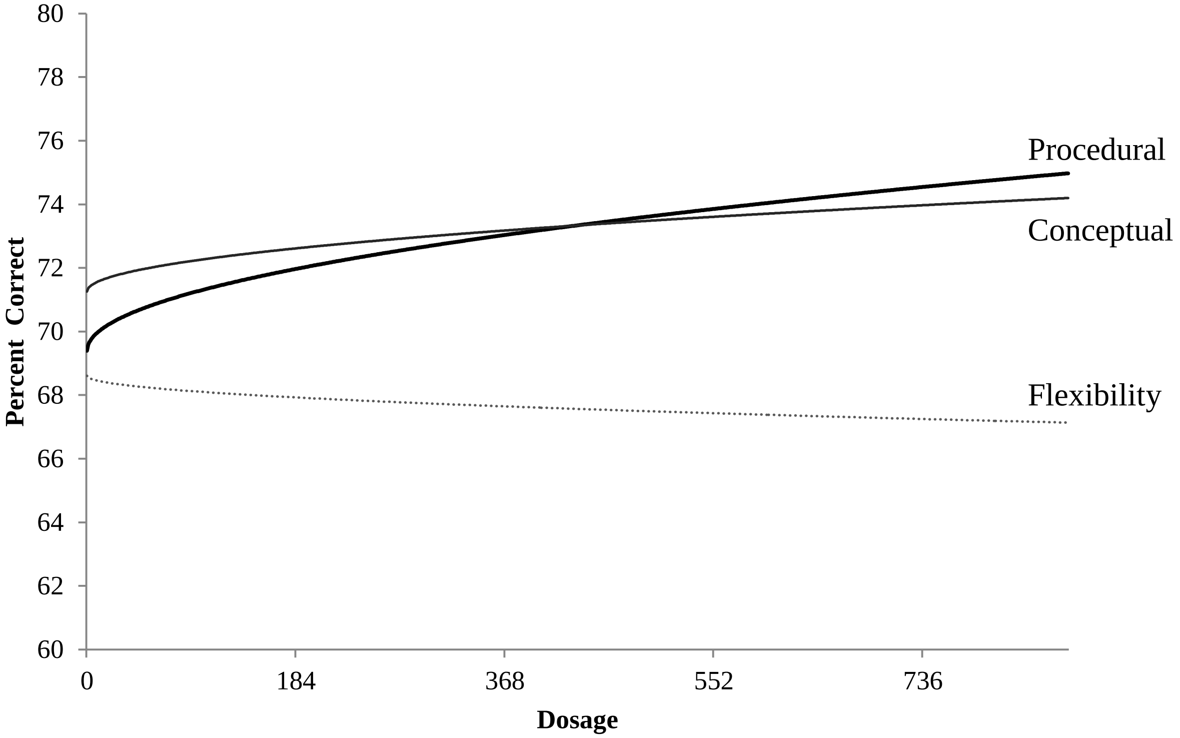

Table 1.

Teacher and Student Characteristics and Prior Knowledge by Condition

	Treatment		Control		Difference Tests	
	M	*SD*	*M*	*SD*	*t*	□^{2}
Teacher-level Demographics						
Age	43.75	9.27	41.74	11.25	0.39	
Years of teaching experience	10.33	6.93	9.26	6.52	-0.08	
Undergraduate degree in mathematics	36.49	48.17	26.96	44.41		1.10
Graduate degree	74.39	43.67	93.69	24.34		5.45*
Student-level Demographics						
White	84.59	36.13	75.22	43.21		18.15****
African-American	2.33	15.10	6.20	24.13		10.91****
Hispanic	4.40	20.53	7.06	25.63		5.09*
Asian	5.96	23.69	7.75	26.75		2.33
Multi-race	2.33	15.10	3.61	18.68		1.02
Native American	0.39	6.23	0.17	4.15		0.56
Free or reduced lunch	19.65	39.76	23.13	42.21		3.71~
Gender (Female)	50.97	50.02	53.45	49.92		0.86
Student-level Prior Knowledge Scores						
Researcher-designed measure	42.71	16.61	43.02	15.44	-0.40	
Acuity	694.86	54.13	698.20	51.94	-1.96~	
6th grade MCAS	250.86	16.96	253.12	15.54	-2.34*	

~*p* < .09, * *p* < .05, ** *p* < .01, *** *p* < .001, **** *p* < .0025 (Bonferroni adjustment for multiple comparisons)

Table 2.

Means and Standard Deviations of Teachers' Scores on the Fidelity Instrument

	Treatment teachers	Control teachers
(A) Adherence to desired implementation model		
1. Teacher asked questions from all three types of reflection prompts (*Understand*, *Compare*, *Make Connections*)	.83 (.38)	-
2. Reflection prompts were used in the desired order	.96 (.21)	-
3. Teacher engaged students in whole-class discussion around *Make Connection* prompts	.89 (.32)	-
4. Teacher displayed the learning objective for the WEP	.86 (.32)	-
(B) Specific instructional practices integral to intervention		
1. Students were exposed to multiple strategies	1.00 (.00)	.38 (.49)
2. Multiple strategies were presented side by side	1.00 (.00)	.12 (.33)
3. Teacher or students explicitly compared multiple strategies	1.00 (.00)	.09 (.29)
4. Use of multiple strategies were intended to highlight common misconceptions, demonstrate efficiency of methods, and/or illustrate why a procedure works	1.00 (.00)	.15 (.36)
(C) General instructional practices related to intervention		
1. Students participated in a mathematical discussion	.89 (.31)	.98 (.14)
2. A portion of the discussion focused on explaining one or more solution methods	.87 (.34)	.90 (.31)
3. Teacher provided concluding summary of major points of the discussion	.88 (.32)	.22 (.41)

Table 3.

Means and Standard Deviations for Each Student Outcome by Condition

	Treatment				Control			
	Pretest		Posttest		Pretest		Posttest	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Research-designed measure	42.71	16.61	66.28	20.78	43.02	15.44	67.09	20.83
Procedural	35.90	17.88	66.07	24.70	34.69	18.18	67.35	23.02
Conceptual	45.59	20.84	63.42	23.19	46.95	19.77	64.48	24.09
Flexibility	46.72	26.21	69.90	23.71	47.45	24.18	69.90	25.73
Acuity measure	694.86	54.13	734.12	62.89	698.20	51.94	737.90	69.23

Table 4.

Estimates and Standard Errors for Outcomes Based on the Offer of the Intervention

	Overall	Procedural	Conceptual	Flexibility
Intercept	-88.06 (14.71)***	-86.47 (20.88)***	-112.07 (17.06)***	-61.43 (19.09)**
Teacher-level				
TRT	0.97 (2.78)	2.54 (3.05)	0.88 (2.76)	-0.63 (3.00)
AGE	0.09 (0.35)	-0.13 (0.40)	-0.01 (0.33)	0.45 (0.33)
YRSEXP	0.39 (0.41)	0.38 (0.46)	0.38 (0.41)	0.42 (0.39)
UDEG	2.41 (2.95)	4.04 (3.45)	0.66 (2.82)	2.71 (2.92)
GDEG	4.57 (2.88)	9.66 (3.73)*	6.87 (2.52)**	-3.68 (3.50)
Student-level				
Asian	3.44 (1.62)*	2.77 (1.82)	1.36 (2.18)	6.63 (2.99)*
African-American	2.05 (1.67)	-0.72 (2.31)	1.67 (1.71)	5.52 (3.05)~
Hispanic	3.78 (2.65)	2.59 (2.24)	4.27 (3.90)	4.50 (3.40)
Multi-race	3.18 (1.55)*	2.74 (2.56)	3.17 (2.31)	3.67 (2.23)
Native American	2.98 (10.74)	-0.12 (8.89)	9.97 (9.68)	-1.89 (16.49)
FR_LUNCH	-0.03 (1.23)	0.27 (1.48)	1.78 (1.58)	-2.50 (1.83)

GENDER(Female)	2.20 (0.76)**	2.58 (0.78)**	0.18 (0.84)	4.19 (1.17)**
CC_PRE	0.20 (0.04)***	0.17 (0.05)***	0.22 (0.06)***	0.20 (0.05)***
ACUITY_PRE	0.08 (0.02)***	0.08 (0.02)***	0.08 (0.02)***	0.08 (0.02)***
MCAS6	0.30 (0.04)***	0.31 (0.06)***	0.40 (0.04)***	0.16 (0.06)*
σ_u	11.10	13.32	10.78	13.16
σ_e	12.56	15.80	15.61	19.18
ρ	0.44	0.42	0.32	0.32

Note: Unstandardized coefficients are shown with standard errors in parentheses. White serves as the reference category.

~ $p < .09$, *$p < .05$, ** $p < .01$, *** $p < .001$

Table 5.

Second Stage IVE Estimates and Standard Errors for Outcomes

	Overall	Procedural	Conceptual	Flexibility
Intercept	-86.65	-83.47	-110.18	-62.32
	(10.77)***	(12.77)***	(12.92)***	(17.06)***
Teacher-level				
S_DOSE	0.08 (0.09)	0.19 (0.11)~	0.10 (0.11)	-0.05 (0.13)
AGE	0.06 (0.11)	-0.19 (0.14)	-0.06 (0.14)	0.47 (0.17)**
YRSEXP	0.40 (0.16)*	0.39 (0.20)~	0.40 (0.20)*	0.41 (0.25)
UDEG	2.12 (1.55)	3.38 (1.94)~	0.32 (1.94)	2.89 (2.35)
GDEG	4.95 (2.31)*	10.45 (2.88)***	7.39 (2.88)*	-3.92 (3.47)
Student-level				
Asian	3.43 (1.53)*	2.78 (1.91)	1.34 (1.90)	6.63 (2.33)**
African-American	2.04 (2.00)	-0.74 (2.52)	1.65 (2.49)	5.53 (3.06)~
Hispanic	3.79 (1.95)~	2.64 (2.44)	4.25 (2.48)~	4.49 (2.98)
Multi-race	3.20 (2.32)	2.79 (2.89)	3.19 (2.88)	3.66 (3.50)
Native American	2.96 (6.47)	-0.10 (8.01)	9.88 (8.62)	-1.88 (9.74)
FR_LUNCH	-0.05 (1.23)	0.24 (1.51)	1.75 (1.56)	-2.49 (1.77)
GENDER(Female)	2.18 (0.71)**	2.54 (0.88)**	0.15 (0.88)	4.20 (1.07)***
CC_PRE	0.20 (0.03)***	0.17 (0.04)***	0.22 (0.04)***	0.20 (0.05)***
ACUITY_PRE	0.08 (0.01)***	0.08 (0.01)***	0.08 (0.01)***	0.08 (0.02)***
MCAS6	0.30 (0.04)***	0.31 (0.05)***	0.40 (0.04)***	0.16 (0.06)**

Note: Unstandardized coefficients are shown with standard errors in parentheses. White serves as the reference category.

~ *p* < .09, **p* < .05, ** *p* < .01, *** *p* < .001

Table 6.

Correlation Between Treatment Teachers' WEP Usage by Type and Treatment Class Average Posttest Gain Scores

	WEP Usage	Correlations			
WEP type	M (SD)	Overall	Procedural knowledge	Conceptual knowledge	Flexibility
Which is better?	6.2 (3.9)	0.14	0.12	0.14	0.24
Which is correct?	6.0 (4.0)	0.14	-0.16	-0.12	-0.06
Why does it work?	5.1 (4.4)	0.35*	0.10	0.18	0.22
How do they differ?	2.9 (3.2)	0.24	0.11	0.15	0.19

*p < .05