Abstract Title Page

Title: Differences in Fidelity of Implementation Measures: What Videos and Surveys Reveal About Algebra Instruction

Author(s): Kelley Durkin, Courtney Pollack, Jon R. Star, & Bethany Rittle-Johnson

Abstract Body

Background / Context:

Many students struggle in mathematics, particularly when learning algebra (Kiernan, 1992). Comparison is a potentially powerful and malleable instructional practice for supporting proficiency in algebra, and mathematics more generally**.** The promise of comparison is supported both by basic cognitive science research and by best practices in mathematics education (Ball, 1993; Gentner & Loewenstein, 2002; Hattikudur & Alibali, 2010; Lampert & Cobb, 2003; Rittle-Johnson & Star, 2007; Silver et al., 2005). When presented with contrasting solution procedures and provided support in comparing them, learners should come to notice critical features of the procedures and abstract key ideas. Our past empirical work has illustrated the potential benefits of comparison, but these interventions were only conducted over a few days' time (e.g., Rittle-Johnson & Star, 2007; Star & Rittle-Johnson, 2009). 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.

Part of this scale-up project involved implementing fidelity measures in both the treatment and control classrooms to investigate what was actually happening in classrooms both with and without our materials. Fidelity of implementation measures are important for assessing the strength of the treatment and fidelity to the program design (Correnti & Rowan, 2007; Hulleman & Cordray, 2009). One such fidelity measure involved teachers filling out surveys about their instructional practices throughout the year (see Table 1). Another fidelity measure involved coding videotapes of lessons that were collected throughout the year (see Table 2 for examples). We investigated the differences between these methods of measuring fidelity of implementation and the value of each for predicting outcomes.

Purpose / Objective / Research Question / Focus of Study:

The current paper investigated the following research questions regarding measures of fidelity:

1) Is there a significant relationship between two different measures of fidelity of implementation: a survey of instructional practices and coded videos of classroom lessons? Does the strength of this relationship differ between treatment and control teachers?
2) Do these measures of fidelity of implementation differ in their predictive value regarding student outcomes?

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 once or 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 supplementary first-year algebra 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 is 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 has a corresponding set of discussion questions, a page that displays the worked example pair's instructional aim, and a student worksheet.

During a one-week professional development, treatment teachers learned about and practiced using the intervention materials. Treatment teachers were instructed to use the materials in a target class one-to-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 and Analysis:

Data was collected in both treatment and control classrooms throughout the academic year. All teachers were asked to videotape themselves using their regular classroom materials about once a month, and treatment teachers were also asked to videotape themselves using our treatment materials once a month. Teachers generally followed these recommendations, although there was variability between teachers (number of videos per teacher ranged from 0 to 20, M = 7). These videos were later coded by trained research assistants, using a detailed coding scheme focused on instructional practices central to the intervention (see Table 2). Mean inter-rater agreement was 84% and relatively good. All teachers also completed a survey on their instructional practices at the end of each semester (see Table 1). In the fall and spring, we

assessed all students' algebra competence using a researcher-developed measure (e.g., Rittle-Johnson & Star, 2007).

From these measures, the following four variables were calculated for each teacher: 1) the mean number, across videos, of the target instructional strategies followed, 2) the mean frequency of how often teachers reported using each instructional practice on the survey, and 3) the mean score of the teacher's class on our algebra competence assessment in the fall and 4) a mean score of the class on this assessment in the spring.

To analyze the relationship between our two measures of fidelity of implementation (coded videos and the survey), we calculated the correlations between these measures for all items and between corresponding items that covered the same practice (e.g., comparing multiple strategies). We calculated these correlations for all teachers, and then separately for the treatment and the control group to see if this relationship differed between groups. We then used regression models to calculate the relative predictive value of each fidelity measure. The dependent variable was the mean score for the teacher's class on the algebra assessment in the spring, and the predictor variables were the class' mean score on the algebra assessment in the fall, the teacher's mean use of the target instructional practices (as coded from the videos), and the mean score from the instructional practices survey. We ran this model for teachers overall and then separately for the treatment group and control group.

Findings / Results:

Fidelity of implementation was high in treatment classrooms and there was some use of target instructional practices in control classrooms; however, the amount of treatment diffusion – how much control teachers were using the instructional practices supported by our supplemental materials – depended on the fidelity measure. Treatment diffusion appeared much higher on the survey than in the lesson videos. As one would expect, our two fidelity measures were significantly related to one another overall, and the correlation was moderately strong (see Table 3). When looking at particular items between measures that assessed the same instructional strategy (e.g., comparing multiple strategies), most item types were also significantly correlated between the two measures. However, when only looking at the control teachers, the two fidelity measures were no longer significantly related to one another. Essentially, control teachers said they were following instructional practices that were not observed in their coded videos. When looking at only the treatment teachers, overall the two measures were related and most item types were related as well. This indicates that what treatment teachers said they did in their classrooms more closely aligned with what was observed in their coded videos. This also suggests that the initial relationship observed between the two fidelity measures overall was being driven by the treatment teachers' results.

In addition, we examined whether our two different measures of fidelity differed in their predictive value regarding student outcomes (see Table 4). The results from our regression models indicate that neither of our fidelity measures was significantly predictive of outcomes. This was true whether we were looking at all teachers, just control teachers, or just treatment teachers. Consequently, the fidelity measures did not significantly differ in their predictive value of outcomes.

Conclusions:

Several implications for future assessment of fidelity of implementation result from this data. First, having multiple methods to measure fidelity can help give you a more complete

picture of what it is happening in control and treatment classrooms. If two measures are related to one another overall, it increases your confidence that the two measures are reliable descriptions of what is happening in classrooms. However, it is important to have a measure of fidelity that is not dependent on teachers' perceptions alone. The current results suggest that teachers in the control condition reported implementing instructional practices that were not observed in their coded videos. There are several potential reasons for this. It is possible that control teachers did not have the same understanding of terms we mentioned in the survey as we did. For example, teachers may have thought something they did was comparison that we would not define as comparison. Also, control teachers possibly reported implementing instructional practices in the classroom, which they knew we were interested in, that they used infrequently. Finally, teachers were asked to think about their instructional practices for the entire academic year when filling out the survey. On the other hand, the coded videos occurred only 7 times on average throughout the year. It is possible that the coded videos we had for some teachers were not representative of what they implemented in their classrooms throughout the year. Whatever the reason for this discrepancy, it seems important to have fidelity measures from multiple perspectives, particularly for the control teachers. In addition, while some methods for measuring fidelity, such as teacher surveys, may be easier to implement, it is important to not rely on them as the only measure of fidelity. Future research should examine the best way to reduce these possible discrepancies between different fidelity measurement methods.

In addition, the current results indicate that our two measures of fidelity did not differ in their predictive value of student outcomes. Consequently, no measure seemed better than the other for predicting outcomes, and no recommendation can be made from this data about one method being used over another. Future work will investigate the possible reasons for the lack of relationship between our fidelity measures and student outcome data, such as limited range on the measures among the treatment teachers (fidelity of implementation was consistently high).

In conclusion, it is important to use a variety of fidelity measures from multiple perspectives when assessing educational interventions. Teachers' self-reports may not match researchers' perceptions of representative lessons, particularly for control teachers.

Appendices

Appendix A. References

Ball, D. L. (1993). With an eye on the mathematical horizon: Dilemmas of teaching elementary school mathematics. *The Elementary School Journal, 93*, 373-397. doi:10.1086/461730

Correnti, R., & Rowan, B. (2007). Opening up the black box: Literacy instruction in schools participating in three comprehensive school reform programs. *American Educational Research Journal, 44*, 298-338. doi:10.3102/0002831207302501

Gentner, D., & Loewenstein, J. (2002). Relational language and relational thought. In E. Amsel & J. P. Byrnes (Eds.), *Language, literacy, and cognitive development: The development and consequences of symbolic communication* (pp. 87-120). Mahwah, NJ: Lawrence Erlbaum Associates.

Hattikudur, S., & Alibali, M. W. (2010). Learning about the equal sign: Does comparing with inequality symbols help? *Journal of Experimental Child Psychology, 107*, 15-30. doi: 10.1016/j.jecp.2010.03.004

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness, 2*, 88-110. doi:10.1080/19345740802539325

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.

Lampert, M., & Cobb, P. (2003). Communication and language. In J. Kilpatrick, W. G. Martin & D. Shifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 237-249). Reston, VA: National Council of Teachers of Mathematics.

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*, 561-574. doi:10.1037/0022-0663.99.3.561

Silver, E. A., Ghousseini, H., Gosen, D., Charalambous, C., & Strawhun, 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. doi:10.1016/j.jmathb.2005.09.009

Star, J. R., & Rittle-Johnson, B. (2009). It pays to compare: An experimental study on computational estimation. *Journal of Experimental Child Psychology, 101*, 408 - 426. doi:10.1016/j.jecp.2008.11.004

Appendix B. Tables and Figures

Table 1
Sample Items from the Instructional Practices Survey

Item Type	Sample Items
Multiple Methods	How often did students see MORE THAN ONE WAY to solve a problem in class on the SAME day?
	How often did you or the students directly compare the different methods, such as identifying similarities and differences in the methods?
Discussion	How often did students participate in a discussion?
	How often did you provide a concluding summary of major points of the discussion (e.g., at the end of the discussion, you provided the major instructional aim or main point of the discussion)?

Note: Teachers responded on a 6-point scale:

* Every day
* 3-4 times per week
* 1-2 times per week
* 1-3 times per month
* Less than once a month
* Never

Table 2
Sample Items from the Video Coding Scheme

Item Type	Sample Items	Coding Detail
Multiple Methods	Were students exposed to multiple strategies?	Check 'Yes' if: More than one way for solving a given problem was present during a single class period. This includes instances where the text presents multiple strategies and the teacher describes what is in the text.
	Did the teacher or students explicitly compare the multiple strategies?	Check 'Yes' if: Regardless of whether the strategies are presented side-by-side, the teacher or student(s) engages the class in thinking about how the two strategies are similar or different. Comparison of strategies is explicit.
Discussion	Did students participate in a discussion?	Check 'Yes' if: The discussion includes question asking and answering between the teacher and students, such that there exists opportunities for the teacher to give feedback on wrong or incomplete answers, even if the teacher doesn't take advantage of these opportunities.
	Did the teacher provide a concluding summary of major points of the discussion or lesson?	Check 'Yes' if: The teacher provided an easily recognizable encapsulation of the major instructional aim of the discussion at the end of the discussion or lesson. This may include a phrase such as, "The take-away from the past few minutes of discussion is...." Alternatively, this may be a statement summarizing the main point of the lesson as a whole, even if the discussion portion of class was not specifically focused on this point.

Table 3
Correlations between Video Coding and Survey Fidelity Measure Items

Items	All Teachers	Control Only	Treatment Only
Overall Items	.325*	-.023	.376*
Exposed to Multiple Ways	-.034	-.177	--
Compared Multiple Ways	.406**	-.214	--
Had Discussion	.320*	.066	.568**
Discussed Multiple Ways	.280*	-.027	.414*
Included Summary	.351**	.022	.445**

* $p < .05$ ** $p < .01$
Note: Some correlations could not be calculated because there was no variation in the video coding measure for those items (i.e., treatment teachers always compared and exposed students to multiple ways in coded videos).

Table 4
Regression Models for Spring Student Assessment Data Outcome

Group	**Parameter**	**Coefficient**	***SE***	***t***
All Teachers	Intercept	10.53	4.00	2.63*
	Fall Assessment Score	0.91	0.19	4.83***
	Survey Measure	-0.13	1.08	-0.12
	Video Coding Measure	-0.01	0.39	-0.01
Control Only	Intercept	3.85	6.67	0.58
	Fall Assessment Score	1.02	0.35	2.89**
	Survey Measure	1.52	1.43	1.06
	Video Coding Measure	0.91	1.64	0.55
Treatment Only	Intercept	-23.12	28.72	-0.81
	Fall Assessment Score	0.92	0.22	4.11***
	Survey Measure	-2.67	1.70	-1.57
	Video Coding Measure	8.36	6.05	1.38

* $p < .05$ ** $p < .01$ *** $p < .001$

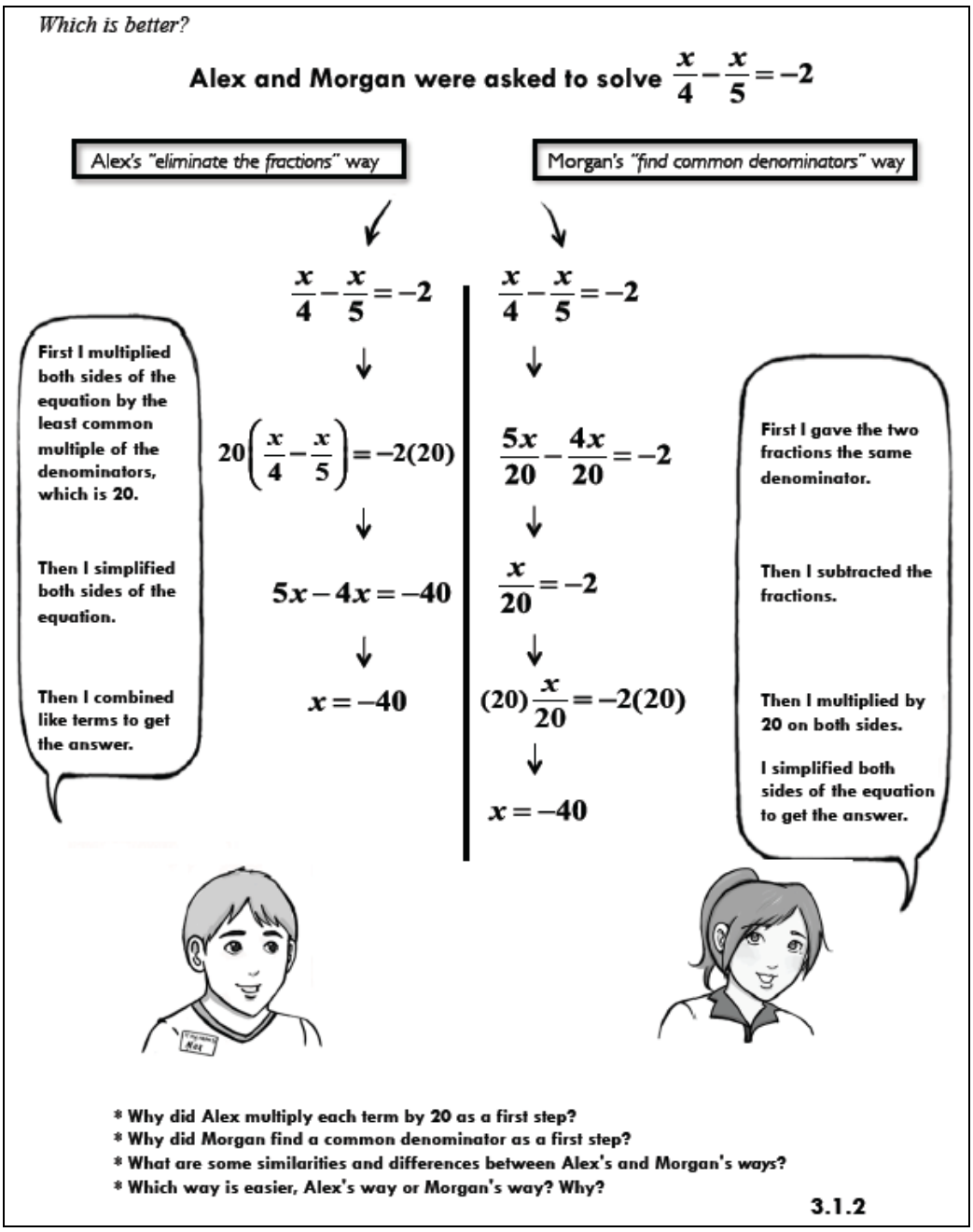

Figure 1. Worked example pair excerpt from the intervention materials.