Abstract Title Page
Not included in page count.

Title:
ASSESSING IMPLEMENTATION FIDELITY: CHALLENGES AS SEEN THROUGH THE LENS OF TWO EXPERIMENTAL STUDIES

Authors and Affiliations:

Rozy Vig, Harvard University
Megan W. Taylor, Sonoma State University
Jon R. Star, Harvard University
Theodore Chao, Harvard University

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

Educational researchers studying the implementation or effectiveness of a curriculum are increasingly asked to do so with an eye towards scientific rigor (O'Donnell, 2008). Implementation fidelity instruments are often used to this end, as a means of evaluating curricular interventions that aim to maintain high validity and reliability (Huntley, 2009; McNaught, Tarr, & Sears, 2010; Munter, 2010; O'Donnell, 2008). The concept of "implementation fidelity" is broadly used to capture the extent to which an intervention is executed as intended by the designers of the intervention (Century, Rudnick, & Freeman, 2010; Huntley, 2005, McNaught, Tarr, & Sears, 2010, Munter, 2010). Though implementation fidelity instruments are often used to assess variability in teachers' implementations of an intervention and can be related to measures of student learning, the form and goals of the work involved in developing, implementing, and evaluating fidelity instruments vary from study to study. In this research report, we discuss two research projects that shed light on two challenges associated with assessing implementation fidelity.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

The first challenge to assessing implementation fidelity is determining how to measure it. What type of instruments should be used and why? On the one hand, direct observation of instruction (live or via video) might intuitively seem to be the best way to assess fidelity, yet such observations are costly. On the other hand, teacher self-report is a cheaper way to document fidelity, yet may be unreliable, especially at large scale. To what extent are these intuitions about assessing fidelity valid? To what extent might the use of multiple measures of fidelity address concerns about the use of a single measure? Which types of measures are 'best' and why? And how do the answers to these questions differ, depending on the goals of a given study?

The second challenge to assessing implementation fidelity is what we refer to as the "fidelity variation dilemma." Should fidelity always be a measurable, valued quality of instruction (where a hoped-for goal would be high fidelity)? Or should fidelity be considered an inevitable product of the adaptation and innovation that is at the core of teaching (where low fidelity may be acceptable or even celebrated)? Should these two ways of conceiving fidelity be viewed as different constructs and therefore be measured separately? Or should adaptation (e.g., positive deviation from the intended implementation) be accounted for in assessments of fidelity? Given this dilemma, how do we then relate measures of fidelity to measures of student learning? What components of an intervention account for variability in student outcomes?

Setting:

Description of the research location.

The two projects in which these questions are explored share certain features. Both sought to evaluate the impact of a middle grades math curriculum unit and both were experimental studies that involved a large number of schools, teachers, and students.

The Contrasting Cases (CC) Project

Population / Participants / Subjects:
Description of the participants in the study: who, how many, key features, or characteristics.
Intervention / Program / Practice:
Description of the intervention, program, or practice, including details of administration and duration.
Research Design:
Description of the research design.
Data Collection and Analysis:
Description of the methods for collecting and analyzing data.
Findings / Results:
Description of the main findings with specific details.

In the CC Project, approximately 100 8th and 9th grade Algebra I teachers were randomly assigned to either a treatment or control group. All treatment teachers attended a one-week professional development institute and were asked to regularly supplement their normal algebra curriculum with worked example pairs that compare two solution methods with paired explanation prompts. Control teachers received no professional development and implemented their normal algebra curriculum without supplementation. Students in all teachers' classrooms completed two pre- and post-assessments: one was a standardized commercial algebra test and the other a researcher-designed assessment that was more closely aligned with the goal of our intervention.

We assessed implementation fidelity in three ways. First, teachers collected videos of their lessons. Treatment teachers were asked to videotape their classes once per month when they were using our materials and once per month when they were not, while control teachers were asked to tape themselves once per month. We created two coding rubrics, one for scoring lessons where our materials were used and one for when our materials were not used. The rubrics asked binary (yes/no) questions about the presence of key features of the intervention. The purpose of the first rubric was to capture variability in implementation of the intervention and relate these scores to student outcome measures. We asked questions such as, "Did the teacher follow the proper order of the discussion phases: understand, compare, and then make connections?" The purpose of the second rubric was to capture treatment diffusion (the extent to which control teachers implemented key features of the intervention) and reduce the risk of underestimating the effectiveness of our intervention. We asked questions such as, "Did the teacher or students explicitly compare multiple strategies?" Members of the research team, who participated in the development of the two coding rubrics, scored the videos. As a second measure of fidelity, teachers completed an instructional practices survey twice per year, once in December and once in May. The survey asked teachers Likert-type questions about instruction in mathematics as related to the principles underlying our intervention (e.g., "How often did students see more than one way to solve a problem in class on the same day?"). The third assessment of fidelity was a teacher log that also involved teacher self-report. Treatment teachers were asked to log each time they used our curriculum, a self-assessment of how closely they adhered to the instructional model that supported our curriculum materials (e.g., "Did you touch on the primary instructional aim of all three discussion phases?").

We report the following four results as related to implementation fidelity, along with some open questions that highlight the fidelity challenges that are the focus of this research report.

First, treatment teachers' fidelity, as determined by our coding of their videos, was very high. Do these high scores indicate that the teachers were implementing our intervention with a high degree of fidelity? Or, do these results indicate that the fidelity rubric for scoring treatment teachers' videos did not adequately and rigorously assess the appropriate level of variability in implementation? Second, treatment teachers' self-reported instructional practices, as measured by their survey and log responses, were significantly and positively correlated with scores from our coding of their videos. Does this correlation indicate that treatment teachers' self-reported practices (which were relatively easy and cheap to obtain) give a reasonable and usable measure of their implementation? Third, treatment diffusion appeared to be higher on control teachers' survey responses, as compared to their fidelity scores from the analyzed videos. Does this result indicate that control teachers' self-reported practices were not an accurate depiction of their instruction? Or, does this result indicate that the rubric used to assess treatment diffusion was too general to be useful? Fourth, the three different types of fidelity measures did not differ in their predictive value of student learning outcomes. What can we conclude about our attempts to measure fidelity of implementation, given that none of the three types of measures clearly linked to our outcomes?

The Transforming Engagement of Students in Learning Algebra (TESLA) Project

Population / Participants / Subjects:
Description of the participants in the study: who, how many, key features, or characteristics.
Intervention / Program / Practice:
Description of the intervention, program, or practice, including details of administration and duration.
Research Design:
Description of the research design.
Data Collection and Analysis:
Description of the methods for collecting and analyzing data.
Findings / Results:
Description of the main findings with specific details.

As part of a large-scale study of motivation in 5th through 8th grade mathematics, 400 teachers from a large, suburban district in the southeast United States implemented a five-day mathematics curriculum (Authors, 2012). An important component of the curriculum was a two-day lesson focused on pattern exploration, built around a task designed to require high-level cognitive demand according to the *Task Analysis Guide* (Stein, Grover, & Henningsen, 1996; Stein, Smith, Henningsen, & Silver, 2000). All students completed two, researcher-designed, pre- and post-assessments: one aligned with patterning problems related to the content goals of the two-day lesson and the other aligned with the broader motivational goals of our intervention. A small subset of teachers (N=9) were videotaped.

These lessons were scored for fidelity using an instrument designed by members of the research team. The instrument asked binary (yes/no) questions about the presence of lesson phase-specific activities, such as if the teacher asked certain questions or engaged students in certain types of discussions. The instrument was piloted and revised by the research team to improve its validity, and in the final version items were weighted so that the scores more accurately reflected key components of "ideal" implementation.

We describe two observations about the use of the fidelity instrument in this study here. First, the fidelity instrument was successful in capturing variation in teachers' implementation.

Overall, fidelity of implementation scores varied from a low of 29 to a high of 50 (out of a possible 73). The 9 teachers varied in which components of the lesson they taught with high fidelity, the order they presented the lesson activities, the materials scaffolding they provided, and the length, focus, and depth of class discussions. With such variability in implementation, what can we say about the effectiveness of the intervention as related to student outcomes? Second, the fidelity instrument failed to capture (what we feel was) the important variation in the ways teachers did or did not maintain cognitive demand. In particular, we noticed a very distinctive "dip" in the cognitive demand of the task in some teachers' lessons. The teachers appeared to implement the task as we had instructed them to do and as the instrument was designed to assess, but as observers, it was clear that something was missing. With improvements to the fidelity instrument, might this kind of variation be captured? Or might there be critical aspects of implementation that are difficult, if not impossible, to capture with a fidelity instrument?

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

Measuring fidelity of implementation is an important component of any evaluation of teacher-implemented curriculum. Two challenges associated with implementation fidelity are determining how to measure fidelity and determining the impact (if any) of teacher adaptation on the development, implementation, and evaluation of fidelity measures.

The CC project speaks to both challenges and highlights the tensions in interpreting fidelity scores when drawn from multiple measures. The notion that more is better is brought into question when the predictive value of the different measures does not differ in determining student outcome measures. However, the use of multiple measures does provide one with different perspectives on teacher adaptation, which can be of considerable value during early stages in the research cycle when the intervention is still in the process of revision.

The TESLA project speaks to the second challenge and highlights that fidelity instruments may be limited in what they can tell us about teaching and learning. For example, when two teachers have the same fidelity score, similarities in their scores might obscure important differences in their classroom practice, differences that may have significant impact on student outcome measures? If the goal of a study is to determine the effectiveness of the intervention, as measured by student outcomes, then teacher adaptation muddies the water. If, however, the goal of a study is to speak to the efficacy of the intervention, than understanding the nuances in teacher adaption become centrally important.

Appendices

Not included in page count.

Appendix A. References

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation, 31*(2), 199-218.

Huntley, M. A. (2005). *Operationalizing the concept of "fidelity of implementation" for NSF-funded mathematics curricula*. In Proceedings of the 2005 National Science Foundation K-12 Mathematics, Science, and Technology Curriculum Developers Conference, (Vol. pp. 38–43). USA.

Huntley, M. A. (2009). Brief report: Measuring curriculum implementation. *Journal for Research in Mathematics Education,* 40, 355–362.

McNaught, M. D., Tarr, J. E., & Sears, R. (2010). *Conceptualizing and measuring "fidelity of implementation of secondary mathematics textbooks: Results of a three-year study.* Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO, May 2010.

Munter, C. (2010). Evaluating Math Recovery: The Impact of Implementation Fidelity on Student Outcomes. (Doctoral dissertation). Retrieved July 15, 2012.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research*, 78(1), 33-84.

Stein, M. K., Grover, B., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal, 33*, 455-488.

Stein, M. K., Smith, M. S., Henningsen, M., & Silver, E. A. (2000). *Implementing standards-based mathematics instruction: A casebook for professional development*. New York: Teachers College Press.

Appendix B. Tables and Figures

Not included in page count.