Abstract Title Page
Not included in page count.

Title: The Use of Program Theory in Mathematics Education Evaluation Research

Authors and Affiliations: Charles Munter (University of Pittsburgh), Paul Cobb (Vanderbilt University), and Calli Shekell (University of Pittsburgh)

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

One purpose of education research is to develop and rigorously evaluate the effectiveness of programs for supporting students' learning and achievement. The Institute of Education Sciences has amplified that purpose (Shadish & Cook, 2009) and attempted to improve the methodological standards for conducting such work—primarily through the What Works Clearinghouse (WWC), which, since 2002, has supported an ongoing effort to synthesize research on the effectiveness of educational interventions, programs, and policies. According to its stringent, methodological standards, only "well-designed and well-implemented" randomized controlled trials and studies employing quasi-experimental designs with equating or matching are included in the WWC's 15 topical syntheses, one of which is mathematics.

In addition to efforts to define and increase methodological standards of evaluation research, over the years, there have also been multiple calls for increased attention to the mechanisms by which programs produce some change—the theorized processes by which inputs and outputs are linked (Bickman, 1987; Cook & Shadish, 1986; Lipsey, Crosse, Dunkle, Pollard, & Stobart, 1985; Lipsey, 1993; Rogers, 2007). Bickman (1987) referred to this as "program theory," and argued that "[t]he nature of the generalizability process requires not only that the nature of the program be explicated but also the nature of the theory underlying the program be explicated" (p. 9). Lipsey (1993) argued that the theory underpinning a program must play a role in each step of an evaluation of that program, from initial design to the interpretation of findings.

Yet, despite repeated calls for increased attention to theory in program evaluation for more than 25 years, recent reviews of the literature suggest that the field has seen little change (Confrey & Stohl, 2004; Coryn, Noakes, Westine, & Schröter, 2010; Weiss, 1997). And, the requirement that evaluators link research designs to the theories underlying the programs they evaluate is not included in the WWC's standards for rigorous evaluations. Given the considerable differences in mathematical goals and theories of learning from which mathematics programs are designed, and the high stakes for students' mathematics learning and academic futures (as well as for the fortunes of program developers), it is important to ask whether the WWC's methodological specifications are sufficient, or whether program evaluation has once again lost (or perhaps never gained) sight of the role of theory in evaluation design and implementation (Bickman, 1987), resulting in overly-constrained and uninterpretable syntheses of otherwise methodologically strong evaluation research (Schoenfeld, 2006).

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

Broadly stated, our purpose was to determine whether calls for theory-based evaluation research have had an impact on the extent to which evaluators of mathematics programs attend to program theory in their design, implementation, and reporting of studies. Specifically, we asked the following questions, some of which are adaptations of those addressed by Coryn et al. (2010), with more specific considerations drawn from Confrey & Stohl (2004). For each report that met stringent methodological standards (i.e., WWC evidence standards), we asked:

1. What type of program theory was articulated (none, sub-theoretical, or theoretical; Lipsey et al., 1985)?
2. What was the quality of evaluators' articulation of program theory (entirely implicit, drawn from a limited number of resources such as developer or publisher descriptions, or drawn from multiple resources and situated in the research literature)?
3. How was the articulated program theory used in the evaluation? Specifically:
   a) To what extent did the research questions concern relationships (e.g., mediating, moderating) between components of the program theory?
   b) To what extent did the evaluation employ a variety (e.g., topic, format, cognitive demand) of outcome measures that are valid measures of curricular intents and are aligned with systemic factors?
   c) Was the level of implementation fidelity assessed? If so, did the assessment pertain to aspects of process (e.g., quality of delivery, whether the program differed across study groups) as well as structure (e.g., adherence to expectations concerning exposure or duration of treatment, or whether particular materials were used) (Dane & Schneider, 1998; Mowbray, Holter, Teague, & Bybee, 2003; O'Donnell, 2008)?
   d) To what extent did analyses both determine whether effects were found on selected outcome variables and explain those results in terms of cause-effect associations between theoretical constructs (i.e., mechanisms)?

Also, based on the findings of previous reviewers regarding relationships between use of program theory and background or program characteristics, we asked a series of secondary questions in order to characterize any trends we identified when addressing the above questions:

4. To what extent does attention to program theory vary by characteristics of the program and evaluator(s), such as the nature of the program's mathematics learning goals and instruction (i.e., back-to-basics, typical, inquiry-based, or blended); type of publication (e.g., peer-reviewed journal or technical report); funding source (e.g., Federal/state agency or publisher/developer); and timing of analysis (i.e., primary or secondary analysis)?

Setting: (Not applicable in this case.)

Population / Participants / Subjects: (We describe the set of reports included in our review.)
Description of the participants in the study: who, how many, key features, or characteristics.

We intentionally limited our investigation to evaluations of K-12 mathematics programs determined by the WWC to have met their own standards for inclusion in ongoing syntheses of the effectiveness of educational programs. Because they were rated by two WWC reviewers as having met standards for inclusion (without or with reservations), we assume that all the evaluation reports in our sample are of relatively high methodological rigor. As of July 1, 2013, 36 research reports of 40 evaluations of 17 general education, K-12 mathematics programs conducted in the last 20 years (i.e., 1993 or after) satisfied these criteria (two reports included two target programs, one report included three).

Intervention / Program / Practice: (Not applicable in this case.)

Research Design: (Not applicable in this case.)

Data Collection and Analysis:
Description of the methods for collecting and analyzing data.

Figure 1 summarizes our framework for analyzing evaluation reports. In addition to drawing on the previously referenced arguments concerning principles of theory-based evaluation research (Bickman, 1987; Coryn et al., 2010; Lipsey, 1993; Rogers, 2007; Weiss, 1997), we follow the work of the committee convened by the National Research Council to review the quality of existing evaluation studies of mathematics curriculum materials (Confrey & Stohl, 2004). Complementing the broader principles of theory-based evaluation research with mathematics-specific criteria, they combined perspectives from both "method-oriented" evaluation (as emphasized by WWC) and "theory-driven" evaluation (as described here).

In general, our coding was limited to what was described in the evaluation reports; we did not make inferences about what evaluators did or did not *understand* about theories underlying the programs of interest, only how they *articulated* and documented the use of program theories. There were three exceptions to these limits of our coding, however. First, if some background information was not available in a report (e.g., source of funding, lead evaluator's area of training), we contacted evaluators to request it. Second, if not sufficiently clear in the report, we occasionally went to publishers' websites or other sources to determine the nature of the program (e.g., traditional, reform, etc.). Finally, although we chose not to necessarily make use of what the WWC terms "additional sources"—related, supporting documents for certain reports—we did extend our analysis to additional sources for four reports in which evaluators explicitly referenced those sources as reporting certain elements of an evaluation relevant to our analysis.

Modeled after the WWC's process, every report was examined and coded independently by two individuals employing the same coding scheme (based on the framework in Figure 1 and research questions listed above), after which discrepancies were resolved. Initially, the first and second authors independently coded three reports and made final refinements to the coding scheme. Using the same three reports and two others, the first and third authors then engaged in a similar exercise to ensure that our applications of the coding were congruent. Then, the remaining 31 reports were all coded independently by the third author and by either the first or the second author, with the first and third authors resolving all discrepancies.

We chose to employ consensus coding because of the relatively small number of reports and limited application of our scheme beyond this analysis. Having at least two individuals score every report renders strict attention to achieving and maintaining reliability standards before and during coding unnecessary. Still, overall rater agreement between the third and either first or second authors on all codes requiring some qualitative judgment coded by the third author was 92% before resolving discrepancies.

After coding was completed and all discrepancies resolved, we examined distributions and descriptives statistics for each variable, as well as cross tabulations of the possible associations identified in research question 4. With respect to the latter, to test for differences in distributions for categories within each characteristic, we employed Fisher's exact test, which is an appropriate alternative to a chi-square test when cells are populated by five or fewer instances.

Findings / Results:
Description of the main findings with specific details.

Table 1 summarizes the results of our analyses pertaining to the first three research questions. We found that 27 of the 36 reports (75%) articulated a theory for the program(s) of

interest, but that 21 of those were limited to describing only structural implementation features, with 25% (n = 9) articulating no theory at all. Regarding the quality of the program theories articulated by evaluators, only 19% (n = 7) drew broadly from research literature to explicate theories underlying the program(s) they evaluated. More (28%, n = 10) provided, at most, a brief description of the program, with no reference to the literature or even developers' intentions. Additionally, evaluators consistently made little use program theory. Typically, research questions and analyses concerned only whether a program led to differences in outcomes (64%, n = 23), of which there was typically only a single mathematics-specific measure (61%, n = 22). Similarly, only 8% of evaluations assessed process as well as structure aspects of implementation fidelity, with nearly half (47%, n = 17) failing to assess implementation fidelity at all.

In table 2 we indicate by which background and program characteristics articulation and use of program theory significantly differed, of which we highlight three examples here. First, given that secondary analysts could not have observed implementation, secondary analyses were unlikely to have included an assessment of fidelity of implementation. The other rows of the table therefore report results from tests limited to primary analyses. Second, evaluations reported in peer-reviewed outlets (journals and dissertations) scored higher in the type and quality of program theory articulated. Last, the results suggest that outside evaluators were more likely than developer-evaluators to assess implementation fidelity and include the data those assessments generate in their analyses.

Conclusions:

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

Overall, our results suggest that while it may have increased in its methodological and statistical rigor, program evaluation in K-12 mathematics has not successfully met calls to attend to underlying theories and mechanisms by which programs are intended to achieve their effects. The majority of evaluations did move beyond "black box" studies in that many specified how a program must be implemented. However, evaluators rarely articulated a testable causal chain for how a program was theorized to achieve its goals. Consequently, most of the evaluations did not investigate questions about mechanisms, and measures were often insensitive to the full array of intended outcomes of a program. Additionally, although we characterized six of the reports as achieving a "theoretical" level of articulated program theory, we found those articulations to be somewhat limited in their attention to context. None of the six reports explicated theoretical differences between the program of interest and comparison and/or other programs, and none of the six reports related program theory to resources necessary for successful implementation (e.g., professional capacity, type and amount of professional development, class size), including relating program theory to teachers' histories with similar programs and the changes and challenges that using a new program entails.

The findings of evaluation studies guide the decisions of policy makers at every level, including the adoption of both curriculum materials and intervention programs. These decisions are consequential for students' mathematics learning and academic futures. It is therefore crucial that evaluators 'get it right' when assessing the effectiveness of such programs. Our analysis indicates that WWC's methodological specifications are inadequate because they overlook understanding and using theory in evaluation design and implementation. In general, evaluation research of mathematics programs needs to improve in its attention to and use of program theory. Both method and theory are important in order to produce valid evidence on which policy and local curriculum adoption decisions can be based with confidence.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Bickman, L. (1987). The functions of program theory. *New Directions for Program Evaluation, 33*, 5-18.

Cook, T. D. & Shadish, W. R. (1986). Program evaluation: The worldly science. *Annual Review of Psychology, 37*, 193-232.

Confrey, J. & Stohl, V. (Eds.). (2004). *On evaluating curricular effectiveness: Judging quality of K-12 mathematics evaluations.* Washington, DC: National Academies Press.

Coryn, C. L. S., Noakes, L. A., Westine, C. D., & Schröter, D. C. (2010). A systematic review of theory-driven evaluation practice from 1990 to 2009. *American Journal of Evaluation, 32*, 199-226.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, *18*(1), 23-45.

Lipsey, M. W. (1993). Theory as method: Small theories of treatments. *New Directions in Program Evaluation, 57*, 5-38.

Lipsey, M. W., Crosse, S., Dunkle, J., Pollard, J., & Stobart, G. (1985). Evaluation: The state of the art and the sorry state of the science. *New Directions for Program Evaluation, 27*, 7-28.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24*, 315-340.

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.

Rogers, P. J. (2007). Theory-based evaluation: Reflections ten years on. *New Directions for Program Evaluation, 114*, 63-67.

Schoenfeld, A. H. (2006). What doesn't work: The challenge and failure of the What Works Clearinghouse to conduct meaningful reviews of studies of mathematics curricula. *Educational Researcher, 35*, 13-21.

Shadish, W. R. & Cook, T. D. (2009). The renaissance of field experimentation in evaluating interventions. *Annual Review of Psychology, 60*, 607-629.

Weiss, C. H. (1997). Theory-based evaluation: Past, present, and future. *New Directions for Program Evaluation, 76*, 41-56.

Appendix B. Tables and Figures

Not included in page count.

Category (Lipsey et al., 1985)	Dimensions of Evaluators' Articulation and Use of Program Theory					
	1) TYPE of program theory articulated (Rogers, 2007; Weiss, 1997)	2) QUALITY of evaluators' articulation of program theory (Rogers, 2007; Weiss, 1997)	3) HOW program theory is used (Rogers, 2007; Weiss, 1997)			
			3a) Research questions (Coryn et al., 2010)	3b) Construct Measurement (Coryn et al., 2010)		3c) Analysis (Coryn et al., 2010)
				i) Outcomes	ii) Fidelity	
Nontheoretical	None (or, provided only a brief description of the program's 'type'—e.g., "student centered," "explicit instruction," etc.)	Entirely implicit	Questions were limited to testing hypotheses concerning the effects of treatment and the magnitude of those effects	Employed a single outcome measure without attempting to differentiate by program intent	Did not assess implementation fidelity	Focused on whether effects were found on selected outcome variables. Beyond reporting limitations of the study, did not offer explanations for the results of the evaluation (Lipsey & Cordray, 2000)
Sub-theoretical	Implementation: specified how the program is carried out (possibly employing a 'logic model' representation) and/or specifies *what* student(s) and teacher should *do*, but not *how* or *why* (in terms of learning opportunities)	Drew from a limited range of sources (e.g., relied solely on publishers' or practitioners' perspectives and interpretations of how programs are intended to work)	Questions beyond those of overall effects concern the relationship between outcomes and implementation factors but not the causal chain	Employed at least two outcome measures, but no rationale for their validity with respect to program intent provided	Measured structural implementation constructs but not process constructs (e.g., confirmed that all components of the logic model *happened*) (Mowbray et al., 2003; Rogers, 2007)	In addition to focusing on whether effects were found on selected outcome variables, reported whether (but not *how* or *why*) there was an association between outcomes and implementation variables
Theoretical	Programmatic: Mapped out the causal chain (White, 2009), including the mechanisms presumed to link program action with student learning, including the moderator and mediator variables associated with that theory (Lipsey, 1993; Lipsey & Cordray, 2000). This mapping might (a) explicate theoretical differences among programs (e.g., goals for students' learning, process of students mathematical learning and of supporting that learning) and (b) relate program theory to resources necessary to implement the program (e.g., professional capacity, appropriate class size, amount and type of professional development—particularly as it relates to teachers' histories with similar programs and the change and challenge that a new program entails) (Confrey & Stohl, 2004)	Drew on multiple resources (logical reasoning, practitioner wisdom, prior evaluations, and social science research—i.e., mathematics education and learning sciences research)	Articulated and prioritized questions concerning the relationships (e.g., mediating, moderating) between components of the program theory (Coryn et al., 2010)	Employed a variety (e.g., topic, format, cognitive demand) of outcome measures that are valid measures of program goals and alignment with systemic factors (Confrey & Stohl, 2004)	Assessed the level of implementation fidelity (to structure *and* process) (Mowbray et al., 2003)	Focused on both whether effects were found on selected outcome variables and on explaining those results in terms of mechanisms (Birckmayer & Weiss, 2000; Lipsey, 1993; Lipsey & Cordray, 2000; Weiss, 1997)

Figure 1. Framework for analyzing research reports in our sample.

Table 1

Results pertaining to program theory articulation and use

Category	1. Type of program theory articulated	2. Quality of program theory articulated	3. Use of program theory			
			a) Research questions	b) Construct measurement		c) Analysis
				i) Outcomes	ii) Fidelity	
Nontheoretical	9 (25%)	10 (28%)	23 (64%)	22 (61%)	17 (47%)	23 (64%)
Sub-theoretical	21 (58%)	19 (53%)	13 (36%)	11 (31%)	16 (44%)	13 (36%)
Theoretical	6 (17%)	7 (19%)	0	3 (8%)	3 (8%)	0

Table 2

Articulation and use of program theory by background or program characteristics

Category	1. Type of program theory articulated	2. Quality of program theory articulated	3. Use of program theory			
			a) Research questions	b) Construct measurement		c) Analysis
				i) Outcomes	ii) Fidelity	
Type of program (curriculum, supplement, practice)						
Nature of program (reform, traditional, back-to-basics)						
Type of publication (journal, technical report, dissertation)	*	**				
Evaluator Role (developer, external evaluator)					*	**
Evaluator's background (psychology, math ed, administration, methodology)						*
Evaluator's institution (academic, private)						
Funding source (publisher, external)		*			*	
Timing of analysis (primary, secondary)	**				*	

Results of Fisher's exact test: $^{**}p < .01$; $^{*}p < 0.05$. Except for "timing of analysis," secondary analyses excluded.