Running head: TREATMENT FIDELITY AND FORMATIVE EVALUATIONS

The Role of Treatment Acceptability, Effectiveness, and Understanding in Treatment Fidelity:
Predicting Implementation Variation in a Middle School Science Program

Abstract

Evaluating a school-based program is a particular challenge when teachers are responsible for implementing the program. Variations in implementation can be difficult to measure and predict. We developed measures to explore variation in treatment implementation and serve as predictors of variation in a school-based science education program. Based on previous work, we focused on assessing treatment acceptability, effectiveness, and understanding among teachers as critical determinants of variations in program implementation using self-reported and objective measures of implementation. All three constructs were associated with implementation of the program. Our measures of these constructs show promise for use in formative and summative evaluations. Our stratification of program elements with implementation predictors can provide a template for future exploration of treatment fidelity.

Keywords: program evaluation; school-based evaluation; evaluation methods; fidelity of implementation

The Role of Treatment Acceptability, Effectiveness, and Understanding in Treatment Fidelity: Predicting Implementation Variation in a Middle School Science Program

Educators, educational researchers, and their funders are eager to understand why apparently effective educational interventions show so much variability in their success during scale up to widespread use.¹ Researchers from a variety of fields have studied treatment fidelity and found that variations in how the program is implemented can often explain differences in treatment effectiveness across program implementations during scale up. Less is known about when and why variations occur, particularly for classroom interventions (Harn, Parisi, & Stoolmiller 2013). This study focused on understanding implementation variation and treatment fidelity among teachers engaged in an instructional intervention in middle school science classrooms.

Treatment fidelity is the extent to which a program or treatment is enacted as originally intended and according to the design of program developers (Gresham & Gansle, 1993; Century, Rudnick, & Freeman 2010; Mowbray, Holter, Teague, & Bybee 2003; O'Donnell 2008; Ruiz-Primo 2006). In various literatures, it is also referred to as Fidelity of Implementation (FOI), treatment integrity, dosage, or degree of program implementation (Ruiz-Primo, 2006). Studies of treatment fidelity address not only variations in how programs are adopted, but also how variations in use affect program effects on the target outcomes.

Treatment fidelity measures have been widely applied in public and mental health fields (Kazdin, 1980; Reimers, Wacker, & Koeppel, 1987) and more recently in school psychology and assessment areas (particularly curriculum-based measurements, e.g., Allinder & Oats, 1997;

¹ Abbreviations used in this article: pedagogical content knowledge (PCK); Pedagogy of Science Inquiry Teaching Test (POSITT); science, technology, engineering, and mathematics (STEM); and South Eastern Consortium for Minorities in Engineering (SECME).

O'Donnell, 2008). However, treatment fidelity is still relatively unexplored in terms of curriculum interventions that focus on content and pedagogy, particularly those outside of curriculum-based measurements and special education applications (O'Donnell, 2008). In some applications, treatment fidelity has been conceptualized as relatively straightforward dosage or degree of implementation. However, in other applications, especially educational interventions, treatment fidelity has been defined as a multidimensional construct including other issues including how well the person implementing the program understands and can competently implement the program elements (Shulte et al., 2009).

Classroom interventions are particularly challenging to evaluate and understand because teachers are responsible for carrying out critical components of the program while maintaining the teacher's own curriculum and professional standards. Multiple demands on a teacher's time means that they often cannot enact a program exactly as intended because of overlapping programs or competing classroom goals (Harn et al., 2013). Because of the complex nature of a school-based program, and the reduced control that program developers have over implementation, researchers and program developers need strategies to track the amount of implementation of the program exhibited by the teachers and to determine why teachers vary in their implementation (Swanson, Wanzek, Haring, Ciullo, & Mc-Culley, 2011).

Given the proliferation of science, technology, engineering, and mathematics (STEM) related school interventions, extending the literature on treatment fidelity to the school-based STEM programs, as represented in this paper, will expand our knowledge not only of such interventions but also of evaluation theory. Our assessment plan drew from recent theory that emphasized the importance of using a variety of measures and a broad conceptualization of treatment fidelity in evaluating instructional interventions (Brandon, Young, Pottenger, &

Tatum, 2009; O'Donnell, 2008). We focused on assessing three constructs related to treatment fidelity—treatment **acceptability**, **effectiveness**, and **understanding**—among teachers as critical determinants of variations in program **implementation**. Our research questions focused on predicting variation in objective and self-reported implementation using measures of treatment acceptability and understanding. We also explored how these three constructs (acceptability, understanding, and implementation) are best measured and used over time to provide formative (i.e., for program improvement) and summative (i.e., documenting program effects) feedback to program developers (Chambers, 1994).

Components and Correlates of Treatment Fidelity

Treatment fidelity and predictive factors have been broken down into component parts in a number of ways by researchers seeking to explain its variability (e.g., Century et al., 2010; Mowbray, Holter, Teague, & Bybee 2003; O'Donnell, 2008; Ruiz-Primo, 2006). The primary dimension of fidelity is implementation—the degree to which the program is actually enacted as intended by the program developers (Century et al., 2010). Theorists consider both structural and process quality of implementation (Mowbray et al., 2003), dimensions that have been applied in practice (Century et al., 2010; Harn et al., 2013). Structural quality of implementation concerns more overt implementation—use of program activities/interventions—while process has to do with the quality of how the program activities/interventions are used or the quality of student-teacher interactions that arise from the program (O'Donnell, 2008). Both structural and process measures of implementation are ideal for understanding treatment fidelity in schools, although process measures are more difficult to assess and require careful definition of the program in terms of process (Harn et al., 2013; Mowbray et al., 2003).

In delving into a multidimensional definition of implementation, Reimers et al. (1987) provided a helpful model for explaining why high-fidelity treatment implementation does not always occur, using factors that are external, but related, to implementation. Most relevant to educational interventions are the concepts of treatment acceptability, perceived effectiveness, and understanding, which Reimers et al. argued would moderate implementation and, therefore, mediate treatment effects. Treatment **acceptability** is based on the *perceived* appropriateness, fairness, reasonableness, and intrusiveness of a treatment to address a particular problem (Kazdin, 1980; Reimers et al., 1987). **Effectiveness** is the perception that the treatment will impact the problem or outcomes of interest. Treatment **understanding** refers to the program implementer's comprehension of general and specific components of the program or treatment and whether their comprehension is sufficient to implement the treatment as intended by the program developers (Reimers et al., 1987). Notably, acceptability is influenced itself by the perception that there is a problem in need of a treatment, with more extreme problems being required to justify more extreme or disruptive treatments (Sterling-Turner & Watson, 2002). For school interventions, administration and community support as well as material and time costs are also determinants of acceptability (Broughton & Hester, 1993; Kurita & Zarbatany, 1991; Witt & Elliott, 1985).

In the classroom, teachers play a critical role in whether curricular programs have significant impacts because they determine whether and how much the program elements occur in the classroom (Allinder, 1996). In this context, acceptability and understanding act as gatekeepers for a program to have any impact on the target outcomes. Because teachers have so many demands on their time, they may not fully implement a program that they do not perceive as acceptable or valuable for a given problem. Likewise, if they do not understand how to

implement the program as intended by the program developers, even the most effective of programs will fail to have an impact on students.

There is some evidence that acceptability influences teachers' implementation of programs. Tanol (2010) applied the concepts of treatment acceptability to the prediction and explanation of variations in treatment fidelity in a year-long classroom intervention. Tanol reported moderate relationships ($r = .28$ to $.53$) between teacher's ratings of acceptability and their fidelity of implementations across time points within the intervention. Similarly, Allinder and Oats (1997) reported that teachers who showed higher acceptability ratings for a curriculum-based measure program used significantly more assessment probes and set higher goals than those with low acceptability ratings (with effect sizes greater than 1SD). In one of the few studies to measure both acceptability and understanding, Kurita and Zarbatany (1991) measured acceptability, time efficiency, and familiarity (which could be considered a self-reported measure of understanding) with six classroom strategies for raising student motivation and reported strong correlations between these measures ($r = .41$ to $.89$).

Timing of measurement of acceptability has been found to be important. Peterson and McConnell (1996) reported weak correlations with overall acceptability and treatment integrity, but measured acceptability pre-treatment, which Tanol (2010) hypothesized may have reduced the correlations. Greshman (2009) speculated that acceptability measured after a treatment has begun will be more accurate because the greater experience teachers have with the intervention. Peterson and McConnell (1996) also noted that acceptability of the treatments was quite high, perhaps indicating restriction of range that might reduce correlations with other variables (indeed the only significant positive correlations observed with treatment integrity were the items that had the lowest average rating for acceptability).

Measures of Components of Treatment Fidelity

The choice of measures of treatment fidelity and its predictors are critical. In particular, self-report measures of treatment fidelity and related factors like acceptability are problematic because of the inherent demand characteristics (acquiescence and/or social desirability effects) such that teachers will be inclined to overestimate their use and quality of implementation and say positive things about the program (Lee, Penfield, & Maerten-Rivera, 2009; Mullens et al., 1999; Wickstrom, Jones, LaFleur, & Witt, 1998). All measures have limitation, of course. Even using observations would overestimate structural integrity measures (e.g., frequency of use) because teachers may increase their use of program components while being observed (at least in the short-term, e.g., Wickstrom et al., 1998). In this study, we report primarily self-report measures of implementation due to the impracticality of observing teachers extensively. The other constructs were measured objectively, when feasible.

Naturally, self-report of treatment understanding is problematic as well, because teachers are unlikely to knowingly implement a program incorrectly or be aware of misunderstandings they have. Objective measures and observations of implementation are more appropriate for this construct. In contrast, acceptability must be self-reported as a perception of the program effects (rather than true effects), because it is the teacher's attitude towards the treatment that impacts use. That said, impression management is still a concern because of the potential for halo effects or a desire to please program developers by indicating positive attitudes (Beckwith & Lehmann, 1975; Randall & Fernandes, 1991).

Another issue in assessing treatment fidelity is a matter of definitions. Program developers must be clear on what elements comprise high fidelity implementation as well as how often and with what quality the elements should be implemented (preferably based on an

established program theory; Allinder, 1996; Century et al., 2010; Mowbray et al., 2003; O'Donnell, 2008). Durlak and DuPre (2008) suggest setting realistic standards for implementation and report that programs with implementation rates of 60% (meaning that individuals complete on average 60% of treatment activities) have been found to have positive effects. In addition to setting a threshold for expected implementation fidelity, documenting variability in implementation is important for anticipating program effects during dissemination and scale-up without frequent intervention by program developers (Durlak & DuPre, 2008). Based on this work, it seems likely that minimum thresholds for understanding are needed as well.

The Current Study

The basis of this paper is an NSF-funded Math and Science Partnership for which the authors serve as external evaluators (Banks, 2012). The focus of the partnership is to provide (1) teacher professional development in inquiry-based pedagogical skills and (2) instructional modules in cutting-edge areas of science (nanotechnology and 3-D imaging). Both components are expected to promote greater science achievement and motivation among students in grades 6-8 in ten struggling school districts in low-income areas of Alabama, a southeastern state in the U.S. Development of the program is on-going; therefore, a critical part of our role is providing formative evaluation information to program directors. As part of our evaluation work, we developed a battery of assessments that includes a survey on the treatment fidelity and acceptability as well as objective and self-report measures of understanding and program effectiveness. Our survey targeted the following issues as suggested by previous research (Tanol, 2010; O'Donnell, 2008; Reimers et al., 1987):

1. Is the program seen as **acceptable** (worth doing and with minimal disruption) by teachers?
2. Is the program seen as **effective** by teachers for the target problem?
3. Do teachers **understand** the program and its components?
4. Do teachers **implement** the program in the way it was intended (e.g., attending PD workshops and using modules)?

We based our measures of treatment understanding, effectiveness, acceptability, and implementation on previous research, gathered the information after teachers had participated in the program for at least one year, and incorporated self-report and objective measures where possible. This led to the need for research on the behavior of our measures of treatment fidelity and predictors of implementation and their psychometric suitability. In this study, our guiding research questions were:

1. Do the newly developed measures show strong psychometric qualities?
2. What are the relationships between the treatment fidelity predictors (acceptability, understanding, effectiveness)?
3. How do contrasting measures of program implementation and understanding (including objective measures) relate to each other and to program acceptability and effectiveness?
4. Which dimensions of treatment fidelity predict degree of treatment implementation in terms of attending PD workshops and using modules?

Questions 1-3 were evaluated using data from year 1 of the project, when implementation of the program was spotty and teachers were still learning about the program. Question 4 was assessed using year 2 data when implementation improved and data on implementation could be more reliably analyzed.

Methods

The population being studied consisted of middle school science teachers in 29 schools across ten school districts in the Black Belt region of Alabama (a historically poor region with an agriculture-based economy) participating in an educational intervention targeted to 6th through 8th grade science teachers (Banks, 2012; Tuskegee University, n.d.). The students in these schools are predominantly African American (80% of our total sample and over 95% in some schools). Free- and reduced rates are well over 90% in most of the schools the program served. The intervention consisted of 2-4 professional development workshops during the year that introduced new instructional modules, a week-long summer workshop (Summer Institute for the South Eastern Consortium for Minorities in Engineering [SECME], Jeffers, Safferman, & Safferman 2004) that introduced modules as well as inquiry-based pedagogical strategies, and instructional modules (provided with necessary materials) developed by interdisciplinary teams of science, engineering, and education faculty at participating universities who worked together to provide modules based on current nanotechnology research that align with state science standards. These teams also involved pre-service science teachers. The themes of the workshops and instructional modules were nanotechnology, 3-D modeling, and inquiry-based science instruction. The workshops primarily focused on introducing the instructional modules and relevant content.

Importantly, decisions to participate in the program were initially made at the district level, but principals and teachers varied substantially in their participation and compliance with the program and evaluation activities. Because of this program structure, this led to immediate issues for identifying participating teachers because the program defined participants as all science teachers in participating schools, but teachers were not compelled to participate in the

program or its evaluation and a substantial number were effectively non-participants in the program as well as non-participants in evaluation data collection activities. As a result, our data is based only on teachers who responded to any data collection efforts. Even so, as the results show, responding teachers still showed substantial variation in the degree to which they participated in program activities and implemented the provided curriculum materials. Therefore, we are still able to explore how this variability is predicted by other measures of program perception, even though the number of non-participating teachers is underestimated.

The instructional modules were developed by university faculty for use by middle school science teachers. The modules were designed to align with state science standards and enrich existing curricula, but were not part of a complete, year-long curriculum. The modules were mostly designed around hands-on inquiry activities and content that required one or two 45-minute class sessions. University faculty collaborators have developed around 10 modules per grade level on a variety of curriculum-relevant topics. Information about the modules is published for teachers on a project website and disseminated at workshops.

Participants

Of the 77 participants in the program in year 1 (i.e., all science teachers in participating schools), 33 completed all or some of the end-of-year assessments administered in spring of 2013 after one full year of initiating the program. In year 2, 79 teachers were eligible to participate in the program, and 36 participants responded to the end-of-year assessments in spring of 2014. There was substantial turnover in many schools, but some consistency in the teachers who participated in the survey—22 completed surveys both years. These 22 teachers represented a wide range of participating schools, coming from 8 school districts and 14 different schools.

All teachers in the program taught some combination of 6th, 7th, and 8th grade science. Teachers with 6th grade classrooms were mixed in terms of whether they served in a departmentalized department and specialized in science or if they were general classroom teachers. All 7th and 8th grade teachers were in science departments. All of the teachers in this sample attended at least one professional development workshop organized by the program during the school year. Half of the respondents in each year attended SECME during the summer. The mode for workshop attendance each year indicated that most teachers attended all of the workshops, though a substantial minority (about 25%) did not attend any workshops during the year.

Table 1

Descriptive statistics (M, SD, %) for demographic variables (both years combined)

	Variable	Descriptive (N=47) ^a
Teaching experience	Years at current school	4.64 (4.72)
	Years teaching	14.13 (13.35)
Gender	Female	82%
	Male	18%
Race/ethnicity	African American	68%
	White	32%
Highest level of education completed	Bachelor's degree	23%
	Bachelor's with additional graduate credits	14%
	Master's degree	55%
	Specialist	9%

Note. ^a Teachers who participated both years are only included once.

Measures

Data collection for treatment fidelity and predictors was organized around two dimensions: the components/predictors of fidelity (rows in Table 2) and major components of the

program (columns in Table 2). We defined the program components fairly globally (Ruiz-Primo 2006) for the purposes of measuring treatment fidelity. The major focus of the program was professional development workshops that introduced instructional modules and training in science inquiry pedagogical methods. Therefore, we assessed the treatment fidelity components and predictors for the program as a whole as well as for specific elements of the program: the professional development workshops (including summer institute), the modules, and the use of inquiry in the classroom. Sampling the key components of the program is essential for a survey that adequately represents different facets of the program and must be adapted to each program. The types of evidence gathered are specified for each cell.

Acceptability and effectiveness measures. Acceptability (8 items) and effectiveness (10 items) were assessed with self-report ratings (described in a later section). Because acceptability and effectiveness overlap significantly when asking questions about specific components of a program, we chose to focus on perceived effectiveness of the overall program and acceptability of specific components (workshops and modules).

POSITT. In addition to researcher-developed measures of acceptability, we administered the Pedagogy of Science Inquiry Teaching Test (POSITT; Schuster, Cobern, Adams, & Skjold 2012), a measure of teachers' orientation towards directive vs. inquiry-focused instruction using vignettes of science teaching. The assessment consists of 16 vignettes describing the use of inquiry-related methods in a science classroom. They then pose a question ranging from asking how the teacher should proceed with a lesson or how the reader would conduct the class differently. Each item consists of four answer choices reflecting direct didactic, direct active, guided inquiry, and open discovery teaching approaches. The measure is intended to reflect attitudes or preferences and is not scored right/wrong. For our purposes, the two directed

methods were classified as reflecting direct (teacher-centered) instructional orientation, while the latter two inquiry answer choices indicated a student-centered instructional orientation. We then created a total score for each teacher that represented the percent of discovery, inquiry-focused responses that were selected relative to the number of direct instruction responses.

Understanding and implementation. Understanding and implementation were measured with both self-report and (for understanding) an objective measure. The self-report measures (3 items each for understanding and implementation) included attitude scales similar to acceptability. See Table 3 later for the items. For understanding, the objective measure was a seven-item, multiple choice, researcher-developed content knowledge test for pedagogical content knowledge (PCK) of inquiry-related skills and project-related nanotechnology and scientific inquiry knowledge (13 and 10 items, respectively; measures studied extensively in Author, 2015). For implementation, self-reported use of instructional modules was gathered as well as objective program data on training attendance. These measures focus predominantly on structural aspects of implementation.

Two self-report methods were used to gather module use. One was a single survey item that asked whether the teacher had been using modules in their classroom. The other was a survey asking teachers to indicate each available module they had used (along with an indication of how much of the modules were used and their judgment of the quality of the module on several dimensions). There were 11 modules available at the end of the first year and around 30 modules available by the end of year two. See Figure 1 for an example of these module-related questions. We believed the checklist approach would be more accurate in gauging actual module use because teachers had to provide several pieces of information about their module use, which would make them less likely to over-report use. We also had clear instructions on the survey that

the evaluation team would not share their individual responses with the program and that they would help the program with honest feedback. Using this checklist data, we classified teachers according to how many modules they used any part of, how many full modules they used, and a dichotomous variable of whether they used any modules at all (to maximize the contrast between module use and non-use).

	If you used this module...				If you did NOT use this module, why not?
	How much of it did you use in your classroom?	Grade level	Would you use this module again?	How beneficial was this module to your students?	
Rock Cycle Microscopy (Gr. 6)	☐ used 1-2 small ideas ☐ used several ideas or activities ☐ used all or almost all of module	☐ 6 ☐ 7 ☐ 8	☐ Yes ☐ Maybe ☐ No	☐ Not at all ☐ Somewhat ☐ Very	☐ Not relevant to my class/content area ☐ Did not have materials or time to implement ☐ Did not have enough details to teach module ☐ Content too complex or confusing ☐ Not aware of this module

Figure 1. Example of a specific module and questions indicating module use. Only the first question on amount of module use was analyzed in this study.

Table 2

Framework for Treatment Fidelity and Associated Constructs

	Components			
	General	Professional Development	Instructional Modules	Inquiry PCK
Acceptability	2 survey questions	2 survey questions	4 survey questions	POSITT scores (16 items)
Effectiveness	10 survey questions	Session evaluations*	Indirectly addressed by survey questions	Not assessed
Understanding	1 survey question	Session evaluations* Nano-specific knowledge test (13 items)	1 survey question Session evaluations*	1 survey question PCK, inquiry knowledge test (7, 10 items)
Implementation/ Compliance	1 survey question	Attendance records	12 survey questions** reported use of modules	Lesson plan analysis and teaching observation***

Notes. * Not linked to teacher data and therefore cannot be analyzed in this study. ** Eleven of these questions are short surveys for each of the 11 different modules available. *** Collected for assessing the process quality of implementation, but not reported here.

Measures in year 2. Because of the difficulty in achieving an acceptable response rate in year 1, we substantially reduced our teacher end-of-year survey in year 2. We measured only the treatment fidelity components and module use. We relied on project records for variables related to workshop attendance.

Year 2 data was used to evaluate the relationship between treatment fidelity dimensions and implementation. To facilitate this analysis, we categorized teachers into four levels of implementation based on patterns in the data: (1) no implementation, (2) attended at least the summer professional development institute (most extensive opportunity to receive professional development), may have attended other PD events, but did not use modules, (3) used modules but did not attend any workshops, including the summer institute, and (4) “full” implementation of module use and at least summer workshop attendance. Because of small numbers in the sample, we also compared a dichotomous variable where teachers in categories 2-4 were grouped together so that implementers/non-implementers could be compared.

Procedures and Analysis

End-of-year surveys were administered in late spring after all professional development events were completed. In year 1, approximately one-third of the sample completed their assessments online in April (at a time and place of their choosing) while the other two-thirds completed the assessments on paper at a summer workshop event in July (this change in method was to increase response rate). In year 2, all assessments were completed online in late April to early May. Assessments were always given with generous or no time limits. Online surveys were administered using Qualtrics. All analyses were conducted in SPSS. Survey scales are available from the authors.

Surveys were completed confidentially, but not anonymously, in order to track survey response over time and match program data to evaluation results (a common limitation in evaluation data collection). Although completing identified surveys could potentially increase teachers' socially desirable responding, our distinction from the program team (and instructions, mentioned above, that indicated individual responses would not be shared with the program) were intended to promote honest responding. With the available sample sizes, factor analytic methods were not feasible. To explore scale dimensionality, we used internal consistency measures (e.g., Cronbach's alpha) as a preliminary method. We also used correlational methods, including binary logistic regression methods, to explore the relationships between scales and especially to look at the predictive value of the measures for program implementation. Group comparisons in the form of t-tests were also used, applying the corrections for unequal variances when appropriate.

Results

Psychometric Qualities (RQ1)

Our analyses began with the first research question relating to the psychometric qualities of our instruments. We addressed this question using the first year of evaluation data (2012-2013). For each scale, we examined the item means and SD for evidence of ceiling effects and internal consistency (Cronbach's alpha) statistics to determine whether the items formed a coherent, unidimensional scale.

Overall, the items contributed to the internal consistency of their intended scales and formed apparently unidimensional constructs. Key exceptions were three items asking about the understanding and involvement of members of the whole school in the project, which failed to load onto their intended scales, but did correlate strongly with each other. We thus formed a

“School-Related Acceptability” scale which seemed to capture the degree to which the teacher perceived the program is understood and accepted by other members of his or her school. Table 3 shows the descriptive and scale-level results.

Table 3

Item-level descriptive statistics for self-report survey questions

Survey items	M	SD	IC	
			Initial	Final
Implementation			0.31	--
Members of our school meet regularly to discuss the progress of our efforts for the NanoBio Science Project.	2.58	1.393	*	
I have used the NanoBio Science Project instructional modules in my teaching.	3.45	1.348	**	
In the future, I plan to use the NanoBio Science Project instructional modules in my teaching.	4.19	1.120	**	
Acceptability			0.61	0.87
Our school has enough staff, time, and other resources to really make the NanoBio Science Project pay off for the school.	2.39	1.321	*	
I would recommend the NanoBio Science Project to other teachers.	4.09	1.07		
Attending NanoBio workshops is worthwhile.	4.24	1.17		
Attending NanoBio workshops is worth the time it takes away from being in my classroom.	4.21	1.02		
I think the NanoBio Science Project instructional modules are appropriate for the content in my class.	4.16	.92		
I think many of the NanoBio Science Project instructional modules could be greatly improved.	2.85	1.25	**	
The NanoBio Science Project instructional modules are easy to implement in my class.	3.67	1.32		
Implementing NanoBio Science Project instructional modules takes too much time.	2.88	1.27	**	
Understanding			0.60	0.80
The NanoBio Science Project mission and vision are understood by members of our school.	2.97	1.42	*	
The training I receive on the NanoBio Science Project instructional modules is sufficient to understand the nano-science concepts.	3.79	1.24		
The training I receive on using inquiry-focused pedagogical methods is sufficient to for me to use those methods in my	3.82	1.33		

Survey items	M	SD	IC	
			Initial	Final
classroom.				
Effectiveness for students: “I believe the NanoBio science project will help me to increase my...”			--	0.98
students’ competence in science.	3.88	1.21		
students’ motivation to understand science in depth.	4.09	1.20		
students’ interest in science.	4.09	1.23		
students’ beliefs in importance of science.	4.03	1.23		
students’ academic performance in science.	3.90	1.30		
Effectiveness for teachers: “I believe the NanoBio science project will help me to increase my...”			--	0.97
competence in teaching science.	4.79	1.24		
confidence in using inquiry-focused teaching methods.	4.70	1.31		
interest in learning about new scientific developments.	4.76	1.23		
beliefs in importance of science.	4.81	1.31		
understanding of cutting-edge science concepts.	4.84	1.19		
School-related Acceptability			--	0.83
Members of our school meet regularly to discuss the progress of our efforts for the NanoBio Science Project.	2.58	1.39		
Our school has enough staff, time, and other resources to really make the NanoBio Science Project pay off for the school.	2.39	1.32		
The NanoBio Science Project mission and vision are understood by members of our school.	2.97	1.42		

Notes. All Ns 31 to 33. Scales ranged 1-5 from “Not at all true” to “Very true” except “Effectiveness for teachers” ranged from “About the same as before” to “Much more than before”. IC=Internal consistency. Estimates for IC provided for initial scale and final scale once problematic items were removed. * Items moved to the School-related Acceptability scale. **Items dropped or analyzed independently due to low inter-correlations

We were concerned about the presence of ceiling effects and low variability in our measures, because of the bias likely in asking active participants in a program about their evaluations of that program. One criterion of ceiling effects is that there should be at least two standard deviations between the mean score and the maximum score (Bracken 2007). Table 4 shows the descriptive statistics for each of the scales in addition to this ceiling effect index. Almost all scales showed relatively high means and little room for variability at the ceiling.

Table 4

Scale-level descriptives and ceiling effect test

Scale	N	Mean	SD	Ceiling effect ^a (>2 no ceiling)
Acceptability	33	4.07	0.90	1.03
Understanding	33	3.53	0.99	1.48
Effectiveness (Students)	32	4.00	1.19	0.84
Effectiveness (Teachers)	33	4.78	1.19	0.18
School Acceptability	33	2.65	1.20	1.97
Any level of use	20	1.65	2.81	NA ^b
Full modules used	20	0.25	0.72	NA ^b
Any module used	20	0.40	0.50	NA ^b
Workshop attendance 12-13	27	4.59	2.39	NA ^b
Used modules (single item)	33	3.45	1.35	1.15
Will use modules (single item)	32	4.19	1.12	0.73
Nano knowledge	24	10.58	2.93	2.19
Inquiry knowledge	24	7.63	3.73	1.44
PCK	16	3.06	1.61	2.44
POSITT	8	4.38	6.25	1.86

Notes. PCK=Pedagogical Content Knowledge test. POSITT is a measure of teacher' orientation towards inquiry-focused teaching. ^a Ceiling effect calculated as maximum score on the scale minus mean score, divided by SD. ^b Ceiling effects were only calculated for survey scales.

The effectiveness for teachers scale was impacted the most by ceiling effects, meaning there was not much variation in teachers' ratings of their benefits from the program because they all indicated that they benefitted greatly from the program. The effectiveness for students scale was the second most affected by ceiling effects, with somewhat more variability and less ceiling than the teacher scale. Although these ratings are an excellent result for the program, they will restrict the range of scores and attenuate any correlations with other measures.

Relationships between Treatment Fidelity and Predictors (RQ2)

Correlations between implementation and predictors of treatment fidelity are presented in Table 5. Acceptability significantly correlated with the other measures of understanding, effectiveness, and school acceptability. For the most part, measures of program acceptability, effectiveness, and other measures, showed positive correlations with each other, indicating that they measure related but distinct constructs. As expected, effectiveness and acceptability showed strong relationships, though intriguingly, this relationship was strong with teachers' impression of student effects ($r=.857, p<.01$) and weaker with their impression of their own impacts by the program ($r=.371, p<.05$).

As expected, acceptability and effectiveness for students were strongly correlated ($r=.857, p < .01$), consistent with the focus on instructional modules in questions about acceptability. Most of the correlations are strong, except for measures related to teacher effectiveness. The correlation between acceptability and teacher effectiveness was quite low ($r=.371, p<.05$). Likewise, effectiveness for students was more strongly correlated to understanding than effectiveness for teachers. These results may be due to the restricted variability of the teacher effectiveness scale due to the ceiling effect demonstrated in the previous section. Another possibility worth exploring is that teachers may focus more on student effectiveness when judging the overall value of a program rather than personal/professional benefits to themselves.

Table 5

Correlations between measures

	Acceptability	Understanding	Effective- Students	Effective- Teachers	School Accept.	Nano Knowl.	Inquiry Knowl.
Understanding	.746**	1					
Effectiveness (Students)	.857**	.679**	1				
Effectiveness (Teachers)	.371*	0.184	.411*	1			
School Acceptability	.409*	.576**	0.349	-0.184	1		
Nano knowledge	-0.125	-0.113	0.043	-0.017	-0.163	1	
Inquiry knowledge	-0.34	-0.218	-0.244	0.041	-0.331	.550**	1
PCK	-0.163	-.517*	-0.23	0.314	-.623*	0.204	0.429
POSITT	-0.069	0.481	0.427	-0.054	0.057	0.208	0.365

Notes. N=33 for treatment fidelity related scales, 23 for knowledge tests, and 15 for POSITT (which was alternated with PCK, so no teachers completed both). * $p < .05$, ** $p < .01$ ^a Because of spiraling forms, teachers received only one of the two assessments: PCK or POSITT.

Comparing Multiple Measures of Implementation and Understanding (RQ3)

It is interesting to note that self-reported understanding correlated negatively with more objective measures of understanding of PCK ($r = -.517$, $p < .05$) as well as negatively with reported module use ($r = -.324$, NS in this sample). Self-reported understanding correlated positively with *intended* module use ($r = .530$, $p < .01$). One possible explanation is that those who did not attempt to use modules overestimated their knowledge of how well they could implement the modules.

In contrast, POSITT measures of attitudes toward inquiry use in the classroom correlated positively with self-reported understanding ($r = .481$). POSITT is again more of a self-report measure, with teachers indicating which of the four given approaches to a vignette they would prefer. A positive correlation indicates that those who report greater understanding of the project also tend to select more student-centered inquiry options on the POSITT survey.

Alarminglly, objectively measured understanding correlated negatively with acceptability ($r = -.125$ to $-.340$). Although not significant with this sample size, these negative correlations

may indicate an unsuspected trend where those with greater knowledge find the program less acceptable (or perhaps those with less knowledge are not rating acceptability with the accuracy they would if they better understood the project). The unexpected negative correlations observed warrant further exploration in larger samples to determine if teachers without extensive knowledge of a program tend to give it the “benefit of the doubt” with higher evaluation scores.

Descriptive Statistics and Relationships with Implementation Measures in Year 2 (RQ4)

In year 2, all of the fidelity-related dimensions, except teacher effectiveness, increased somewhat over the previous year. See Table 6. This is consistent with our observation that implementation increased substantially in year 2, although implementation was still far from the desired level. In year one, just 8 out of 77 teachers reported using modules. In year 2, 26 out of 79 teachers (or 26 out of 36 that completed surveys) reported using at least one module. In addition to the increased means, we also observe a narrowing of the ratings (i.e., smaller SDs). This mitigated the effects that mean changes would have on the ceiling effects of the scales—all of the scales maintained roughly the same amount of ceiling.

Measures of implementation are problematic for year-long programs such as this one. Direct measurement of module use is often not feasible as it would require extensive monitoring, but self-report is subject to a number of potential biases from respondents. Because year 1 implementation was weak, we used year 2 implementation data to study the relationship between the treatment fidelity dimensions and implementation in the form of instructional module use and attendance at workshops.

Table 6

Scale-level descriptives and ceiling effect test from year 2 data

Scale	N	Mean	SD	Ceiling effect ^a
-------	---	------	----	-----------------------------

				(>2 no ceiling)
Acceptability	37	4.42	0.59	0.99
Understanding	37	3.81	0.73	1.64
School Acceptability	37	3.15	0.93	2.00
Effectiveness (Students)	37	4.20	0.82	0.97
Effectiveness (Teachers)	37	4.44	0.72	0.78
Count of module use (self-report)	38	1.61	1.70	38
Module use (yes/no used any module)	40	0.68	0.47	39
Count of module use (program records)	69	0.28	1.04	69
Attended summer workshop (yes/no)	69	0.25	0.43	NA
Other workshops (0-3 events)	69	0.74	0.83	NA
Implementation variable (4 levels)	69	.00 ^b	NA	NA
Implementer (yes/no)	69	0.22	0.42	NA

Notes. ^a Ceiling effect calculated as maximum score on the scale minus mean score, divided by SD. ^b Ordinal variable, median reported.

The correlations between treatment fidelity dimensions were similar to those seen in year 1 of the program (Table 7). All of the fidelity-related dimensions correlated significantly with the involvement/implementation variables with the exception of school acceptability. Personal acceptability showed the strongest relationships with implementation variables.

Acceptability of the program was significantly correlated with module use when recorded as a dichotomous variable ($r = .352$, $p < .05$), but not when an absolute count of module use was used ($r = .207$). Teacher effectiveness was also associated with module use (whether this is an effect of using modules—teachers felt like they benefitted from them—or simply a covariation is unclear). Attending workshops was not associated with the treatment fidelity dimensions.

Table 7

Pearson correlations between measures in year 2 data

	Acceptability	Understanding	School Accept.	Effective. (Students)	Effective. (Teachers)	Mod (any)	Mod (0/1)	Module evaluations	SECME attendance	Workshop attendance
Understanding	.617**									
School accept.	.209	.538**								
Effective (students)	.756**	.444**	.080							
Effective (teachers)	.516**	.348*	.050	.360*						
Module use (count)	.207	.126	.204	.082	.183					
Module use (0/1) ^a	.352*	.179	.265	.226	.394*	.650**				
Module evaluations ^b	.214	.108	.274	.039	-.048	.453**	.249			
SECME attendance	.272	.168	.097	.098	.169	.295	.140	.271*		
Workshop attendance	.374*	.308	.171	.243	.185	.418**	.103	.355**	.789**	
Involvement (0/1)	.422**	.344*	.147	.214	.380*	.414**	.489**	.466**	.532**	.464**

Notes. N=36 * $p < .05$, ** $p < .01$ ^a Point-biserial correlations calculated where appropriate. ^b

Teachers were asked to submit evaluations of modules used. This variable reflects the number of evaluations submitted and is a more conservative of modules used as it required more effort for teachers to document their use of modules and was less sensitive to over-reporting.

When the teachers were divided into implementer and non-implementer groups (or compliant and non-compliant), significant differences were found in terms of acceptability ($t(35) = 2.75$, $p < .01$, $d = 0.9$), understanding ($t(35) = 2.17$, $p < .05$, $d = 0.7$), and teacher effectiveness ($t(35) = 2.55$, $p < .05$, $d = 0.8$). See Table 8. All of these differences showed large effect sizes, indicating important differences in these dimensions for teachers who implemented at least part of the program versus those who were nominally participating, but not implementing any part of the project in year 2. It is interesting that, despite these differences, non-implementers still give relatively high ratings of the program even though that they choose not to implement any part of the program.

Table 8

Differences in Treatment Fidelity Facets by Implementation

	No implementation	Implementation	d	t	Sig. (2-
--	-------------------	----------------	---	---	----------

Involvement_SECME_01	N	M	SD	N	M	SD	tailed)		
Acceptability	20	4.19	0.63	17	4.68	0.41	0.92	2.75	.009
Understanding	20	3.58	0.76	17	4.08	0.60	0.72	2.17	.037
School Acceptability	20	3.02	0.83	17	3.29	1.03	0.29	0.88	.386
Effectiveness (Students)	20	4.04	0.95	17	4.39	0.61	0.44	1.35	.188
Effectiveness (Teachers)	20	4.19	0.83	17	4.73	0.41	0.82	2.55	.016

To explore the predictive utility of the fidelity-related measures, a binary logistic regression model was used to determine which measures were associated with a teacher being classified as an “implementer” vs. “non-implementer”. In the empty model, 54.1% of teachers were correctly classified (Wald=0.243, $p>.05$). The final model (using a forward-conditional variable selection procedure), included only the overall acceptability of the program as a significant predictor, and increased the accuracy of the model to 70.3% ($\chi^2(1)=7.739$, $p<.01$; Nagelkerke $R^2=.252$; Wald=5.552, $p<.02$). Acceptability was strongly related to being an “implementer” as those with 1 unit higher acceptability score were over 7 times as likely to be in the implementer group ($\text{Exp}(\beta)=7.259$). None of the other fidelity-related measures (understanding, perceived effectiveness, school acceptability) were significantly related to implementation in this model once acceptability was included. When other dependent variables (such as number of modules used, number of workshops attended, or dichotomous versions of these variables) were tested similarly, acceptability was again the best and only significant predictor of implementation as measured by the other indicators.

Discussion

Our research questions focused on the feasibility of measuring implementation and predictors of treatment fidelity using self-report and the initial psychometric and validity evidence for self-report measures of treatment acceptability, effectiveness, and understanding.

Previous research (Century et al., 2010; Mowbray et al., 2003; O'Donnell, 2008) has established the need for such measures, but appropriate self-report measures would be beneficial to this area of evaluation. Our purpose was to explore how these constructs are best measured to provide formative and summative feedback to program developers.

Among three predictors of treatment fidelity that we measured (acceptability, understanding, and student effectiveness), there was a strong positive relationship in year 1 ($r = .68$ to $.86$). Not surprisingly, effectiveness for students and acceptability of the program were the most strongly related ($r = .86$) and this may reflect a single underlying attitude towards the program. The strong correlation of the various attitudes measure may reflect a pervasive acquiescence bias, leniency, or halo effect. However, these moderate to strong relationship between facets are consistent with previous work (Tanol, 2010; Kurita & Zarbatany, 1991).

The other two potential predictors (teacher effectiveness and school acceptability) showed modest relationships with the other fidelity facets ($r = .18$ to $.57$). In looking at ceiling effects, we found teacher effectiveness had the least room for variability, with the likely effect of attenuating its correlations with other measures. This explanation is made more plausible by year 2 data, where the ceiling effect for this scale was less pronounced and correlations with other scales were stronger. Prior work has not distinguished between effectiveness for particular stakeholders, but these initial findings indicate that there may be differentiation between the perceptions of effects by teachers for themselves versus for their students.

In addition to self-reported attitudes, we gathered objective measures of understanding in year 1 in the form of multiple-choice tests of pedagogical content knowledge related to inquiry and directly tested knowledge of inquiry terms and methods. . These measures, though lacking power in this study, appeared to show weak or negative correlations with acceptability ($r = -.13$

to -.34), understanding ($r = -.11$ to $-.52$), and effectiveness for students ($r = -.24$ to $.04$). This indicates that perhaps either the participants are not aware of their level of understanding or that perceptions of program usability (i.e., subjective understanding) are independent of objective knowledge of program content. It may also reflect problems with the measures themselves, which requires further validity research (initial validity explorations are reported in Author, 2015). Future research should consider whether subjective and objective understanding reflect separate constructs and explore whether it is feasible to measure understanding with self-report. Our initial results indicate that self-report is not effective as a proxy for objective measures.

The ultimate goal of this study was to predict implementation. Evidence from year 2 of the program provided such predictive validity evidence. Using year 2 data, we found that acceptability, understanding, and teacher effectiveness were all associated with whether a teacher implemented the program (either by attending the summer workshop or using one or more modules). These results are consistent with prior work where correlations with implementation dimensions were similar in magnitude (Tanol, 2010; Kurita & Zarbatany, 1991). Small sample size precluded the use of ANOVA with more levels of implementation. Acceptability was the most sensitive construct, showing large differences between teachers who did and did not implement the program. It is important to note that since we used measures from the end of the school year, it is not clear whether acceptability causes implementation or is an outcome of implementation. Likely, it is a combination of both. Longitudinal studies could explore this phenomenon, but would require a larger sample and less teacher turnover than our project provided.

One limitation of this study is the use of primarily self-report measures of implementation. Because of the need to use self-report, the measures of module use (though not

workshop attendance) was also impacted by positive response biases like social desirability (see previous work on self-report vs. objective observations, Mullens et al., 1999; Wickstrom et al., 1998). This may inflate overall scores as well as inflating the correlations between measures. In the future, we hope to gather more objective measures of module use that will allow better estimation of the relationship between objective and self-report measures of components of treatment fidelity.

Even accounting for over-reporting of implementation, it should be noted that the implementation level of the full project seems to be much lower than even the teacher surveys indicate. We must assume that most of the 42 nominally participating teachers who did not return surveys were not implementing the project (or at least not with high fidelity though they might make some use of the materials). Of the 42 non-implementing teachers, just 16 attended one or more of the workshops (most attended just the workshop at the beginning of the year; just 1 teacher attended all three but did not submit a survey). This indicates that the program is not being implemented with high fidelity by many teachers in participating schools. We suspect this is a common challenge for classroom-based interventions where teachers have many other competing programs and instructional materials. Additional work is needed to explore how to increase response rate among non-implementing teachers who are very likely to be missing from the survey data. Future research with other, larger intervention programs (particularly other Math-Science Partnership programs) is also needed to explore how these results vary or replicate for other contexts and programs

Surprisingly, despite low implementation, the teachers rated the program as beneficial to themselves and their students and are almost universally positive about the program. We suspect this paradox is common in educational interventions because teachers are inundated with

seemingly good (or well-intentioned) interventions but lack the knowledge, motivation, or time to implement the program as intended. The many demands placed on teachers in the classroom no doubt leads to many competing demands. Thus, they may find programs attractive and want to say positive things about it, even though they are not actually implementing the program themselves. Measuring treatment fidelity and understanding its predictors is therefore likely complicated by the response biases at work on self-report measures. Encouraging honest assessments of program usefulness and implementation is therefore critical when objective measures of treatment fidelity and its components are not possible. Related to this issue, another line of research needed is an exploration of teachers' decisions about selecting instructional interventions when they are presented with multiple competing, but attractive, choices. Understanding the process used and the tradeoffs they make in designing their instruction could inform program development. The role of school administration involvement (and administration support/encouragement of program participation) should also be explored.

Conclusions

We believe the evaluation literature is benefitted by authentic descriptions of evaluations that are far from ideal, but representative of feasible educational evaluation practice. Our program is still a work in progress as is our evaluation of it. As the program continues to evolve, we believe our measures of treatment fidelity and beliefs that predict implementation will continue to be valuable in understanding when the program is being implemented and the quality of that implementation. Gathering evidence on treatment implementation has important implications for formative program evaluation. In this program, the developers took the low rate of implementation in year 1 seriously and enacted new mechanisms in the next year of the program to set clearer expectations for module use and make access to instructional modules

easier. Such positive outcomes are a key goal in formative evaluations and encouraging for this line of research. Predicting low implementation may be a powerful tool for improving treatment fidelity.

References

Author, 2015

- Allinder, R. M., & Oats, R. G. (1997). Effects of acceptability on teachers' implementation of curriculum-based measurement and student achievement in mathematics computation. *Remedial & Special Education, 18*, 113-120.
- Allinder, R.M. (1996). When some is not better than none: Effects of differential implementation of curriculum-based assessment. *Exceptional Children, 62*(6), 525-535.
- Banks, C. (2012, February). *Tuskegee University Math and Science Partnership (MSP): A NanoBio Science for the Alabama Black Belt Region*. Poster presented at the annual MSP grantees meeting, Washington, DC. Retrieved April 15, 2014 from <http://posterhall.org/msp2012/posters/55>
- Beckwith, N. E., and Lehmann, D. R. (1975). The importance of halo effects in multi-attribute attitude models. *Journal of Marketing Research, 12*, 265-275.
- Bracken, B. A. (2007). Creating the optimal preschool testing situation. In B. A. Bracken, and R. J. Nagle (eds.), *Psychoeducational assessment of preschool children* (4th ed., pp. 137-154). Mahwah, NJ: Lawrence Erlbaum Associates.
- Broughton, S.F., & Hester, J.R. (1993). Effects of administrative and community support on teacher acceptance of classroom interventions. *Journal of Educational and Psychological Consultation, 4*, 169-177.
- 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.
doi:10.1177/1098214010366173

- Chambers, F. (1994). Removing confusion about formative and summative evaluation: Purpose versus time. *Evaluation and Program Planning*, 17(1), 9-12.
- Durlak, J.A., & DuPre, E.P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, 41:327–350. DOI 10.1007/s10464-008-9165-0
- Gresham, F. M. (2009). Evolution of the Treatment integrity concept: Current status and future directions. *School Psychology Review*, 38(4), 533-540.
- Gresham, F. M., & Gansle, K. A. (1993). Treatment integrity of school-based behavioral intervention studies: 1980-1990. *School Psychology Review*, 22, 254-273.
- Harn, B., Parisi, D., & Stoolmiller, M. (2013). Balancing fidelity with flexibility and fit: What do we really know about fidelity of implementation in schools. *Exceptional Children*, 79(2), 181-193.
- Jeffers, A. T., Safferman, A. G., & Safferman, S. I. (2004). Understanding K-12 engineering outreach programs. *Journal of Professional Issues in Engineering Education and Practice*, 130(2), 95-108.
- Kazdin, A.E. (1980). Acceptability of alternative treatments for deviant child behavior. *Journal of Applied Behavior Analysis*, 13, 259-273.
- Kurita, J. A. & Zarbatany, L. (1991). Teachers' acceptance of strategies for increasing students' achievement motivation. *Contemporary Educational Psychology*, 16, 241-253.
- Lee, O., Penfield, R., & Maerten-Rivera, J. (2009). Effects of fidelity of implementation on science achievement gains among English language learners. *Journal of Research in Science Teaching*, 46(7), 836-859.

- Mowbray, C., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, 24(3), 315-340.
- Mullens, J.E., Gayler, K., Goldstein, D., Hildreth, J., Rubenstein, M., Spiggle, T., Walking Eagle, K., & Welsh, M. (1999). *Measuring classroom instructional processes: Using survey and case study field test results to improve item construction* (Working Paper 1999-08). Washington, DC: Policy Studies Associates, Inc. Retrieved July 1, 2013 from <http://files.eric.ed.gov/fulltext/ED434144.pdf>
- 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.
- Randall, D. M., & Fernandes, M. F. (1991). The social desirability response bias in ethics research. *Journal of Business Ethics*, 10(11), 805.
- Reimers, T.M., Wacker, D.P., & Koepl, G. (1987). Acceptability of behavior interventions: A review of the literature. *School Psychology Review*, 16(2), 212-227.
- Ruiz-Primo, M.A. (2006). A Multi-method and multi-source approach for studying fidelity of implementation (CSE Report 677). SEAL, Stanford University/CRESST.
- Schuster, D., Cobern, W.W., Applegate, B., Schwartz, R., Vellom, P., & Undreiu, A. (2007). Assessing pedagogical content knowledge of inquiry science teaching— Developing an assessment instrument to support the undergraduate preparation of elementary teachers to teach science as inquiry. *Proceedings of the National STEM Conference on Assessment of Student Achievement*, hosted by the National Science Foundation and Drury University, Washington DC, October 19-21, 2007.

- Shapiro, E. S., & Eckert, T. L. (1994). Acceptability of curriculum-based assessment by school psychologists. *Journal of School Psychology, 32*(2), 167-183.
- Shulte, A.C., Easton, J.E., & Parker, J. (2009). Advances in treatment integrity research: multidisciplinary perspectives on the conceptualization, measurement, and enhancement of treatment integrity. *School Psychology Review, 38*(4), 460-475.
- Sterling-Turner, H. E., & Watson, T. S. (2002). An analog investigation of the relationship between treatment acceptability and treatment integrity. *Journal of Behavioral Education, 11*, 39-50.
- Swanson, E., Wanzek, J., Haring, C., Ciullo, S., & Mc-Culley, L. (2011). Intervention fidelity in special and general education research journals. *The Journal of Special Education, 45*, 1-11. doi: 10.1177/0022466911419516
- Tanol, G. (2010). Treatment fidelity: relation to treatment acceptability and change over time. University of Minnesota (dissertation).
- Tuskegee University (n.d.) Math and Science Partnership. [Program description and instructional modules]. Retrieved from http://www.tuskegee.edu/math_and_science_partnership_msp.aspx
- Wickstrom, K., Jones, K., LaFleur, L., & Witt, J. (1998). An analysis of treatment integrity in school-based behavioral consultation. *School Psychology Quarterly, 13*, 141–154.
- Wilder, D. A., Atwell, J., & Wine, B. (2006). The effects of varying levels of treatment integrity on child compliance during treatment with a three-step prompting procedure. *Journal of Applied Behavior Analysis, 39*(3), 369-373.
- Witt, J. C., & Elliott, S. N. (1985). Acceptability of classroom intervention strategies. In T. R. Kratochwill (ed.), *Advances in school psychology* (pp. 251-288). Hillsdale, NJ: Lawrence Erlbaum.

Running head: TREATMENT FIDELITY AND FORMATIVE EVALUATIONS

The Role of Treatment Acceptability, Effectiveness, and Understanding in Treatment Fidelity:
Predicting Implementation Variation in a Middle School Science Program

Joni M. Lakin

David M. Shannon

Auburn University

Author Note

Joni M. Lakin, Ph.D., Department of Educational Foundations, Leadership, and Technology, Auburn University. David M. Shannon, Ph.D., Department of Educational Foundations, Leadership, and Technology, Auburn University.

Correspondence concerning this article should be addressed to Joni Lakin, Department of Educational Foundations, Leadership, and Technology, (+1)334-844-4930, 4036 Haley Center, Auburn University, AL 36849, USA. Email: joni.lakin@auburn.edu

Vitae

Joni M. Lakin, Ph.D. The University of Iowa, is Assistant Professor of Educational Foundations, Leadership, and Technology at Auburn University. Her research interests include educational assessment, educational evaluation, and increasing diversity in STEM fields. Her

recent work has focused on educational measurement and program evaluation for STEM educational interventions.

David M. Shannon has a Ph.D. in Educational Research and Evaluation Methodology from the University of Virginia and is currently the Humana-Sherman-Germany Distinguished Professor at Auburn University. He teaches graduate courses in measurement, research, and evaluation and his research interests include student and teacher evaluation and methodological issues.

Funding Acknowledgment

This material is based on work supported by the National Science Foundation under Grant No. MSP-1102997.