Studies in Educational Evaluation 49 (2016) 51–65

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

Defining and assessing FoI in a large-scale randomized trial: Core components of values affirmation☆

Dominique N. Bradley[a], Evan P. Crawford[b,*], Sara E. Dahill-Brown[c]

[a] Educational Leadership and Policy Analysis, University of Wisconsin-Madison, 253 Education Building, 1000 Bascom Mall, Madison, WI 53706, United States
[b] Political Science, Fellow, Wisconsin Center for Education Research, University of Wisconsin-Madison, 110 North Hall, 1050 Bascom Mall, Madison, WI 53706, United States
[c] Politics and International Affairs, Wake Forest University, 325 Kirby Hall, PO Box 7568, Winston-Salem, NC 27109, United States

ARTICLE INFO

Article history:
Received 5 October 2015
Received in revised form 11 April 2016
Accepted 28 April 2016
Available online 20 May 2016

Keywords:
Implementation fidelity
Self-Affirmation
Values-Affirmation
Randomized controlled trial
Achievement gap
Scaled-up replication
Survey

ABSTRACT

A growing body of work suggests that values affirmation can serve as a simple, powerful tool for reducing achievement gaps. The dramatic results of these studies have been shared with and discussed by educators, researchers, and policy-makers, spurring excitement about deploying the intervention in schools around the country. Scholars grasp of the mechanism by which the intervention alters student achievement is limited. We develop a framework for assessing fidelity of implementation by identifying the most crucial elements of an ideal classroom-administered values affirmation. We apply this framework to data from a district-wide randomized trial of values affirmation. Our descriptive analysis shows that fidelity varied across schools, teachers, and over time. We believe that out results make a strong case for future implementations to take fidelity into account. Assessment of fidelity of implementation using a critical components framework will ensure better understanding of variation in the impacts of values-affirmation. Our data support integrating teachers more fully into the process of delivery, though we strongly caution that the integration of these written interventions into regular curriculum must be handled carefully. Lastly, even with the threats to the fidelity of delivery and stealth we find significant impacts of the intervention on the intended audience (Borman, Grigg, & Hanselman, 2016). These significant positive impacts despite low fidelity of some criticial components suggests that there may be aspects of the intervention that are less important as they are defined and understood in the original protocol.

1. Introduction

Studies suggest that values affirmation exercises, in which individuals are instructed to identify, reflect on, and write about their core values, can serve as simple, yet powerful tools for reducing achievement gaps (Cohen, Garcia, Apfel, & Master, 2006; Cohen, Garcia, Purdie-Vaughns, Apfel, & Brzustoski, 2009; Miyake et al., 2010; Sherman et al., 2013). However, dramatic results do not always manifest (Borman, 2012; Kost-Smith et al., 2012; Dee, 2014). Our district-wide randomized trial of values affirmation among 7th graders revealed important, but relatively modest, impacts that varied across schools. These findings contrast with the significant and substantively larger impacts produced by earlier studies (Cohen et al., 2006; Miyake et al., 2010). This is not an unusual story in the study of educational interventions (Makel & Plucker, 2014). What might explain these divergent findings? In this paper, we define and measure fidelity of implementation for values affirmation interventions, and investigate variability in implementation across schools and classrooms as one possible contributing factor in these inconsistent effects. Our aims are twofold: we contribute to the specification and development of this promising classroom intervention, and simultaneously model a process that will inform the scaling and field implementation of other social-psychological classroom interventions.

Fidelity has not often been an explicit consideration in most prior studies of values affirmation. Research-based classroom interventions often employ small samples, highly controlled

☆ The authors share equal responsibility for this manuscript and are listed alphabetically. Research on this paper was supported by grants from the Institute of Education Sciences, U.S. Department of Education (R305A110136 and R305C050055) and Spencer Foundation (201500044). Findings and conclusions are those of the authors and do not necessarily reflect the views of the supporting agencies.

* Corresponding author.
E-mail addresses: dust@wisc.edu (D.N. Bradley), crawford3@wisc.edu (E.P. Crawford), dahillse@wfu.edu (S.E. Dahill-Brown).

http://dx.doi.org/10.1016/j.stueduc.2016.04.002
0191-491X/

implementations, and close contact with the teachers who deliver the intervention, all of which ease monitoring processes and ensure high levels of control, obviating a need to specify and measure implementation fidelity. We therefore begin by employing a core implementation components approach to reexamine the literature on values affirmation and draw from it the most crucial elements of a classroom-administered values affirmation intervention. Because our study evaluates a scaled-up replication of the intervention developed by Cohen et al. (2006, 2009), we pay particular attention to their hypotheses about which aspects of the intervention are most crucial to improve student outcomes. Our examination therefore supports the development of fidelity measures for such an intervention. In so doing, it serves to illuminate challenges to implementation that arise in the field, raising questions about the conditions necessary for a values-affirmation intervention to positively affect student outcomes. These are important steps to take to facilitate this interventions' deployment across several schools. On whole, specifying and measuring fidelity is important for translating educational research into sustainable classroom practice, and the process we model here is one that might be productively utilized by educational researchers developing other interventions.

In the following pages we discuss approaches to the study of fidelity and apply them to the literature on values affirmation. We then describe our measurement tools, analysis, results, and conclude with a discussion of implications for future research as well as for development of these exercises for broader use.

We find substantial variation in fidelity across schools and classrooms, particularly in terms of teacher delivery of the intervention. By the fourth affirmation exercise, notable declines in both student and teacher engagement with the exercise is observed. Our evidence reveals that several components of implementation were executed with fidelity, such as the time and place of the intervention. At the same time, we identify a tension between two core implementation components of the ideal intervention, stealth and low-stress environment; this internal conflict may explain the variation in teacher delivery that took place, despite the detailed training and manuals provided.

2. Theoretical framework

2.1. Conceptualizing and measuring fidelity in the context of values affirmation

Fidelity (also referred to as program integrity) can be broadly defined as the extent to which an intervention is implemented in accordance with the intentions of the designers. A substantial amount of work suggests that failure to achieve fidelity explains the disappointing results of many promising programs (Dusenbury, Brannigan, Falco, & Hansen, 2003; Mihalic, 2004). The extent to which fidelity varies over the course of an intervention can substantially mediate efficacy. This has been referred to as an "implementation gap," and is especially likely when fidelity is not adequately addressed during research design (Durlack & DuPre, 2008; Lipsey, 2009). Therefore, when researchers are piloting, evaluating, and scaling interventions, measuring fidelity is important to establish internal validity and avoid compromising external validity – as well as to maximize statistical power for detecting effects and causal heterogeneity (Cook & Poole, 1982; Chen & Rossi, 1983; Dumas, Lynch, Laughlin, Phillips Smith, & Prinz, 2001; Maynard, Peters, Vaughn, & Sarteschi, 2013). Proper evaluation of implementation fidelity can also help intervention developers make adjustments to program design that lead to improved fidelity in the future (Lakin & Shannon, 2015).

Monitoring, documenting, and measuring fidelity is important for setting expectations regarding which components of an intervention are likely to be successfully transported to other sites, as well as identifying which components are likely to pose challenges for scale-up and broader implementation (Fagan, Hanson, Hawkins, & Arthur 2008; Esbensen, Matsuda, Taylor, & Peterson 2011). This is particularly crucial in designing interventions for use by human service based organizations such as schools, complex institutions "with hundreds of thousands of practitioners situated in a variety of provider organizations that function within uniquely configured state and federal service systems" (Fixen, Blase, Naoom, & Wallace, 2009; p. 532). Identification of challenging components can lead to beneficial changes in the intervention before scale-up has been initiated.

During the 1970s policy analysts, evaluators and methodologists, many of whom were concerned with policy implementation in educational settings, began earnestly investigating the disconnect between interventions as conceived and their implementation (Lipsky, 1971; Pressman & Wildavsky, 1973; Berman & McLaughlin, 1976; Fullan & Pomfret, 1977; Hall & Loucks, 1977; Berman, 1978; Sechrest, West, Phillips, Redner, & Yeaton, 1979). Several studies later focused on meta-analyses assessing the impact of fidelity of implementation on intervention effects. Their findings indicated that interventions designed with measurement tools for implementation identified positive, measureable impacts of fidelity on desired outcomes (Mihalic, 2004; Blase & Fixsen, 2013). However, education and policy scholars debated the merits of maximizing fidelity when doing so would limit the ability to make necessary intervention adaptations for local contexts, thereby potentially jeopardizing the quality of the intervention (Mowbray, Holter, Teague, & Bybee, 2003). Policy scholars established that implementers – often 'street-level' bureaucrats like teachers and police officers – might subvert the original intent of an intervention but could also profitably adapt it to a local context (Elmore, 1979). Methodologists, less concerned with success than accurate estimation, pointed out that effect size estimates were also potentially misleading given the heterogeneous quality of implementation efforts. These early scholars generally recommended designing policies to constrain implementers as a way of ensuring higher fidelity of implementation.

The concept of fidelity has since come to mean more than simply the extent to which implementers deliver an intervention as intended; it is now widely regarded as a multidimensional construct referring to aspects of delivery and receipt, in which fidelity can vary by programmatic characteristics as well as by features of the settings in which interventions are being placed (Stein et al., 2008; Zvoch, 2012). The meaning of fidelity thus varies widely in its manifestations across context and discipline. Several frameworks have emerged, among them the core implementation components approach (CIC), the five dimensions approach, and the structure and process approach (Dane & Schneider, 1998).

The CIC approach builds on Hall and Hord's emphasis on the "building blocks" of an intervention (1987, p. 117) and seeks to address implementation fidelity and preserve the integrity of an intervention by identifying its most "essential and indispensable" elements, those that directly impact the intended outcomes (Wallace, Blase, Fixen, & Naoom, 2005; Fixen et al., 2009; Protheree, 2009). Many scholars and agencies have similarly emphasized a focus on "critical components," "essential characteristics," and "critical parts," among others (Century, Rudnick, & Freeman, 2010). For example, in an effort to mediate the tension between quality implementation and the potential hazards of an overly stringent intervention design, the U.S. Department of Education (USED, 2009) stated "Quality implementation can be defined as the effective delivery of a program's core components to its target audience". They recommended that in order to ensure

quality of implementation and limit the potential impact of adaptations on intervention design, the core components of the intervention necessary to achieve the intended outcomes should be identified.

The five dimensions approach originated from a close review of substance abuse prevention program studies and proposed a set of five elements to define fidelity of implementation (Dane & Schneider, 1998). In studies where program integrity was explicitly considered, the authors identified: adherence, exposure, quality of delivery, participant responsiveness, and program differentiation.[1] Overall, the five dimensions approach succeeds in identifying possible measures (i.e. dosage) as well as important, common aspects of implementation that should be considered in the assessment of fidelity (i.e. participant responsiveness), and points to a possible need for planned differentiation, but in spite of its relative popularity as a guiding framework (Sanetti & Kratochwill, 2009), the dimensions themselves are not analytically distinct from one another and do not adequately support the process of considering and systematically measuring fidelity. Adherence, as defined by Dane and Schneider, is defined similarly to the umbrella concept of fidelity; for that reason the terms have at times been used interchangeably (Dusenbury et al., 2003; Lynch & O'Donnell, 2005). Scholars in the health sciences often define adherence in way that is closer to Dane and Schneider's concept of participant responsiveness (Gearing et al., 2011). In particular, we believe that 'quality of delivery' as defined by Dane and Schneider fails to distinguish adequately between the intervention itself and the environment in which it takes place.

While we acknowledge that "the definition of what is contextual and what is part of the intervention can always shift" (Century et al., 2010, p. 208), we believe it is the purpose of a framework to help distinguish between antecedent environmental conditions which may moderate both fidelity and effect size and the processes that constitute fidelity of implementation itself. For example, in the context of a values affirmation, the level of stereotype-threat present in a school, as well as which identities and stereotypes were salient in that school, might constitute an antecedent environmental condition that could mediate the efficacy of a values affirmation intervention but which would not be appropriately identified as an element of implementation fidelity. We therefore turn our attention to the CIC framework and to the structure and process scholarship, keeping these five dimensions and the measures they recommend in mind.

Rather than conceptualize fidelity as a construct with common features that exist independent of the varied theories guiding interventions, the CIC approach provides a template for an intervention-specific process of defining and assessing fidelity rather than promoting a generalizable model of fidelity.[2] This approach recommends using the particular theories undergirding the intervention in order to determine the most essential elements necessary in order for a particular treatment to have its intended effect. Those who employ a CIC approach also recommend that program evaluators explicate ways an intervention can 'break' during implementation. The CIC approach guides evaluators to go beyond simply determining which essential components must be supported. They must also consider which adaptations are acceptable and therefore encouraged as well as which behaviors constitute an unacceptable adaptation (Waltz, Addis, Koerner, & Jacobson, 1993; Huntley, 2005). In an educational setting – where teachers are implementers who will need to innovate and where some aspects of the intervention may not be completely unique to the intervention (i.e. practices that take place daily in schools) – the clear identification of an intervention's most critical ingredients as well as the overt identification of opportunities for flexibility may be especially useful (Harn, Parisi, & Stoolmiller 2013). These distinctions may also guide the development of measurements and training materials.

Structure and process scholars provide some additional guidance for defining the core components of an intervention by recommending a distinction between structural elements, such as those related to pre-existing resources and process-related components more directly related to intervention delivery (Wang et al., 1984; Mowbray et al., 2003; Lynch & O'Donnell, 2005; Lastica & O'Donnell, 2007; Century et al., 2010). We conceive of core structural components as conditions or actors over which researchers have little control (structure). These should be specified in a model of the intervention, as structural components may influence selection of test-cases, design of the intervention, fidelity itself, and how the intervention affects the targeted population. We conceive of the core process components of implementation as those over which researchers and implementers have more influence during design, delivery, and receipt.[3] For example, if a school district mandated that values-affirmation exercises be administered by all homeroom teachers, it could be imagined that special instructions would be created to help homeroom teachers deliver the intervention in a way that would maintain fidelity. The mandate that homeroom teachers deliver the intervention is a structural component that is not controlled by the teachers, the students, or researchers. The design of the intervention itself along with the instructions created for teachers and their ability to correctly follow those instructions would all be considered core *process* components.

In what follows, we apply the CIC approach, to conceptualize fidelity for a values affirmation intervention designed to reduce stereotype threat and decrease achievement gaps among middle school students. We identify core (essential) components of the intervention, identify areas where differentiation may be acceptable, and identify possible breakpoints.

2.2. Core structure and process components for values affirmation

The implementation of the values affirmation intervention suggests the need for a specific set of behaviors on the part of the intervention's designers, the implementers themselves – typically classroom teachers, and the targets of the intervention – the students. To identify these core components, we conducted a review of prior values affirmation studies with an eye towards specific elements of the intervention that had been empirically linked to outcomes in prior works, or in cases where the empirical connections were less certain but consistent with data and strongly supported by theory.

The first core process component centers on *intervention design (fit)*. Values affirmations administered in classroom and laboratory settings have typically offered students or participants a list of

[1] For a detailed description of each of the five dimensions, see Dane and Schneider (1998; p. 45).

[2] While the five dimensions approach can be clearly tied back to the work of Dane and Schneider, the core components approach, once named and defined, can be observed in decades of prior work on fidelity and program integrity. Hall and Hord (1987) for example, emphasize the necessity of clearly delineated "building blocks of the innovation" (p. 117), while Bond, Evans, Salyers, Williams, & Kim (2000). write about "critical ingredients" (p. 76), and Mowbray et al. (2003) refer to "fidelity criteria".

[3] We do not directly replicate any of the conceptual strategies employed by previous 'structure and process' studies, though we recognize in our MWAP intervention the necessity of a teacher-centered delivery as a core structural component., Here, we focus on analysis of process components through measurement tools developed specifically for MWAP. A first pass is taken to examine the variation in how teacher-centered delivery impacts various process components, though a more rigorous analysis is planned for future research.

important 'values' from which they are instructed to select the most personally important. In order for self-affirmation to be triggered, this list needs to reflect the salient values of the particular student body at whom it is targeted. Additionally, the entirety of the exercise should also be written at a reading level and in a language that matches the literacy level of the students (e.g. Sherman et al., 2013). If the values do not resonate, the literacy level is too high, or they cannot read the language in which the intervention is delivered, students are unlikely to engage in the cognitive processing that inoculates them against stereotype threat. The *process* of tailoring the design of the intervention is by definition a response to a *structural* component that the researchers have no control over—such as the relative achievement level or diversity of language present in a particular school population. This design element represents a differentiation of the values affirmation intervention that is essential, rather than merely acceptable. Though it poses tremendous challenges for drawing conclusions across studies and for offering generalizable recommendations based on experimental findings, the issue of fit is of fundamental significance for the affirmation process and must therefore be a component of fidelity. The second and third core process components that we derive from studies of values affirmation focus on delivery of the intervention and the ways that teachers frame the intervention when presenting it to students.

Values affirmation works to reduce the impact of stereotype threat in part by reminding a student of the diverse, positive aspects of herself and in so doing reducing the stress or anxiety that she may experience as a result of threat in one domain (academics) based on one aspect of her social identity – or an intersection of those identities (Sherman et al., 2013). This mechanism by which values affirmation has its impact on students points to a possible breakpoint; the intervention's implementation must not induce *stress* in students. To avoid creating stress for students Cohen and his co-authors (Cohen et al., 2006, 2009) trained teachers to avoid referring to the writing exercise as an evaluative instrument, to emphasize that students' written work would remain private even from the teacher, and therefore constituted an ungraded opportunity to free-write about values.

We draw our third core process component from the growing consensus that the efficacy of a values affirmation depends in part on the *stealth* with which the intervention is presented and framed to students (Yeager & Walton, 2011). Of the implemented values affirmations that obtained positive effects, most were delivered in a stealthy fashion, disguised as regular classroom assignments. Stealth emerges as an essential feature of implementation not only from theory and norms among experimental social psychologists, but also recent experimental evidence supporting these intuitions. Informing students that a values affirmation should be beneficial and thereby compromising stealth, appears to attenuate affirmation's benefits (Sherman et al., 2009). Therefore, in order to achieve *stealth* a values affirmation should be delivered by a teacher in a classroom as a regular writing assignment and refrain from referring to the activity as beneficial to their academic performance. Teachers might also be cautioned to avoid referring to the activity as 'special', connected to 'research', or in any other way out of the ordinary.

During our review of prior studies, we identified a significant tension between maintaining low-stress and stealth, at least as Cohen et al. (2006, 2009) have recommended operationalizing them during implementation. A teacher introducing a classroom activity would rarely emphasize its ungraded nature and almost never promise not to read it. While these promises could reduce the *stress* associated with the values affirmation exercise, they would simultaneously highlight the intervention as out of the ordinary, compromising the *stealth* of the intervention, prompting students to ask probing questions about its purpose, and potentially reducing student engagement. Another recent, experimentally administered values affirmation in a middle school setting suggests that the guarantee of privacy may not be necessary in order to obtain effects. The effect of the intervention was larger for students assigned to the condition in which teachers read their essays (Bowen, Wegmann, & Webber, 2013). Based on these findings, we suggest that *stealth* must be balanced against efforts to minimize *stress*.[4] We use our data to assess the impact of breakages in the stress and stealth components in an effort to further clarify what acceptable adaptations might be made to attenuate the tension between these two components. A fourth core process component centers on the *timing* of the intervention. Stereotype threat interacts with achievement to stimulate a recursive, or cascading process that produces a downward performance trajectory; in short, threat and early academic failures reinforce one another (Cohen et al., 2006, 2009; Sherman et al., 2009; Obradovic, Burt, & Masten, 2010). Values affirmation can buffer students against these early failures, increasing resilience, and initiating a recursive process in the opposite direction. These insights suggest the first essential aspect of intervention timing: *values affirmations should be delivered to students near the beginning of the school year*, before a downward performance trajectory can be initiated. Again, at least one study of values affirmation validates this expectation (Cook, Purdie-Vaughns, Garcia, & Cohen, 2012).[5] Appropriate timing of the intervention should be qualitatively determined based on the particular context of a school, in part relating the fulfillment of this component to the fit, discussed above.

The fifth and final core process component necessary for achieving fidelity to the theoretical ideal of values affirmation rests on *the extent and nature of student engagement with the treatment*. While some fidelity scholars would submit that behaviors of the participants (alternately dubbed participant responsiveness or adherence) are outcomes rather than measures of implementation fidelity, in this intervention we view these intermediate outcomes as indicative of the extent to which students are in fact receiving the treatment. Fidelity of implementation, in our conceptualization, is a product of core components, but these components can be ordered and may interact with one another. The extent of student engagement and the nature of this engagement are certainly likely to be affected by core process components discussed above, but student engagement remains a crucial and distinct step that must also be taken in the process of completing a values affirmation. At baseline, in order to receive the treatment, a student must first receive the treatment exercise as intended by the researchers. This component may be assumed away in a small, laboratory setting, but should not be taken for granted during a scaled-up replication that is randomized at the student level and reliant on a teacher-centered delivery. Values-affirmation requires a particularly deep processing that cannot be triggered absent authentic student engagement. Recent work even suggests that the content of a students' written work mediates the effect of the intervention (Shnabel, Purdie-Vaughns, Cook, Garcia, & Cohen, 2013).

[4] We should also note that in order to ensure privacy and to maintain teacher blindness to treatment condition, Cohen et al. delivered exercises to students in envelopes with their names written on the outside. In many educational contexts, we believe that this mode of administration is likely to evoke standardized testing procedures and could simultaneously induce stress and compromise the stealth of the values affirmation. Middle school teachers administering the intervention in St. Paul, Minnesota found the envelopes to be distracting and recommended they be eliminated (Borman, 2012).

[5] The same study suggests that timing may matter more than dosage. Thus far, though easily measurable over the standard course of four administrations, dosage effects do not appear to be particularly significant mediators of values affirmation's impact.

In addition to these 5 core process components, we identify one core structural component: teacher-led implementation. In part, teacher implementation is necessary in order to maintain *low-stress* and *stealth*. Researchers or unknown proctors might compromise both of those conditions, and previously successful implementations of values-affirmations in school have made use of classroom teachers. However, we also suggest that teacher-led implementation be considered a core structural component of the intervention, as a practical measure. If the promise of values affirmation is to be realized in American schools, then teachers are most likely to be the individuals administering the exercises and potentially best situated to employ affirmation with large groups of students. As best practice in implementing educational interventions, intervention designers should consult with educators when tailoring the intervention, as teachers will have the best information about salient threats, literacy levels, and language abilities. Designers should also carefully plan training and support for teachers to ensure that delivery coincides with the primary mechanisms necessary to achieve successful implementation, should coordinate with teachers to achieve timely implementations, and should identify ways for teachers to encourage sincere student engagement.

Having identified these core elements, which are summarized in Table 1 and may be thought of as a gold standard against which any actual implementation is evaluated, we now turn to the focus of this analysis, the Madison Writing and Achievement Project (MWAP).

3. Intervention design and procedures to support, monitor, and measure fidelity

3.1. Madison Writing and Achievement Project (MWAP)

MWAP implemented a two-year, district-wide, randomized field trial of values affirmation. The project sought to replicate and scale up Cohen et al.'s intervention. The treatment exercise asked students to select two or three values from a list and then to write about when and why those values were important in their lives. Cohen et al. (2006, 2009) administered the writing exercise to two cohorts of seventh grade students across four classrooms and found a 40% reduction in the Black-White achievement gap among the treatment students. To our knowledge, this is the first district-wide replication of this values affirmation writing intervention.

The MWAP intervention took place in an urban school district in the Midwest that fosters some of the largest ethno-racial achievement gaps in the country – between Whites and Blacks (Vanneman, Hamilton, Anderson, & Rahman, 2009), as well as between Whites and Hispanics (Hemphill & Vanneman, 2011). The intervention was implemented over two cohorts of seventh grade students across eleven schools. The first year of implementation, with which this study is concerned involved 91 classrooms, 45 teachers, and 1049 students. Though the initial intervention on which we based this replication (Cohen et al., 2006, 2009) entailed treating students with the values-affirmation one and two times, subsequent implementations treated students between three and five times over the course of the academic year in an effort to reinforce the initial treatment (Cook et al., 2012). Students in the MWAP study were asked to complete a sequence of three or four writing exercises in class over the course of the school year.

The intervention processes, and their relationship to the core components of fidelity, are described below and illustrated in Fig. 1. The core structural component identified above, teacher led implementation, was a central consideration at all points during intervention design and constant fact during implementation. Fig. 1 also illustrates the manner in which the core process components necessary for maintaining fidelity were differentially important, depending on the stage of the implementation. For example, assessing the 'fit' between the values listed in the treatment exercise and the language skills and literacy levels of the students needed to occur early on, but then required only some attention and maintenance throughout the year, whereas maintaining stealth and minimizing stress to students were central aims each time an intervention was scheduled with and then administered by teachers.

3.2. Intervention design and fit

Consented students were randomized within school to either the treatment or control condition; non-consented students received a third, neutral written exercise in a similar format.

Table 1
Core Process Components and Gold Standard Elements of Implementation.

Core Process Component	Gold Standard for Implementation
Intervention Design (Fit)	☐ Reflects salient values of student body ☐ Appropriate reading level for student ☐ Language of delivery
Low-Stress	☐ Described as ungraded ☐ Promised privacy ☐ Not described as a test or associated with research ☐ Delivered by a known teacher ☐ Regular classroom setting ☐ Students ignorant of variation in exercises
Stealth	☐ Not described as "good for you" ☐ Not described as a test or associated with research ☐ Is described as ordinary. ☐ Delivered by a known teacher ☐ Classroom setting where writing would take place ☐ Teacher blind to student treatment condition ☐ Students ignorant of variation in exercises
Timing	☐ First, implemented early in the school year ☐ Later, prior and proximate to an exam or other stressful event
Student Engagement	☐ Students select values from list ☐ Write about values' personal importance

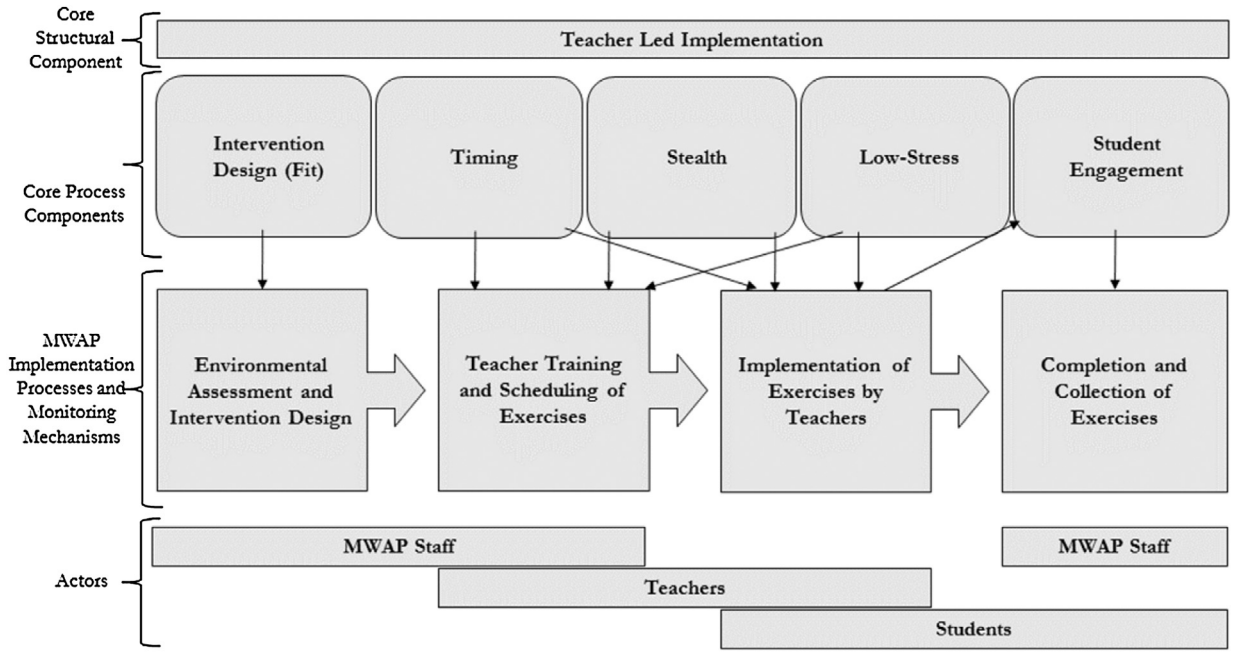

Fig. 1. Values Affirmation: Core Components and Implementation Processes.

The first two treatment exercises were structured response questions in which students were given a list of values and asked to select two to three values that were important to them. In keeping with Cohen et al.'s design, the third and fourth treatment exercises prompted students to respond to a question about a specific value. When possible, this value was one they had written about previously. Across all treatment exercises students were prompted to reflect upon their choice of value or the value provided, and instructed to write a response explaining why that value was important to them.

Students in the first control condition were provided with the same list of values as students in the treatment condition but asked to select two to three values that were not important to them. They were then prompted to reflect on why the values they selected may not be important to them and asked to write about why they might be important to someone else. Students in the second control condition (which will be referred to as the Neutral condition) were given an expository prompt that encouraged, for example, describing how they opened their locker or got ready for school. This neutral exercise was also provided to unconsented students.

As the treatment condition was randomized at the school level, students in *the same classroom* would typically be completing different exercises. Based on consultation with teachers, administrators, and student records, students were provided with Spanish language or low-literacy versions of each exercise as needed to ensure full participation and engagement from all students.

3.3. Low-stress and stealth

To facilitate an implementation that minimized stress and maintained stealth, program assistants, who liaised with each school, collaborated with school staff and faculty to identify the most appropriate classrooms for completing the writing exercise. Schools were asked to accommodate the implementation of the exercises in classes where a creative writing exercise would be a common type of assignment. Actual implementation classrooms varied between Language Arts, Homeroom, and World Language classrooms.

Program assistants also provided brief trainings to implementing teachers prior to the intervention delivery, and delivered written implementation instructions to teachers with each exercise. Training of teachers included introduction of a script to be used during exercise delivery implementation, familiarity with appropriate responses to common student questions, and an introduction to teacher implementation reports, herein referred to as Exercise Fidelity Reports. During pre-intervention sessions, teachers were asked to avoid telling their students that the exercises were "good for you", part of research, or beneficial in any way, and were encouraged to treat the exercises as a normal classroom activity. For example, if a teacher normally put an agenda on the board, we encouraged the teacher to include "free write" or "writing activity" on the board so as to maintain a sense of normalcy for the students. To further support teachers in this effort, a list of potential student questions coupled with suggested answers were provided for teachers to review ahead of time. To frame the exercise as part of a normal class activity (decreasing student stress and maintaining stealth), teachers were asked to allow 15–20 min to complete the exercises independently and then continue on with regularly planned class activities. These instructions were re-emphasized on printed instruction forms delivered to teachers with each subsequent exercise. Written instructions also re-emphasized the importance of privacy for students while completing the exercises. Included in exercise delivery packets were exercise fidelity reports to be filled out by implementing teachers for each intervention. Reinforcement of these constructs in instructions was an important element of maintaining a low-stress environment for students.

In addition to printed teacher instructions, each exercise was printed with an identically formatted cover sheet including a printed name and study ID number. The cover sheet was intended to buttress the students' sense of privacy. This is a slight departure from Cohen's original design in which student exercises were delivered to students in personalized envelopes, and an entire classroom's completed exercises were sealed in a larger envelope. The change in delivery method was made in response to feedback from teachers in a pilot study conducted by the MWAP group. In that pilot, teachers overwhelmingly reported that the envelopes

were cumbersome and conspicuous – triggering students to ask probing questions about the exercises. Additionally, by pre-packaging the exercises before delivery, the likelihood of breaking the "blindness" of teachers and students was minimized. Teachers were also asked to not look at completed exercises and to avoid bringing attention to any difference between exercise versions.

3.4. Timing

Timing protocols were based directly on the original intervention design established by Cohen et al. (2006, 2009); the first intervention should occur close to the beginning of the school year and the subsequent exercises should occur close to high stress testing, in this case the state and district standardized testing. Members of the research team worked with classroom teachers and school Learning Coordinators to schedule a drop-off date for each exercise implementation with the understanding that the exercises should be completed close to, and certainly *prior to* a standardized test. Teachers and school liaisons typically identified an ideal window of several days during which students should complete the exercises so as to accommodate different schedules.

3.5. Student Engagement

Though it is arguably the most important component, student engagement is the element of the intervention over which the research team has the least direct control and which is most likely to be influenced by teachers' manner of delivering the intervention. In order to trigger the buffering psychological processes, students must identify values that resonate in their lives, and write about them in a way that is affirming. However, all of the above efforts work to facilitate student engagement. The project team worked to ensure that the exercises were presented in an age-appropriate and skill-appropriate fashion, referenced relevant values, provided training to teachers, and emphasized the importance of student engagement (as described above). We discuss our approach to measuring engagement in more detail below.

4. Methods and materials for measuring fidelity

4.1. Exercise fidelity reports

In addition to training protocols described above, teachers were asked to complete exercise fidelity reports for each exercise administered. These reports were delivered with the exercise instructions for each intervention. Reports consisted of a one-page form that each teacher was asked to fill out immediately after each exercise implementation. The number of reports an individual teacher was asked to complete varied with the number of class periods in which there was an intervention assigned. Though this is a limited measurement, as the data are self-reported, these forms provide indication of the extent to which *stealth* and *low-stress* were achieved as well as providing insights into the extent of the teacher's compliance with the intervention guidelines. These forms serve as a useful indicator of breaks from the core components of stealth and low-stress, stemming from the actions of students and/or teachers.[6]

The fidelity reports asked teachers to record the number of students talking during the writing exercise, asking about research, noticing a difference between exercises, and asking for additional explanation for exercise instructions. The reports also asked teachers to make note of other classroom activities taking place that day, and included a space for other relevant comments. Teachers were asked to return the completed reports in the same packet as the completed exercises.

By collecting this information we are able to estimate how successful each teacher was at creating the conditions necessary for successful implementation as defined in our core components design, *stealth and low-stress,* as well as the extent to which both the teacher and the research team achieved fidelity for the first three core process components: *intervention design* (did students ask for additional explanation, a possible indication of poor reading level or language fit), *low-stress* (did students ask about research or ask if it was graded/a test), *stealth* (did students notice a difference in exercises or otherwise comment that it was out of the ordinary). These exercise fidelity reports comprise a chief data-gathering method to measure fidelity for *intervention design, low-stress,* and *stealth.*

We also use receipt of these forms to measure the extent of teacher compliance with the intervention protocols. This measure must be understood as a reflection only of the teacher's willingness to comply with the instructions received in pre-intervention training and in written form with the students' interventions. This measurement is relevant to our overall analysis because we believe teacher compliance is likely to influence student engagement. Teacher compliance is defined as *partial* if a teacher returns the fidelity forms for each exercise administered with the student reaction information completed, but without other classroom activities mentioned. Teacher compliance is defined as *complete* if both student reaction information and listed other classroom activities are provided. (The "other comments" section of the reports was described to teachers as an optional response section.) By requesting teachers to document events during implementation, the fidelity reports may also act as an implementation check, heightening the sense of accountability to conform to protocol, and increasing the probability of fidelity.

Conversely, weaknesses of relying on the fidelity reports are the potential threats to internal validity in the information reported and the inability to determine breaks in fidelity if teachers failed to complete the report. Further, internal validity could be threatened if teachers see the fidelity reports as a potential source of judgment by the researchers regarding their abilities. If teachers are resistant to implementing the intervention, perhaps in an effort to protect their classroom time, they may also resist by not completing, or partially completing, the fidelity report. Failing to complete fidelity reports may therefore occur at random, but non-return or non-completion may also be a form of systematic resistance to intrusion on regular classroom time. Even given these weaknesses, the completed classroom fidelity reports allow us to assess the implementation procedures in a general way and serve as a data source hereto unavailable in any other trial of this intervention. Further, recent work also suggests a high degree of accord between teacher self-reported measures of fidelity and those determined by an independent observer in the classroom setting (Foster 2011). Given the variation between classes in reporting disruptions, some reporting none at all and some reporting multiple disruptions, we are lead to believe that there is evidence of heterogeneity in implementation between classrooms in these reports that could directly affect overall implementation fidelity.

4.2. End-of-year survey

In addition to measuring the level of teacher compliance to exercise implementation, we also address breaking of the core component of *stealth* and *low-stress* using a self-reported, reflective, teacher survey given to participating teachers at an

[6] For example, a teacher can, despite instruction against it, introduce the activity as part of a research project. Students could also, despite teacher instruction, talk to their peers and compare exercise activities.

Table 2
Student Response Examples for Coding by Treatment Condition.

	Values Affirmation	Control	Neutral
FTI-T = 1	My friends and family are really important to me. When I need some one to talk I can go to them. I love music! I play music when I'm happy or sad. Music feels like a second world to me. And when there's no music I sing. My great grandma is 81 years old and she taught me to 'live life to the fulles'. Somethings never happen twice so I appreciate everyday of my even if it's a bad day.	i think this values that i pickt are very important to me because i like soccer and im good at it. and the thing about listining to helps me with my homework. and if some ever would care it would be *my parents because they like to see me succed* in my soccer game so when i grow up i can be a pro soccer player	During summer, I listened to music. I listened to all the types of music I have on my ipod, which is . . . Bachata, Hip-hop, some Rap (Nicki Minaj, Eminem, Lil' Wayne + Drake.) Some of my moms music, but I don't know what category they go in. Last time I listened to my music, was yesterday. Listening to music, helps me relax
FTI-T = 0	1) Thats Who i am–!2) It might be important to be religious if you work in a church.It might be important to be good at art if you like to do art or if you're an artest.	I sat and witchd TV i think it was the news. the presh speech, was wet watch wipeout. theo, football that i dident relly care about, then i went to bedThe End	In People program goverment would have been important to my friend. Becuase she was in a group about law and gurement. So I think it would be important to her to know about that stuff so she could do well in that group.

Underlined text indicates part of response in which student engages in an affirmation. Note that it was possible for students to affirm values even in conditions where it was not explicitly encouraged.

end of the year debriefing. The surveys were administered to teachers during the start of the meeting, and then followed up by group debrief sessions with project. The survey asks teachers to select from a list of descriptors they used to present the exercises and to explain in general what student reactions to the exercises were. Teachers were also asked how they believed students viewed the exercises, as out of the ordinary or as part of the normal classroom activities. Last, teachers were asked to report what they thought the purpose of the MWAP study was in an open response. Our indicators of teacher induced fidelity breakages in these surveys are: (1) Teacher reported describing the activity as research or "good for you" (2) Students thought the exercises was out of the ordinary (3) Teachers report knowing the correct hypothesis. Though debriefing sessions were requested of all participating teachers, a little less than half (20/45) chose to participate. This hampers our ability to rely on the survey data as anything more than an indication of how teachers who chose not to participate in the debrief *might* have been implementing and perceiving the exercises. However, we find that the information highlighted in these reports helpful in providing a supplemental description of implementation fidelity.

4.3. District reports and exercise receipt files

In order to evaluate the success of timing the interventions close to, and prior to, stress-inducing events, we recorded the date (or date range in the case of a school implementing their exercises over the course of several days) of each implementation and compare that to the date window of the next scheduled standardized test for those students.

4.4. Exercise coding for student engagement

A potential fidelity break at the student level is whether or not students were deeply engaged with the writing activity. A coding scheme developed by the MWAP research team enabled content analysis of all exercises. Two variables were identified as indicative of basic student engagement with the exercise: whether or not the student *identified at least one value* and whether or not the student *explained why that value is important to him or her*. Together, these two items comprise *Fidelity to Instruction* for students assigned to the treatment condition (hereafter referred to as *FTI–T*), An exercise must have been coded a "1" for both variables in order to be a "1" for *FTI-T*. Thus, for example, if an exercise were coded a "1" for *Identifies a Value* but a "0" for *Explains Why Value is Important to Self*, that exercise would be coded a "0" for *FTI-T*. A perfect implementation would result in all of the treatment students coded as a "1" for *FTI-T* for each exercise, while all of the control and neutral students would be coded as a "0" for *FTI-T* for each exercise, though our content analysis finds that some students not in the treatment condition satisfied these conditions.

Another measure of student engagement is the extent to which the students engaged in self-affirmation, since that is the psychological process the exercise is designed to promote. For an exercise to be coded a "1" for *self-affirmation*, the exercise must have identified a value from the list provided and either been coded a "1" for *values affirmation* or *attribute affirmation*. Self-affirmation is considered to have occurred even if the student engaged in attribute affirmation or values affirmation. Students who engaged in attribute affirmation and not values affirmation may have identified a value, but rather than explaining why the value is important to her or him, they may explain how they engage in the value, such as "being good at sports."[7]; Theoretically, the effect of a lack of student engagement on treatment impacts would be manifest in the degree to which treatment students complete the treatment exercise as instructed and the control/neutral students do *not* engage in the same process of self-affirmation that the treatment condition is designed to prompt. That analysis can be accomplished with the aforementioned binary variable of *self-affirmation*. Examples of student responses meeting these different conditions are presented in Table 2.

5. Results

5.1. Teacher compliance, stealth, & low stress

We begin with measures of teacher compliance as these are an overall indicator of: (1) Our ability to measure the potential threats to fidelity of implementation and (2) As an indicator of teacher investment in delivering the exercises as intended. Teacher compliance was operationalized as the partial or full completion and return of classroom fidelity reports. Partial completion included only information about student behaviors (talking, questioning, etc.) during implementation. Teachers were completely compliant within each exercise implementation if they returned the fidelity report with all sections of the report filled out, with the exception of the optional "other comments" field.

Table 3 portrays teacher return of exercise fidelity reports and compliance across the four implementations of the values affirmation treatment. More than 85% of teachers returned fidelity

[7] The vast majority of exercises that met the criteria for self-affirmation did so as a result of having engaged in values affirmation, which is unsurprising given the instructions of the prompt.

Table 3
Teacher Return and Compliance with Fidelity Reports.

	Exercise 1	Exercise 2	Exercise 3	Exercise 4	Total
Partial% (n)	88.2% (75)	85.9% (73)	95.0% (38)	87.1% (74)	88.1% (260)
Complete% (n)	68.2% (58)	63.5% (54)	80.0% (32)	60.0% (51)	66.1% (195)
Implementations% (N)	85	85	40	85	295
Schools (N)	11	11	7	11	11

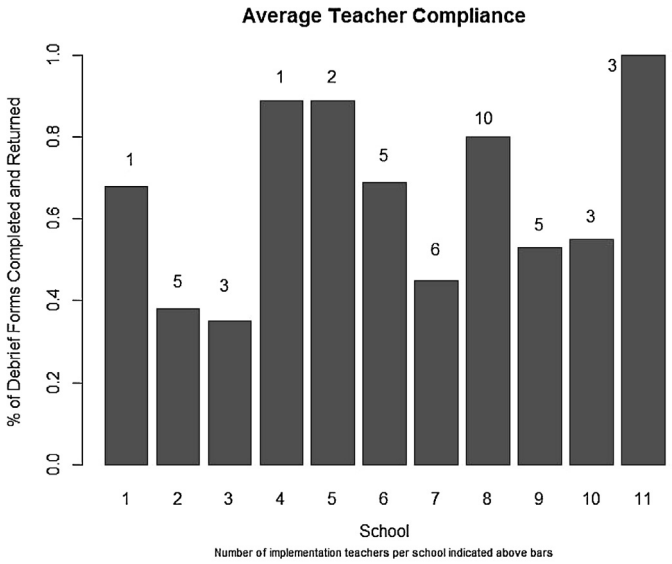

Fig. 2. Average Teacher Compliance by School over All Implementations.

reports that were at least partially completed. However, only 6% of teachers returned complete forms during the first exercise, and with the exception of the compliance rate during exercise 3, the percentage declined steadily to reach 60% by exercise 4. We find as we disaggregate return rates by school, there are several schools driving compliance rates downwards. Schools 2, 4, 7, and 9 did not complete the third exercise implementation (Fig. 2). School 4 had relatively high compliance rates over 80% overall, however this school had only four classroom implementations, all supervised by the same teacher, a lesser responsibility than most of the teachers in our sample. The exclusion of these schools in the third exercise eliminated three of the four lowest compliance schools. Clearly, there is heterogeneity in the rates of compliance between schools. Further, within schools, teachers who were not compliant in one exercise were also likely to be non-compliant in subsequent exercise administrations.

The responses taken from the returned exercise fidelity reports were used to measure *low stress* and *stealth*. Specifically, teachers who report students asking about research and/or noticing differences between assignments are considered to have student-induced breakage. If teachers report on the exercise fidelity reports that students perceive the exercises to be outside of normal classroom activities, part of research, or notice that their classmates have different exercises from each other we consider it an indication that there has been a potential break to the intervention. We also operationalize classroom disruption as classrooms in which students were reportedly talking during the exercise. Because students complete exercises in a group setting, if a teacher reports at least one student asking about research or differences we consider the entire classroom exposed to the same possible fidelity breakage.

Despite some fatigue in complete teacher compliance with fidelity reports over time, important feedback on classroom context with respect to potential breaks in *low-stress* and *stealth* may be obtained from the partially completed forms. Additional feedback on these two core components comes from the end of year teacher survey. Tables 4 and 5 report these results.

There is clear evidence that *stealth* may have been compromised. Of the 45 classes that participated during the year, 18 were language arts, 24 were homeroom, and 3 were world-language classes, all of which were settings in which teachers and principals expected the values affirmation writing exercises to blend with regular coursework. Nonetheless, on the end-of-year survey, sixty percent of responding teachers stated that their students did not view the exercises as a normal classroom activity. On that same survey fifty percent of responding teachers recalled describing the exercise as research, 30% recalled describing it as 'good for you', and 20% recalled describing it to students as 'something we have to do', though each of these had been cautioned against during small group trainings and in the instructions provided with each round of the intervention. Additionally, 21% of exercise fidelity reports indicated that students noticed differences between treatment and control exercises. This points to not only a break in *stealth* (because differences in exercises could signal to students that this is not normal), but also in *intervention design*, given that efforts were made by the researchers to ensure treatment and control exercises looked similar. Finally, nearly sixteen percent of exercise fidelity reports indicated that at least one student asked if the exercise was part of research, and over half of those reports indicated that *more than one* student asked about research. No teachers reported connecting the intervention to stereotype threat, suggesting that teacher blindness to the research hypothesis was maintained. Cohen et al. have indicated that it is imperative that teachers and students should remain unaware of the actual hypothesis underlying the research design.

There is less evidence that fidelity breaks occurred with respect to the *low stress* component. On the end-of-year survey, only one

Table 4
Teacher Reports from Debrief Forms.

Students . . .	Exercise 1	Exercise 2	Exercise 3	Exercise 4	Total
Asked for explanation	69.7% (52)	56.2% (41)	34.2% (13)	56.8% (42)	56.9% (148)
Talked during exercise	58.7% (44)	65.8% (48)	52.6% (20)	59.5% (44)	60.0% (156)
Noted differences among exercises	25.3% (19)	32.9% (24)	5.3% (2)	16.2% (12)	21.5% (57)
Asked about research (at least 1)	12.0% (9)	17.8% (13)	15.8% (6)	17.6% (13)	15.8% (41)
Returned (N)	75	73	38	74	260
School (N)	11	11	7	11	11

Table 5
Teacher Reports from End of Year Survey.

Teacher Described Exercise As . . .		Teachers Described Students As.	
Creative writing	55% (11)	Believing the exercise normal	60% (12)
Research	50% (10)	Appearing comfortable	60% (12)
Important activity	40% (8)	Exhibiting concentration	50% (10)
Good for you	30% (6)	Demonstrating boredom	40% (8)
Private journaling	25% (5)	Showing curiosity	35% (7)
Something have to do	20% (4)	Seeming confused	35% (7)
Writing test	5% (1)	Appearing anxious	5% (1)
Teachers Surveyed	20		20

teacher indicated that they described the exercise as a test, and only once teacher described students' affect while taking the exercises as anxious. While 60% of exercise fidelity reports indicated that students talked during the exercises, and this could be seen as inducing stress via class disruptions, it could also be argued that middle school students talking during any given classroom activity is not abnormal and thus not necessarily indicative of stress. Also, while 57% of the reports indicated that students asked for additional explanation when completing the exercises, qualitative feedback from teachers indicated that students largely required clarification with respect to the control exercise as opposed to the treatment. The proportion of students who asked for explanations or clarifications also diminished notably from almost 70% during the first round of interventions to 56.8% during the last round.

5.2. Timing

Exercises were timed close to the start of school and prior to the district's first round of standardized MAP tests for most schools, though proximity to the start of the school year varied, as the district staggered testing times to ensure adequate bandwidth. Eight of the eleven schools administered the values-affirmation within the first two weeks of school, ten before the first MAP tests. Time between administration of the intervention and the MAP assessment varied from one day to one week. One school was a notable outlier on exercise 1, implementing its first exercise well after the first standardized test due to delays with the consent process (Fig. 3). Exercise two followed closely, within six weeks after exercise 1 and was implemented shortly before the Wisconsin Knowledge and Concepts Examinations (WKCE). Exercise 3, which was optional, took place during January or February, depending on the schools' writing assessment schedule. Exercise 4 was timed in the late spring to precede the last MAP assessment of the year.

5.3. Student engagement

When assessing the proportions of students who follow instructions and/or engage in self-affirmation, the denominator could be the number of students who actually received, and thus actually had the opportunity to complete the exercise or the number of students who were supposed to have received the exercise. The former is the focus of our discussion of fidelity, while the latter is acknowledge as relevant for main impacts analysis. Table 6 reports these numbers. About 92% of students returned their assigned exercise. Less than one percent returned the wrong form.

We next assess the extent to which students met the criteria for self-affirmation. Because all exercises, regardless of condition, were coded for the same variables, comparing rates of affirmation across treatment conditions serves as a validity check to ensure that not only were treatment students engaging in values affirmation, but also that control and neutral students were *not* engaging in values affirmation at similar rates. If it were found that a substantial number of control and neutral students were engaging in self-affirmation even though the exercise assigned to them was not designed to prompt such a response, it would suggest that findings in the main impact analysis would be conservative, given that a large portion of the comparison group would have engaged in self-affirming behavior similar to the treatment students.

Self-Affirmation. The proportion of students who engaged in self-affirmation is reported in Fig. 4 (by actual receipt of exercise) and Fig. 5 (by condition assignment). Students who actually received the treatment exercise self-affirmed at a high rate across all four exercises, although there was a notable decline between

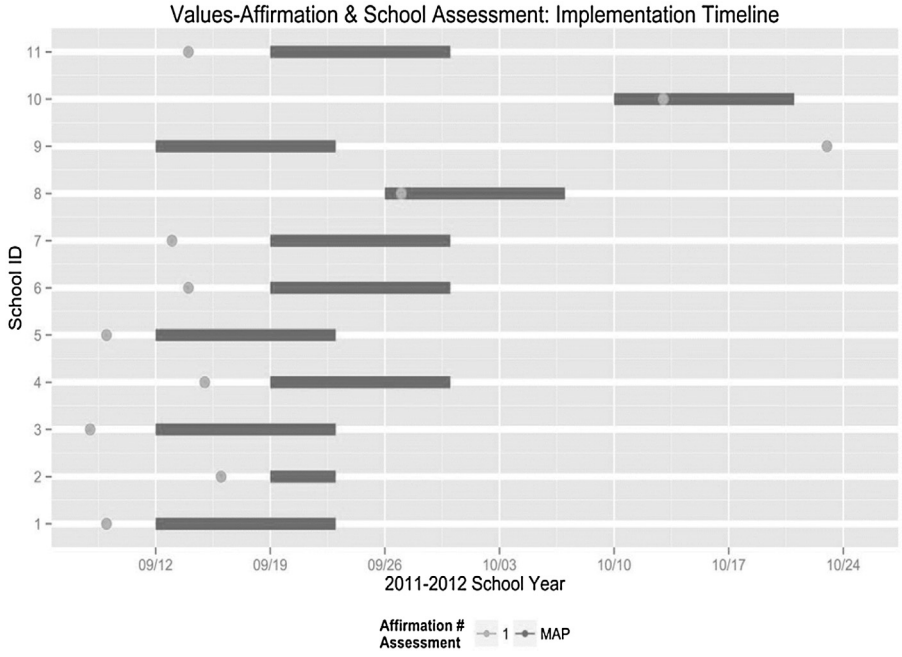

Fig. 3. Proximity to Start of the School Year and First Standardized Tests.

Table 6
Student Receipt and Return of Exercises.

Student Returned . . .	Exercise 1	Exercise 2	Exercise 3	Exercise 4	Total
Assigned Exercise	95.0% (995)	92.0% (963)	91.6% (636)	88.6% (928)	91.8% (3522)
Different Exercise	0.6% (6)	0.8% (8)	0.7% (5)	1.1% (11)	0.7% (30)
No Exercise	4.4% (46)	7.3% (76)	7.6% (53)	10.3% (108)	7.4% (283)
Total Exercises	1047	1047	694	1047	3835

the first exercise, when 88% of students affirmed, and the last exercise when 71% of students in the treatment condition affirmed. On average, the control students only engage in self-affirmation five percent of the time, peaking at ten percent in the first two exercises and dropping to less than five percent in exercises three and four.

Because of the nature of the neutral prompt during the first exercise, it is particularly important to check for possible values affirmation amongst those students. During the first administration of the intervention 54% (31 of 57) of those students met the criteria for *self-affirmation*. In other words, more than half of the students who received the neutral exercise were coded as having engaged in the same self-affirmation process as the students who received the treatment exercise. Given that half of the exercise 1 neutral students were randomized into the comparison group for impact analysis, this point may be relevant if future analyses reveal more insight into whether or not our measure of self-affirmation is a valid proxy for successful treatment or if dosage (the number of exercises) has any impact on results. For the purposes of evaluating fidelity of implementation, we have no indication that variation in how teachers delivered the exercises should play a role in explaining why or why not the students in the neutral condition engaged in self-affirmation, especially given the simplicity of the prompt in the neutral condition.

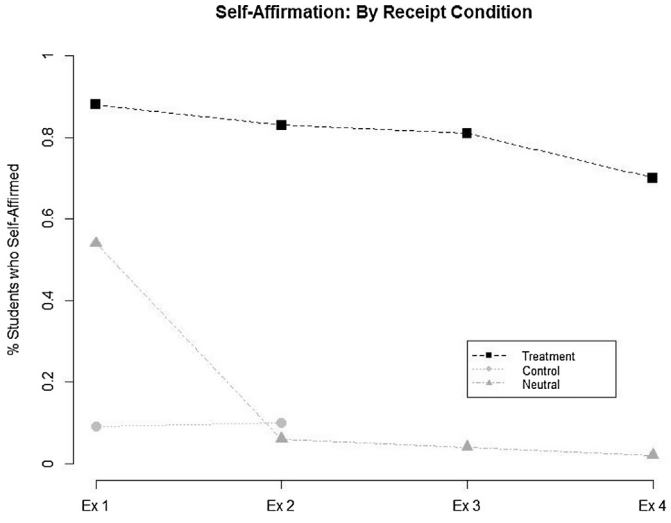

Fig. 4. Rates of Self-Affirmation by Receipt Condition.

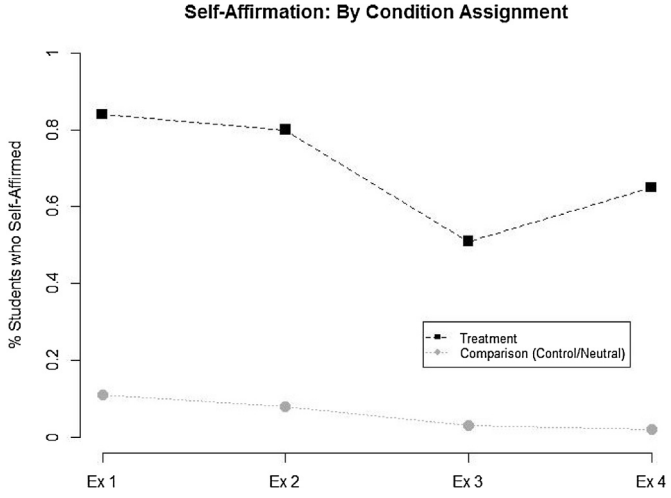

Fig. 5. Rates of Self-Affirmation by Condition Assignment.

Across Schools. A logical concern is the potential for variation across schools. Fig. 6 displays student self-affirmation amongst treatment students by school. Increased variation over time is readily apparent. In exercise 1, self-affirmation amongst treatment students ranged from a low of 81% in School 2 to a high of 92% in School 5 (mean = 87.6%). By exercise 2, this range had expanded, with a low of 73% in School 2 and a high of 88% in School 4 (mean = 82.3%). By exercise 4, only 48% of treatment students self-affirmed in school 1, while 87.5% did so in school 4 (mean = 70%). Future analysis that incorporates variation in the aforementioned breaks in *low stress* and *stealth* may be used to investigate the extent to which those core components impact variation in *self-affirmation* across schools.

Mapping Fidelity to Instruction on to Self-Affirmation. With respect to student exercises that meet the criteria for following treatment prompt instructions (identifying a value and explaining why it is important to oneself) and self-affirming, students may fall into one of three categories: (1) both *following instructions* and *affirming,* (2) *affirming* only (doing so despite not following instructions), and (3) neither *following instructions* nor *affirming.*

Fig. 7 reports the proportions of treatment students that fall into each of the 3 categories. For example, in exercise 1, 59% of students followed instructions and self-affirmed. An additional 35% self-affirmed without having followed the treatment instructions. About 6% of students did neither. Over time there is a steady decline in the total proportion of students who self-affirm, and this

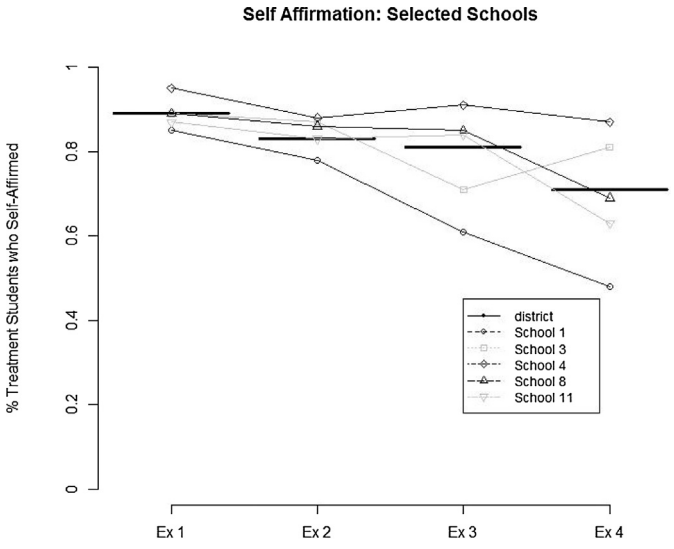

Fig. 6. Rates of Self-Affirmation among Treated Students in Selected Schools.

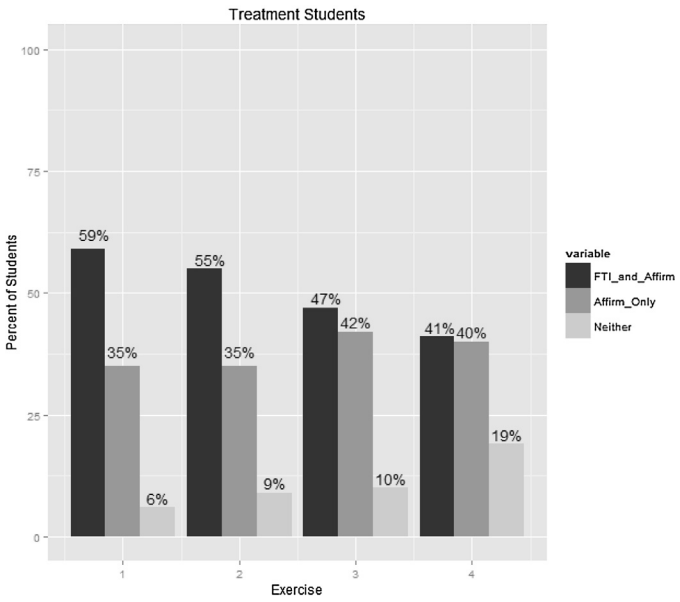

Fig. 7. Fidelity to Instruction and Self-Affirmation by Exercise.

coincides with a decline in the proportion of affirming students that also follow instructions. In exercise 1, 63% of students who self-affirmed followed instructions, compared with 50% in exercise 4. These results provide actual evidence that following treatment prompt instructions is a sufficient but not necessary pathway to self-affirmation.

6. Discussion

Using a core components approach (Waltz et al., 1993) and drawing on Cohen et al. (2006) original intervention design and key intervention components, we analyzed fidelity of the MWAP implementation to the theoretical ideal of a classroom-administered values affirmation. From this we identify five core components of *intervention design (fit), low-stress, stealth, timing, and student engagement* and use them as a framework for analysis and interpretation of our results. Our results point to the fact that while some components may stand alone in the implementation process (e.g. timing), other components are actually closely related and in turn can impact one another. For example, fidelity to the ideal of low-stress must be balanced against stealth (normalized activity), creating a natural tension between the two components in which high fidelity in one component may induce low fidelity in another component. Below, we discuss how these core components may or may not be dependent on one another, where the balance between components can create issues of fidelity, and how successful the MWAP project was in maintaining fidelity to each of the core components. Further, we identify how the fidelity of each component is reliant on the actions of differing actors. We include in the discussion future plans for analysis and implications for a larger scale-up of this particular writing intervention.

6.1. Fidelity to intervention design

Given the limited incidences of students needing additional explanation, language mismatch in exercises, and high rates of student engagement with respect to self-affirmation, we feel confident that the *intervention design* was met with high fidelity, particularly considering the magnitude of the scale-up of this intervention. The intervention design was developed and implemented by the research team. Interaction with teachers in how they carry out the research design separates the measurement of fidelity to *intervention design* from measurements of *stealth* and *low-stress*.

6.2. Fidelity to timing

The execution of the *timing* of the exercises, with the exception of one school, was true to the original protocol design. Small variation in timing existed, but given the number of schools participating can be thought of as a natural results of the size of the scale up. Rather than the research team mandate a specific date for each exercise, collaboration with individual schools enabled the timing of exercises to be close to the testing dates for each school. We would suggest that allowing the timing to vary slightly is not only acceptable; it is necessary to the success of the intervention design at scale. While the *intervention design* and *timing* components met high standards of fidelity, potential "breaks" related to *low-stress, stealth,* and *student engagement* do exist.

6.3. Fidelity to low stress and stealth, via a teacher-Centered delivery

Teacher training materials and protocols were established and implemented by the research team to ensure the highest possible fidelity to the components of *low-stress* and *stealth*. The packaging of exercises was modified from the original design to further normalize the appearance of exercises. In teacher training sessions, teachers were instructed to never refer to the exercises as tests or research. Upon delivery of the exercises instructions given in training were reinforced through additional written instructions and the classroom exercise fidelity report, to be completed by teachers. Despite these measures, our analysis revealed variation in teacher compliance to the established protocols, including how teachers introduced the exercises and teacher control of the classroom environment. Teachers reported having difficulty introducing the exercises without relating to students that they could be beneficial to them, or "good for them". Well over half of teachers reported that the exercises seemed "out of the normal" to their students. Despite the efforts of the researchers to modify the research design to lessen the non-normal appearance of the exercises the tension between the components of *stealth* and *low-stress* was not easily attenuated. From these findings we conclude that despite a structured and standardized protocol teachers often found their own method of integrating the exercises into their instructional activities. These findings resonate with Elmore's (1979) discussion of how street level bureaucrats develop coping strategies to implement new policy measures while balancing previous burdens. This suggests that future implementers may need to either find a way to exert greater control of the implementation of the intervention while keeping teachers engaged or find a way to integrate the creative strategies used by teachers to normalize the activities to the classroom. Teacher input with respect to integrating the exercises more seamlessly into the curriculum could not only reduce the likelihood of a break in *stealth* but also lessen the frequency of questions from students that put teachers in a position to justify the exercises.

Our measurements of the extent of threat to *low-stress* delivery and *stealth* components are limited by the realities of scaling up a research project into the everyday classrooms of teachers; particularly varied teacher compliance rates and weak participation in end of the year teacher debriefings. However, even given these breakages in the core components we find significant positive impacts of the intervention on the intended audience (Borman et al., 2016). These significant positive impacts, despite low fidelity of some core components, suggest that there may be aspects of *low-stress* and *stealth* that are not as essential to the

Table 7
Teacher Reports from End of Year Survey: Differences by Student Stereotype Threat Status.

	Potentially-threatened students whose teacher described exercises as . . .	Non-threatened students whose teacher described exercises as . . .
Creative writing	52.8%	68%
Research	49.7%	41.8%
Important activity	30.0%	32.2%
Good for you	35.2%	27.7%
Private journaling	34.2%	26.5%
Something we have to do	18.1%	14.3%
Writing test	8.8%	14.3%
N	193	313

success of the intervention as they are defined and operationalized in the original intervention protocol.

6.4. Student and teacher engagement

We have shown that overall students and teachers exhibit fatigue as the number of interventions increases. We maintain that the drop in FTI-T and compliance, or "leveling out" effect, is an expected trend given the number of times students and teachers are asked to engage in the intervention, or intervention fatigue. Our interpretation of the importance of the teacher's role in delivery speaks to other recent findings that suggest integration of teachers could bolster the effect of the intervention rather than hinder it (Bowen et al., 2013). Engaging teachers in the process of implementation, by adding the exercises to curricular modules or otherwise, may lead to a condition where students would be more likely to self-affirm. Further, if this is the case it implies that the exclusion of teachers from modifying and integrating the intervention in their own way could have served to reduce the magnitude of effects in this intervention and possibly other replications. The trends of fatigue, or leveling-out, we see in this replication could have disproportionately impacted students who are particularly vulnerable to stereotype threat, given that these "potentially-threatened" students are the target audience for these interventions,[8] if they were disproportionately served by non-compliant teachers (Cohen et al., 2006, 2009). Increasing the chance of self-affirmation for minority students could potentially increase the magnitude of desired positive effects overall. As a first look at whether or not potentially-threatened students were disproportionately served by teachers who break protocol when delivering the exercises, we delineate the results of the end-of-year survey (reported first in Table 5) by student stereotype threat status in Table 7. Potentially-threatened students appear to have been served disproportionately by teachers who did not follow some of the implementation protocols (describing the exercises as "research" or "good for you"). Future analysis of data that specifically investigates whether or not breaks in these particular core components has an affect on student outcomes will help inform whether or not the differences witnessed here are substantively meaningful. In either case, documenting the extent to which implementers of interventions follow protocol *and* investigating whether or not targeted subgroups are disproportionately exposed to those breaks is an important foundational step in evaluating fidelity of implementation.

[8] We define "potentially threatened" as minority students attending schools in which there are more white students than minority students. In this scenario we assume that there is more potential for a student to feel the effects of stereotype threat, thus making the intervention potentially impactful.

7. Conclusion

Our analysis of fidelity has provided a road map for further analysis and informs our recommendations for future intervention designs. First, variation exists in student FTI-T and self-affirmation regardless of the condition the student was assigned. There existed some treatment students who never met FTI-T in any of the exercises. Impact analysis using student self-affirmation results in place of assignment condition (Treatment vs. Control vs. Neutral) may provide insight into the validity of our self-affirmation measure and whether or not the self-affirmation coding scheme is a sufficient measure of the fidelity of implementation for student engagement. Further, is our operationalization of FTI-T also a measure of the values affirmation process that is believed to be taking place? Results of such analysis would help inform discussion as to whether exposure to the treatment (student receiving the treatment exercise) is necessary and sufficient, or if the benefits of values affirmation are only manifested through the actual writing process, which would be revealed in the self-affirmation measure, regardless of the student actually being exposed to the treatment.

Secondly, we believe that we have sufficient data to develop a more complex understanding of how the structural component of teacher-centered delivery, exhibited in stealth and low-stress core components, may serve to increase or dampen the effects of the intervention. From our analysis we have been able to identify those components that exhibit relationships to one another as well as which components are potentially less critical to successful intervention implementation. From this analysis we will be able to further operationalize the structural component of teacher-centered delivery, exhibited in the core components of low-stress and stealth, in particular and investigate the connections between these components and student outcomes.

Though preliminary, we believe our data provides evidence that further integration of teachers into the implementation process could be beneficial to students, particularly those students potentially experiencing stereotype threat. Of course this has implications for how these exercises may be used by a school district as a targeted intervention to close achievement gaps. Based on our findings we support the concept of scaling out the interventions beyond a research capacity, but we caution that the integration of these written interventions must be handled carefully. In Bowen et al.'s (2013) article the researchers incorporate teachers into the process of implementation by telling students their teachers would read the responses and find much larger impacts than among the comparison group. However, the extent of this integration of teachers into the implementation process potentially changes the mechanism of the intervention from intrinsic affirmation to extrinsically motivated affirmation. Further, incorporating the teacher in a potentially evaluative manner, simply by having it explicitly known that teachers would read responses, negates the low-stress emphasis of delivery in

favor of creating complete normalcy of the exercises. We suggest that to attenuate the tension between normalized activity and minimized stress, exercises could potentially be integrated into normal language arts curriculum with the expectation that only the student will be aware of what was written but will be evaluated on completion, similar to private journaling exercises. In this way, motivation to internalize values affirmation remains with the student while the delivery remains with the teacher and is accepted as a normal classroom routine.

References

Berman, P., & McLaughlin, M. W. (1976). Implementation of educational innovation. *Educational Forum*.

Berman, P. (1978). The study of macro- and micro- implementation. *Public Policy*, *26* (2), 157–184.

Blase, K., & Fixsen, D. (2013). Core intervention components: identifying and operationalizing what makes programs work: ASPE research brief. *US Department of Health and Human Services*.

Bond, G. R., Evans, L., Salyers, M. P., Williams, J., & Kim, H. K. (2000). Measurement of fidelity in psychiatric rehabilitation. *Mental Health Services Research*, *2*, 75–87.

Borman, G. (2012). *Examination of a self-affirmation intervention in St. paul public schools*, *Vol. IX*, St. Paul, MN: The Council of Great City Schools44 [Retrieved from http://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/85/St_Paul_Self-Affirmation.pdf.

Borman, Grigg, & Hanselman (2016). An effort to close achievement gaps at scale through self-affirmation. *Educational Evaluation and Policy Analysis*, *31*(1), 21–42.

Bowen, N. K., Wegmann, K. M., & Webber, K. C. (2013). Enhancing a brief writing intervention to combat stereotype threat among middle-school students. *Journal of Educational Psychology*, *105*(2), 427–435. http://dx.doi.org/10.1037/a0031177.

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. http://dx.doi.org/10.1177/1098214010366173.

Chen, H.-T., & Rossi, P. H. (1983). Evaluating with sense the theory-Driven approach. *Evaluation Review*, *7*(3), 283–302. http://dx.doi.org/10.1177/0193841X8300700301.

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: a social-psychological intervention. *Science*, *313*(5791), 1307–1310. http://dx.doi.org/10.1126/science.1128317.

Cohen, G. L., Garcia, J., Purdie-Vaughns, V., Apfel, N., & Brzustoski, P. (2009). Recursive processes in self-affirmation: intervening to close the minority achievement gap. *Science*, *324*(5925), 400–403. http://dx.doi.org/10.1126/science.1170769.

Cook, T. J., & Poole, W. K. (1982). Treatment implementation and statistical power a research note. *Evaluation Review*, *6*(3), 425–430. http://dx.doi.org/10.1177/0193841X8200600305.

Cook, J. E., Purdie-Vaughns, V., Garcia, J., & Cohen, G. L. (2012). Chronic threat and contingent belonging: protective benefits of values affirmation on identity development. *Journal of Personality and Social Psychology*, *102*(3), 479–496. http://dx.doi.org/10.1037/a0026312.

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.

Dee, T. S. (2014). Social identity and achievement gaps: evidence from an affirmation intervention. *Journal of Research on Educational Effectiveness*1–20. http://dx.doi.org/10.1080/19345747.2014.906009.

Dumas, J. E., Lynch, A. M., Laughlin, J. E., Phillips Smith, E., & Prinz, R. J. (2001). Promoting intervention fidelity: conceptual issues, methods, and preliminary results from the EARLY ALLIANCE prevention trial. *American Journal of Preventive Medicine*, *20*(Suppl. 1), 38–47. http://dx.doi.org/10.1016/S0749-3797(00)00272-5.

Durlack, 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*(3-4), 327–350. http://dx.doi.org/10.1007/s10464-008-9165-0.

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. *Health Education Research*, *18*(2), 237–256. http://dx.doi.org/10.1093/her/18.2.237.

Elmore, R. F. (1979). Backward mapping: implementation research and policy decisions. *Political Science Quarterly*, *94*(4), 601. http://dx.doi.org/10.2307/2149628.

Esbensen, F.-A., Matsuda, K. N., Taylor, T. J., & Peterson, D. (2011). Multimethod strategy for assessing program fidelity: the national evaluation of the revised G.R.E.A.T. Program. *Evaluation Review*, *35*(1), 14–39. http://dx.doi.org/10.1177/0193841X10388126.

Fagan, A. A., Hanson, K., Hawkins, J. D., & Arthur, M. W. (2008). Bridging science to practice: achieving prevention program implementation fidelity in the community youth development study. *American Journal of Community Psychology*, *41*(3–4), 235–249. http://dx.doi.org/10.1007/s10464-008-9176-x.

Fixen, D. L., Blase, K. A., Naoom, S. F., & Wallace, F. ([212_TD$DIFF]2009). Core implementation components. *Research on Social Work and Practice*, *19*(5), 531–540.

Foster, L. H. (2011). *Fidelity: snapshots of implementation of a curricular intervention (Ph.D.)*. United States, Virginia: University of Virginia Retrieved from http://search.proquest.com.go.libproxy.wfubmc.edu/docview/908430898/abstract?accountid=14868.

Fullan, M., & Pomfret, A. (1977). Research on curriculum and instruction implementation. *Review of Educational Research*, *47*(2), 335–397. http://dx.doi.org/10.3102/00346543047002335.

Gearing, R. E., El-Bassel, N., Ghesquiere, A., Baldwin, S., Gillies, J., & Ngeow, E. (2011). Major ingredients of fidelity: a review and scientific guide to improving quality of intervention research implementation. *Clinical Psychology Review*, *31*(1), 79–88. http://dx.doi.org/10.1016/j.cpr.2010.09.007.

Hall, G. E., & Hord, S. M. (1987). *Change in schools: facilitating the process*. SUNY Press.

Hall, G. E., & Loucks, S. F. (1977). A developmental model for determining whether the treatment is actually implemented. *American Educational Research Journal* 263–276.

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. http://dx.doi.org/10.1177/001440291307900204.

Hemphill, F. C., & Vanneman, A. (2011). Achievement gaps: how hispanic and white students in public schools perform in mathematics and reading on the national assessment of educational progress. *Statistical analysis report. NCES 2011-459*. National Center for Education Statistics Retrieved from http://eric.ed.gov/?id=ED520960.

Huntley, M. A. (2005). Operationalizing the concept of fidelity of implementation for NSF-funded mathematics curricula. *Proceedings of the 2005 national science foundation K-12 mathematics, science, and technology curriculum developers conference*38–43 Retrieved from http://www.powershow.com/view/20cdd-MzRkO/Operationalizing_the_Concept_of_Fidelity_of_Implementation_for_NSFFunded_Mathematics_Curricula_powerpoint_ppt_presentation.

Kost-Smith, L. E., Pollock, S. J., Finkelstein, N. D., Cohen, G. L., Ito, T. A., & Miyake, A. (2012). Replicating a self-affirmation intervention to address gender differences: successes and challenges. *AIP Conference Proceedings*, *1413*(1), 231–234. http://dx.doi.org/10.1063/1.3680037.

Lakin, J. M., & Shannon, D. M. (2015). The role of treatment acceptability, effectiveness, and understanding in treatment fidelity: predicting implementation variation in a middle school science program. *Studies in Educational Evaluation*, *47*(December), 28–37 Retrieved from http://www.sciencedirect.com/science/article/pii/S0191491X15000553.

Lastica, J., & O'Donnell, C. (2007). Considering the role of fidelity of implementation in science education research: fidelity as teacher and student adherence to structure. *Annual meeting of the american educational research association* Retrieved from http://www.gwu.edu/~scale-up/documents/Adherence%20to%20Structure_NARST_2007.pdf.

Lipsey, M. W. (2009). The primary factors that characterize effective interventions with juvenile offenders: a meta-analytic overview. *Victims & Offenders*, *4*(2), 124–147. http://dx.doi.org/10.1080/15564880802612573.

Lipsky, M. (1971). Street level bureaucracy and the analysis of urban reform. *Urban Affairs Quarterly*, *6*(June), 391–408.

Lynch, S., & O'Donnell, C. (2005). The evolving definition, measurement, and conceptualization of fidelity of implementation in scale-up of highly rated science curriculum units in diverse middle schools. *Annual meeting of the american educational research association* Retrieved from http://www.gwu.edu/~scale-up/documents/AERA_GWU_Fidelity_4.12.05.pdf.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty replication in the education sciences. *Educational Researcher*. http://dx.doi.org/10.3102/0013189X14545513.

Mihalic, S. (2004). The importance of implementation fidelity. *Emotional and Behavioral Disorders in Youth*, *4*(4), 83–104.

Maynard, B. R., Peters, K. E., Vaughn, M. G., & Sarteschi, C. M. (2013). Fidelity in after-School program intervention research: a systematic review. *Research on Social Work Practice*. http://dx.doi.org/10.1177/1049731513491150.

Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010). Reducing the gender achievement gap in college science: a classroom study of values affirmation. *Science*, *330*(6008), 1234–1237. http://dx.doi.org/10.1126/science.1195996.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: development, measurement, and validation. *American Journal of Evaluation*, *24* (3), 315–340. http://dx.doi.org/10.1177/109821400302400303.

Obradovic, J., Burt, K. B., & Masten, A. S. (2010). Testing a dual cascade model linking competence and symptoms over 20 years from childhood to adulthood. *Journal of Clinical Child and Adolescent Psychology: The Official Journal for the Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division 53*, *39*(1), 90–102. http://dx.doi.org/10.1080/15374410903401120.

Pressman, J. L., & Wildavsky, A. (1973). *Implementation: how great expectations in washington are dashed in Oakland*. Berkeley, CA: University of California Press.

Protheree, N. (2009). Fidelity of implementation. *Principal's Research Review*, *4*(6), 1–8.

Sanetti, L. M. H., & Kratochwill, T. R. (2009). Toward developing a science of treatment integrity: introduction to the special series. *School Psychology Review*, *38*(4), 445–459.

Sechrest, L., West, S. G., Phillips, M. A., Redner, R., & Yeaton, W. (1979). Some neglected problems in evaluation research: strength and integrity of treatments. *Evaluation Studies Review Annual*, *4*, 15–35.

Sherman, D. K., Cohen, G. L., Nelson, L. D., Nussbaum, A. D., Bunyan, D. P., & Garcia, J. (2009). Affirmed yet unaware: exploring the role of awareness in the process of self-affirmation. *Journal of Personality and Social Psychology*, *97*(5), 745–764. http://dx.doi.org/10.1037/a0015451.

Sherman, D. K., Hartson, K. A., Binning, K. R., Purdie-Vaughns, V., Garcia, J., Taborsky-Barba, S., et al. (2013). Deflecting the trajectory and changing the narrative: how self-affirmation affects academic performance and motivation under identity threat. *Journal of Personality and Social Psychology*, *104*(4), 591–618. http://dx.doi.org/10.1037/a0031495.

Shnabel, N., Purdie-Vaughns, V., Cook, J. E., Garcia, J., & Cohen, G. L. (2013). Demystifying values affirmation interventions writing about social belonging is a key to buffering against identity threat. *Personality and Social Psychology Bulletin*, *39*(5), 663–676. http://dx.doi.org/10.1177/0146167213480816.

Stein, M. L., Berends, M., Fuchs, D., McMaster, K., Saenz, L., Yen, L., et al. (2008). Scaling up an early reading program: relationships among teacher support fidelity of implementation, and student performance across different sites and years. *Educational Evaluation and Policy Analysis*, *30*(4), 368–388.

U.S. Department of Education (2009). *Implementing research-based prevention programs in schools*. U.S. Department of Education Retrieved from http://www.ed.gov/print/admins/lead/saftey/training/implementing/prevention.html.

Vanneman, A., Hamilton, L., Anderson, J. B., & Rahman, T. (2009). Achievement Gaps: How Black and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress. Statistical Analysis Report. NCES 2009-455. National Center for Education Statistics. Available from: ED Pubs. P.O. Box 1398, Jessup, MD 20794-1398. Tel: 877-433-7827; Web site: http://nces.ed.gov/help/orderinfo.asp Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED505903

Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. S. (1993). Testing the integrity of a psychotherapy protocol: assessment of adherence and competence. *Journal of Consulting and Clinical Psychology*, *61*(4), 620–630.

Wang, M. C., Nojan, M., Strom, C. D., & Walberg, H. J. (1984). The utility of degree of implementation measures in program implementation and evaluation research. *Curriculum Inquiry*, *14*(3), 249–286. http://dx.doi.org/10.2307/3202214.

Yeager, D. S., & Walton, G. M. (2011). Social-Psychological interventions in education they're not magic. *Review of Educational Research*, *81*(2), 267–301. http://dx.doi.org/10.3102/0034654311405999.

Zvoch, K. (2012). How does fidelity of implementation matter? using multilevel models to detect relationships between participant outcomes and the delivery and receipt of treatment. *American Journal of Evaluation*, *33*(4), 547–565. http://dx.doi.org/10.1177/109821.