Abstract Title Page
Not included in page count.

Title: Inside the Black Box of Self-Affirmation: Which Parts of Affirmation Exercises Are Critical for Treatment Efficacy?

Authors and Affiliations:
Christopher S. Rozek, crozek@wisc.edu- Contact
University of Wisconsin-Madison

Paul Hanselman, paul.hanselman@uci.edu
University of California, Irvine

Rachel C. Feldman, rcfeldman@wisc.edu
University of Wisconsin-Madison

Erin A. Quast, eaquast@wisc.edu
University of Wisconsin-Madison

Evan P. Crawford, crawford3@wisc.edu
University of Wisconsin-Madison

Geoffrey D. Borman, gborman@education.wisc.edu
University of Wisconsin-Madison

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

Steele and Aronson (1995) hypothesized that underperformance in academics by minority students might be partially due to a newly identified phenomenon: *stereotype threat*. Stereotype threat was defined as the anxiety or fear that an individual might experience because of the negative stereotypes about a group associated with that individual. For example, Steele and Aronson (1995) found that African American participants in laboratory studies showed worse performance on an academic test when stereotype threat was experimentally induced. This underperformance was theorized to be due to the fear of confirming the negative stereotype about African Americans students in school.

Both Nguyen and Ryan (2008) and Walton and Spencer (2009) demonstrated the large amount of evidence for the existence of stereotype threat and also for the efficacy of an intervention to alleviate the effects of stereotype threat: self-affirmation. Self-affirmation is an intervention that allows students to buffer themselves from the deleterious effects of stereotype threat by bolstering their sense of self-integrity (Steele, 1988). A typical self-affirmation exercise consists of writing about the importance of values students select from a list. This exercise can take just a few minutes to complete, but field research has shown that it can reduce the size of the racial achievement gap in schools by up to 40% (Cohen et al., 2006).

Critically, cues in the social context are necessary in order to see the harmful effects of stereotype threat as well as the benefits of self-affirmation. That is, self-affirmation will only promote academic achievement in contexts in which stereotype threat is present. For example, in a district-wide randomized controlled trial, Hanselman et al. (2014) found that self-affirmation only benefited minority students in schools with two important characteristics related to stereotype threat: schools with relatively larger racial achievement gaps and relatively fewer minority students. In those high threat contexts, stereotype threat should be greatest, and thus, self-affirmation should have the biggest impact.

Although there is much support for the existence of stereotype threat and for the efficacy of self-affirmation to combat stereotype threat and reduce racial achievement gaps, there is little consistent evidence for the mechanisms underlying stereotype threat and self-affirmation. In a systematic review of self-affirmation effects, McQueen and Klein (2006) noted that most studies rely on effects on performance to know if self-affirmation works, and there is little consistency in effects on mediators or manipulation checks.

Some studies (e.g., Shnabel et al., 2013) have shown evidence of mechanisms, but these effects have not yet been consistently replicated. The reason for the difficulty in finding a mediator for self-affirmation might be because numerous processes are involved. Stereotype threat might contribute to a host of negative effects that self-affirmation alleviates through a progressive succession of mechanisms. Cohen et al. (2009) found long-lasting effects of self-affirmation and suggested that recursive processes were involved, such that self-affirmation might affect a constellation of positive coping mechanisms that reinforce each other over time.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

In the current study, we examined students' written responses from the aforementioned Hanselman et al. (2014) district-wide scale-up study of self-affirmation to investigate how what

students wrote on self-affirmation exercises might mediate the effects of treatment on students' GPA. Instead of focusing on a particular psychological process that might underlie some of the positive effects of self-affirmation (e.g., social belonging in Shnabel et al., 2013), we tested the effect of treatment compliance, which we hypothesized might be an all-encompassing predictor of all of the subsequent distinct psychological and behavioral processes that mediate the long-term effects of self-affirmation on academic performance. Treatment compliance was defined as whether or not students wrote about one of the values listed on their exercises being important to them. We hypothesized that this is the most basic goal of self-affirmation exercises (i.e., a self-affirming writing expression); therefore, this variable should be a strong predictor of both short and long-term treatment effects on GPA. To test this hypothesis, we utilized instrumental variable analysis to create treatment on the treated estimates based on treatment compliance.

Setting:

Description of the research location.

Research occurred in all eleven middle schools in the Madison Metropolitan School District from 2011-2014.

Population / Participants / Subjects:

Description of the participants in the study: who, how many, key features, or characteristics.

We focused on the 310 African American and Latino students who were consented assented for the study (out of 910 total students in the study). These are the students for whom stereotype threat should be relevant, so we refer to them as potentially threatened students. Students entered the study in seventh grade and were followed through ninth grade for the purposes of this analysis.

Intervention / Program / Practice:

Description of the intervention, program, or practice, including details of administration and duration.

The treatment exercise consisted of the self-affirmation intervention developed by Cohen and colleagues (2006), which asked students to write about values they thought were important from a set list. Control exercises typically asked students to write about values they found to be unimportant but might be important to others. Students were randomized within schools, and teachers administered packets with identical cover pages to treatment and control students. All instructions for completing the exercises were included in the packet to be read by the students, and students were given up to four exercises to complete during the school year.

Research Design:

Description of the research design.

Intention to Treat (ITT) estimates of the self affirmation intervention effects reflect one policy-relevant parameter: the average effects if the same intervention were implemented widely. Another informative parameter is the effect of compliance with the intervention, which provides an indication of the effects of the desired self-affirmation responses themselves. The magnitude of this Treatment on the Treated (TOT) estimate speaks to the basic social processes at play and helps to gauge how effective this type of intervention might be if improved to reach all students.

Although previous reports have shown the ITT impacts of self-affirmation on seventh grade GPA for potentially threatened students (Borman, Grigg, & Hanselman, 2014) and that these effects were concentrated within high-threat school contexts (Hanselman et al., 2014), the current study examined the effects of self-affirmation on GPA from seventh through ninth grade

and explored treatment on the treated analyses via a treatment compliance variable, which was coded from students' written responses.

Data Collection and Analysis:

Description of the methods for collecting and analyzing data.

Exercise Coding. The research team developed a qualitative coding scheme with the intent of identifying treatment compliance (i.e., a self-affirming writing expression), which was defined in terms of if students 1) referred to one of the values listed in the intervention and 2) stated that the value was important to the themselves. See Figure 1 for the distribution of this variable across the four exercises.

Instrumental Variables Estimates. For the purposes of this study, we define compliance as engaging in self-affirming writing, based on our coding of the students' written exercises. So construed, compliance was largely determined by individual students' receipt of and response to the writing exercise. For instance, students absent during the relevant class period did not have a chance to engage in self-affirming writing, while some who were present simply did not follow directions. Conversely, students in the comparison conditions may have engaged in self-affirming writing despite the alternate prompt. Non-compliance of either type is expected to cause ITT estimates to understate the "pure" TOT effects.

To estimate TOT impacts, we fit two-stage least square models of the effect of self-affirmation using treatment group as an instrument. This technique identifies exogenous variation in self-affirmation related to experimental group (first stage) and estimates the effect of this variation on ultimate outcomes (second stage). Formally, we specify the following first stage equation: $M_i = \Pr(SA_i = 1) = \beta_0 + \beta_1(Treatment_i) + \boldsymbol{B}(\boldsymbol{X}_i) + \varepsilon_i$ (1)

Here the probability that student i engaged in self-affirmation writing (SA_i) is modeled as linear function of treatment status and a vector of covariates ($\boldsymbol{X}_i$). The covariates include prior GPA (2011), gender, LEP status, Special Education designation, Free/Reduced Lunch Eligibility, and a school indicator.

The second stage equation is then: $Y_i = \beta_{o'} + \beta_{1'}(\widehat{M_i}) + \boldsymbol{B}'(\boldsymbol{X}_i) + u_i$ (2)

Where Y_i is the outcome (GPA score), $\widehat{M_i}$ is the predicted value of self-affirmation for student i based on equation 1, and $\beta_{1'}$ is the main parameter of interest: the effect of self-affirmation writing.

An identifying assumption for this approach is that the instrument affects outcomes only through the endogenous variable of interest. In this case, this means that the effects of the intervention do not occur through any pathway except self-affirming writing. Though untestable, this assumption is most plausible if writing is a good proxy for the underlying social psychological responses targeted by the intervention. Because it is not clear a priori what level of writing is the best indication of meaningful self-directed affirmation, we test a range of three measures of self-affirmation writing: First, a minimal compliance definition is whether a student ever self-affirmed in any exercise. Second, we consider self-affirmation specifically during the second exercise, which came at a critical time before state accountability tests. Finally, we consider a "maximum dose" definition in which compliance represents self-affirming in each of the three core exercises (1st, 2nd, and 4th).

It is important to note that IV methods provide unbiased estimates specifically for the population of individuals for whom the instrument induced change. If treatment effects are heterogeneous, then IV estimates provide a local average treatment effect only across the population of compliers to the intervention. The estimates may not apply to two types of non-

responsive subjects: always-takers, who will engage in self-affirmation whether assigned the treatment or not, and never-takers, who will not, regardless of condition. While we cannot determine any individuals' status amongst these groups, we can estimate the size of each group in the population if we make the reasonable assumption that there are no universal non-compliers (often called "defiers"). The prevalence of self-affirmation among control students provides an estimate of the share of always-takers, while the share of non-self-affirming students in the treatment condition indicates how many of the population are never-takers. Assuming no universal non-compliers (often referred to as defiers), then the remaining population represents the size of the complier population. We present estimates of the size of each group to put the TOT in context.

Findings / Results:

Description of the main findings with specific details.

Results of two-stage least squares analyses are presented in Tables 1, 2, and 3. As is typical, TOT estimates tend to be larger than ITT estimates, suggesting some dilution of the "pure" effect of self-affirmation due to non-compliance. There are three key findings. First, there are detectable effects of self-affirmation intervention for potentially threatened students overall across all years. The magnitude of these effects vary with the measure of self-affirming writing: the most inclusive measure yields estimates close to ITT, because a majority of students are compliers, while the effects according to the most demanding criteria are twice as large, coinciding with the fact that no comparison students and a minority of treatment students wrote self-affirming responses for all three core exercises. Second, all of the effects of self-affirmation writing are concentrated in putatively more threatening schools contexts, those characterized by relatively smaller African American and Latino populations and larger achievement gaps. There are no effects of self-affirming writing in any year in low threat contexts. The difference in effects across school types is not driven by differences in compliance across these groups, however. Rather, the effects themselves are different, supporting the theory that the practice of self-affirming writing itself is beneficial only in specific contexts. Third, like the ITT treatment effects, the estimated effects of self-affirming writing grow larger over time, suggesting lasting and increasing benefits of the writing activity over time. The largest effects are seen in ninth grade for most specifications, so it is possible that the effects of the writing will persist into future grades.

Conclusions:

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

What do students need to write about in order to show positive effects from self-affirmation exercises? Our results indicated that self-affirming writing expressions (i.e., simply saying that a value is important) are critical for self-affirmation treatment efficacy. Although we found generally high levels of compliance, not all students successfully complied with the treatment. Furthermore, treatment compliance was associated with the positive effects of self-affirmation for up to three years after the exercises were given. This knowledge could be used in future studies as an immediate measure of treatment efficacy. Future studies might also use this measure of treatment compliance to design even more effective self-affirmation exercises and to encourage higher levels of treatment compliance.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Borman, G.D., Grigg, J., Hanselman, P. (2014). *An Effort to Close Achievement Gaps at Scale through Self-Affirmation*. Manuscript submitted for publication.

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science*, *313*, 1307– 1310.

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*, 400-403.

Hanselman, P., Bruch, S. K., Gamoran, A., & Borman, G. D. (2014). Threat in context: School moderation of the impact of social identity threat on racial/ethnic achievement gaps. *Sociology of Education*, *87*(2), 106-124.

McQueen, A., & Klein, W. P. (2006). Experimental manipulations of self-affirmation: A systematic review. *Self and Identity*, *5*(4), 289-354.

Nguyen, H. D., & Ryan, A. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, *93*(6), 1314-1334.

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.

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In L. Berkowitz (ed.), *Advances in experimental social psychology* (Vol. 21, pp. 261-302).

New York: Academic Press.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797-811.

Walton, G. M., & Spencer, S. J. (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science*, *20*(9), 1132-1139.

Appendix B. Tables and Figures

Not included in page count.

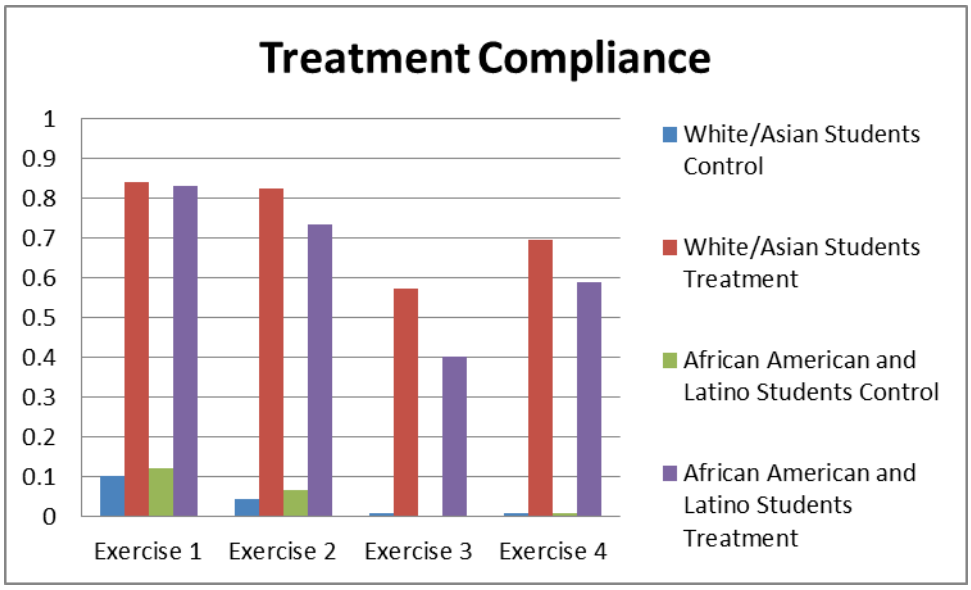

Figure 1. Distribution of treatment compliance across four exercises for both treatment and control students as well as White/Asian students and African American/Latino students.

Table 1. Instrumental Variables Estimates of the Effects of Self-affirmation Writing on 7th Grade GPA and Estimated Proportion of Always-takers, Compliers, and Never-takers.

	ITT	Ever SA	Ex 2 SA	Full SA
All Schools	0.0671+	0.0859*	0.0989+	0.169+
	(0.0365)	(0.0432)	(0.0530)	(0.0912)
N	310	310	310	310
Always-takers		0.182	0.0584	0
Compliers		0.786	0.685	0.397
Never-takers		0.0321	0.256	0.603
Low Threat Context Schools	-0.0209	-0.0255	-0.0275	-0.0500
	(0.0347)	(0.0354)	(0.0378)	(0.0697)
N	136	136	136	136
Always-takers		0.121	0.0303	0
Compliers		0.836	0.727	0.371
Never-takers		0.0429	0.243	0.629
High Threat Schools	0.131*	0.176*	0.203*	0.325*
	(0.0417)	(0.0446)	(0.0638)	(0.108)
N	174	174	174	174
Always-takers		0.227	0.0795	0
Compliers		0.749	0.653	0.419
Never-takers		0.0233	0.267	0.581

SA = Self-affirmation writing; Standard errors in parentheses.

+ $p < 0.1$; * $p < 0.05$

Table 2. Instrumental Variables Estimates of the Effects of Self-affirmation Writing on 8th Grade GPA and Estimated Proportion of Always-takers, Compliers, and Never-takers.

	ITT	Ever SA	Ex 2 SA	Full SA
All Schools	0.144*	0.184*	0.214*	0.371*
	(0.0427)	(0.0543)	(0.0708)	(0.129)
N	298	298	298	298
Always-takers		0.182	0.0541	0
Compliers		0.784	0.679	0.387
Never-takers		0.0333	0.267	0.613
Low Threat Context Schools	-0.0209	-0.0255	-0.0275	-0.0500
	(0.0347)	(0.0354)	(0.0378)	(0.0697)
N	136	136	136	136
Always-takers		0.121	0.0303	0
Compliers		0.836	0.727	0.371
Never-takers		0.0429	0.243	0.629
High Threat Schools	0.218*	0.298*	0.351*	0.570*
	(0.0525)	(0.0625)	(0.0933)	(0.191)
N	164	164	164	164
Always-takers		0.238	0.0833	0
Compliers		0.737	0.629	0.400
Never-takers		0.0250	0.287	0.600

SA = Self-affirmation writing; Standard errors in parentheses.

+ $p < 0.1$; * $p < 0.05$

Table 3. Instrumental Variables Estimates of the Effects of Self-affirmation Writing on 9th Grade GPA and Estimated Proportion of Always-takers, Compliers, and Never-takers.

	ITT	Ever SA	Ex 2 SA	Full SA
All Schools	0.160+	0.203*	0.235+	0.424+
	(0.0835)	(0.0963)	(0.120)	(0.229)
N	282	282	282	282
Always-takers		0.176	0.0423	0
Compliers		0.788	0.686	0.371
Never-takers		0.0357	0.271	0.629
Low Threat Context Schools	0.0705	0.0855	0.0912	0.164
	(0.0806)	(0.0827)	(0.0906)	(0.161)
N	130	130	130	130
Always-takers		0.115	0.0164	0
Compliers		0.842	0.737	0.377
Never-takers		0.0435	0.246	0.623
High Threat Schools	0.245+	0.328*	0.383*	0.682+
	(0.137)	(0.149)	(0.192)	(0.373)
N	152	152	152	152
Always-takers		0.222	0.0617	0
Compliers		0.750	0.642	0.366
Never-takers		0.0282	0.296	0.634

SA = Self-affirmation writing; Standard errors in parentheses.

+ $p < 0.1$; * $p < 0.05$