Bolstering Trust and Reducing Discipline Incidents at a Diverse Middle School: How Self-Affirmation Affects Behavioral Conduct During the Transition to Adolescence

Kevin R. Binning[1], Jonathan E. Cook[2], Valerie Purdie-Greenaway[3], Julio Garcia[4], Susie Chen[1], Nancy Apfel[5], David K. Sherman[6], Geoffrey L. Cohen[7]

[1] University of Pittsburgh, Department of Psychology & Learning Research and Development Center

[2] Pennsylvania State University, Department of Psychology

[3] Columbia University, Department of Psychology

[4] Stanford University, Graduate School of Education

[5] Yale University, Department of Psychology

[6] University of California, Santa Barbara, Department of Psychological and Brain Sciences

[7] Stanford University, Graduate School of Education & Department of Psychology

Corresponding Author: Kevin Binning, 210 S. Bouquet St, Pittsburgh, PA 15213

Email: kbinning@pitt.edu; Phone: 412 624 7473

Abstract

A three-year field experiment at an ethnically diverse middle school (N = 163) tested the hypothesis that periodic self-affirmation exercises delivered by classroom teachers bolsters students' school trust and improves their behavioral conduct. Students were randomly assigned to either a self-affirmation condition, where they wrote a series of in-class essays about personally important values, or a control condition, where they wrote essays about personally unimportant values. There were no behavioral effects of affirmation at the end of 6th grade, after students had completed four writing exercises. However, after four additional exercises in 7th grade, affirmed students had a significantly lower rate of discipline incidents than students in the control condition. The effect continued to grow and did not differ across ethnic groups, such that during 8th grade students in the affirmation condition on average received discipline at a 69% lower rate than students in the control condition. Analyses of student climate surveys revealed that affirmation was associated with higher school trust over time, a tendency that held across ethnic groups and partially mediated the affirmation effect on discipline. Repeated self-affirmation can bolster students' school trust and reduce the incidence of discipline in middle school, findings with both theoretical and practical implications.

Bolstering Trust and Reducing Discipline Incidents at a Diverse Middle School: How Self-Affirmation Affects Behavioral Conduct During the Transition to Adolescence

Students' transition to middle school is often marked by a confluence of academic, developmental, and social stressors (Eccles et al., 1993). Although some stress is necessary and healthy to build resilience, stressors that cause students to question their fundamental adequacy as a student can be destructive. Such threats to students' self-integrity can cause them to disengage from school and slide into downward academic trajectories (Major, Spencer, Schmader, Wolfe, & Crocker, 1998; Steele, 1992). Self-affirmation interventions in education have been used as a means to bolster self-integrity and thereby improve students' educational outcomes (e.g., Cohen, Garcia, Apfel, & Master, 2006; Cohen, Garcia, Purdie-Vaughns, Apfel, & Brzustoski, 2009; Cook, Purdie-Vaughns, Garcia, & Cohen, 2012; Martens, Johns, Greenberg, & Schimel, 2006; Powers et al., 2016; Sherman et al., 2013; Tibbetts et al., 2016).

Self-affirmation interventions usually involve brief exercises in which people write essays about personally important values, such as relationships, religion, or athletics. By providing a "psychological time-out" (Lyubomirsky & Della Porta, 2010) to identify and reflect on their values, the interventions help people adopt a more expansive self-view from which specific threats appear to be more manageable (Cohen & Sherman, 2014; Critcher & Dunning, 2015; Sherman, 2013; Steele, 1988). The first field experiments in schools used within-classroom random assignment and found that affirmations can improve academic performance among students belonging to a negatively stereotyped subgroup, African Americans (Cohen et al., 2006), with benefits (e.g., higher grade point average; GPA) that persisted for several years after the intervention (Brady et al., 2016; Cohen et al., 2009; Goyer et al., 2017). Replications and extensions showed that affirmations can confer benefits to students theorized to be under

psychological threat, including women undergraduates in physics classrooms (Miyake et al., 2010), Latinx American[1] middle school students (Sherman et al., 2013), first generation college students in biology classes (Harackiewicz et al., 2014), and women MBA students (Kinias & Sim, 2016).

In the present research, we suggest that while self-affirmation theory and research has broad relevance to many domains (e.g., Ferrer & Cohen, 2018; Sherman, Lokhande, Müller, & Cohen, in press), self-affirmation research in education has been almost singularly focused on one important behavioral outcome, performance (e.g., grades, test scores). This focus has yielded insight into group achievement gaps, but it has largely overlooked another pressing problem in US secondary schools and society more broadly, the ethnic group discipline gap (Gregory, Skiba, & Noguera, 2010; Okonofua, Walton, & Dweck, 2016). Discipline problems tend to spike among students-in-general during middle school (Skiba et al., 2011), but disciplinary citations tend to soar among students of color. In one analysis of 3.5 million public school students, African American secondary school students were suspended at over three times the rate as White students (Losen, Hodson, Keith, Morrison, & Belway, 2015). Latinx students were suspended at one and half times the rate of White students (Losen et al., 2015). Students' experiences with discipline can be highly consequential, serving in some cases as an inflection point that changes their self-narratives (see Wilson, 2011) and increases their likelihood of subsequent encounters with the criminal justice system (Hirschfield, 2009; Pettit & Western, 2004).

The present study examines the effects of a randomized affirmation intervention on students' discipline incidents at an ethnically diverse middle school. As we outline below, when

[1] We use Latinx as a gender-neutral term to refer broadly to all people identified as Hispanic, Chicana/o, and/or Latina/o.

students have behavioral problems, such as when they violate school rules and norms, we argue that they may do so, in part, as a means to defend and bolster their sense of self-integrity. Thus a well-timed values affirmation intervention that bolsters self-integrity may alleviate pressure to defend the self and reduce students' behavioral conduct problems. To understand how this process may unfold psychologically, we borrow from research implicating students' trust in their teachers as a potentially critical predictor of problem behaviors (Gregory & Ripski, 2008; Tyler & Degoey, 1996; Yeager, Purdie-Vaughns, Hooper, & Cohen, 2017). That is, if students do act out as a means to protect and bolster the self, they may do so because they have lost trust in their school's authorities, such as their teachers and administrators.

Trust, Discipline, and the Self

Too often, the use of school discipline has the opposite of its desired intentions, making students more rather than less likely to act out in the future (Amemiya, Fine, & Wang, 2019; Okonofua, Paunesku, & Walton, 2016). Trust is critical in this process because students' trust in teachers and administrators (hereafter referred to as school trust) shapes their perceptions and reactions to disciplinary behavior (Way, 2011). Discipline from an untrusted source may backfire because it is seen as illegitimate, contributing to a recursive cycle in which the distrust increases, triggering more behavior problems, more disciplinary responses, and more distrust (see Cohen & Sherman, 2014; Goyer, Cohen, Cook, Master, Apfel, Lee, Henderson, Reeves, Okonofua, & Walton, in press; Sherman & Cohen, 2006; Yeager et al. 2014). As noted, discipline is not meted out evenly across ethnic groups (see Okonofua, Walton, Eberhardt, 2016). Real and perceived violations of procedural and relational trust, such as discriminatory behavior, unwarranted punishment, and biased treatment (Bies & Tripp, 1996), can cause trust to deteriorate.

The present research theorizes that trust involves a willingness to be vulnerable to another person (Rousseau, Sitkin, Burt, & Camerer, 1998). It is a "cognitive leap" (Gambetta, 1988), in which people assume that another person will meet obligations and expectations to accord one a measure of respect and regard (Cohen & Steele, 2002). To trust puts people at risk of being subjected to biased treatment—which, in addition to its material costs, can threaten people's sense of being valued by the larger group to which they belong (Tyler, Degoey, & Smith, 1996). By contrast, distrust is a way to protect the self from this vulnerability. It is often an adaptive response to the social reality of discrimination and bias. We suggest that, regardless of the actual trustworthiness of teachers and other authority figures in school, students may be reluctant to trust because doing so puts the self at risk. Accordingly, we hypothesize that self-affirmations that support self-integrity might help to foster and maintain trust.

Indeed, some evidence already suggests that affirmations bolster trust and improve behavioral conduct. In a field experiment, affirmed participants reported higher prosocial feelings and their teachers' rated them as more prosocial for at least three-months after the intervention (Thomaes, Bushman, de Castro, & Reijntjes, 2012). However, the evidence is not uniform, as another one-year intervention did not find hypothesized effects of affirmation on students' problem behaviors (de Jong, Jellesma, Koomen, & de Jong, 2016). This means that while affirmation can affect trust and discipline, it will not necessarily do so, and more research is needed to understand when, why, and for whom affirmation interventions may yield positive effects.

Trust and Discipline in Diverse Settings

Although school trust is important for all students, some circumstances make it relatively more difficult for some students to trust. For example, students of color may wonder if others are

biased by negative stereotypes about their group (Crocker & Major, 1989). This can lead them to mistrust critical feedback (Cohen, Steele, & Ross, 1999; Yeager et al., 2014) and increases the chances that teacher intentions will be misread (Yeager et al., 2017). Moreover, research finds that feeling stereotyped increases norm deviance (Belmi, Barragan, Neale, & Cohen, 2015), and that people tend to avoid, disengage from, or de-identify with domains where they feel negatively stereotyped (Major, Spencer, Schmader, Wolfe, & Crocker, 1998; Schmader & Sedikides, 2018; Steele, 1992). Feeling disrespected or ostracized increases the likelihood of aggression (Twenge, Baumeister, Tice, & Stucke, 2001; Thomaes, Bushman, de Castro, Cohen, & Denissen, 1999; Dodge & Somberg, 1987) and reduces feelings of school identity and individual self-worth (Huo, Binning, & Molina, 2010). In one recent set of studies (Yeager et al., 2017), school trust among African American and Latinx American students declined during middle school in tandem with perceptions of group-based discrimination.

The present study was conducted in a middle school setting with a high level of ethnic diversity, raising a number of theoretically and practically important considerations for examining affirmation, trust, and discipline. In light of prior research, we were particularly interested in whether the effects of the affirmation intervention on trust and disciplinary incidents would be strongest among Black and Latinx students. We note, however, that non-stereotyped students may also grow mistrustful of academic authorities following a difficult transition to middle school. Research has shown that while declines in trust tend to be particularly steep among Black and Latinx students, it also tends to decline among White students (Yeager et al., 2017). More generally, regardless of race, students from working-class backgrounds have shown steady declines in trust and social capital in American society (Putnam, Frederick, & Snellman, 2012). Accordingly, one possibility is that students could benefit from the protective effects of

affirmation regardless of race, particularly during the sensitive developmental period that characterizes middle school and the transition to adolescence.

Present Research

The present report examines the possibility that affirmation affects disciplinary behavior and, if so, the extent to which a possible psychological driver of disciplinary misconduct, trust, explains that effect. The longitudinal design encompasses the middle school career of one cohort of students, from 6th through 8th grade. Thus the study focuses on a stressful developmental window that confronts adolescents with a variety of threats and times of vulnerability (see Eccles et al., 1993). By repeatedly delivering affirmation exercises throughout this window and measuring trust and discipline along the way, we hoped to maximize the chance that affirmation would exert a favorable effect. That is, we sought to provide multiple entry points for the "trigger and channel" effects described in Goyer et al. (2017). When interventions trigger positive psychological processes at times of stress, the effects of these processes can be channeled into positive behavioral consequences. The study used within-classroom random assignment to deliver the affirmation just after students transferred to middle school. Students then maintained their initial assignment and received the intervention repeatedly, up to 9 times total during their middle school tenure. Discipline data were drawn from official school records, and we tested the following hypotheses.

1. Affirmation will forestall reductions of school trust during the school year.
2. Affirmed students will have a lower rate of yearly discipline incidents as recorded in administrative records.
3. The affirmation effect on trust will mediate the affirmation effect on discipline incidents.

4. These effects will be especially strong for Black and Latinx students compared to White and Asian students, thereby reducing the magnitude of the discipline gap.

Method

Participants

The study took place at a middle school serving a medium-sized suburban town on the Eastern seaboard of the United States, in which approximately one third of students received free or reduced-price lunches. At the beginning of the first year of the study, we sought consent to participate from all students in 6th grade. The study used an active consent procedure in which parents provided consent (or not) for their children to participate in the longitudinal study. In all, 55% of 6th graders (N = 163) provided consent and enrolled in the study. Using administrative records the sample was 50% female and 50% male. The sample was also 48% White, 39% Black/African American, 7% Latinx American, and 6% Asian/Asian American, a breakdown that approximated the demographics of the school population. The ethnic group classification was based on administrative records, except in one case where a student who was listed as "other" in the administrative records was placed in the White category based on self-reported ethnicity. To simplify the ethnic group analyses while retaining the full sample, ethnicity was dichotomized into groups in line with academic stereotypes, resulting in students from ethnic groups with negative intellectual stereotypes (i.e., African Americans and Latinx Americans) and those from groups with positive intellectual stereotypes (i.e., Whites and Asian Americans). Although this full-sampling approach departs from previous studies that analyzed just African American and White students (e.g., Cohen et al., 2006), all results remain virtually unchanged when the analyses were limited to African American and White students (N = 141), and thus we opted for the more inclusive classification.

Attrition and Missing Data

Over the course of the three-year study, 18 students were lost to attrition (see Table 1). There was no heterogeneous attrition, as nine were lost in the affirmation condition and nine were lost in the control condition. In all, 89% of the original students who began the study as 6th graders continued to participate as 8th graders. For the analyses reported below, we used all available student data, and we handled missing data by using likelihood-based estimation on whatever data for each participant were available (Enders, 2010). Parallel analyses that included only complete cases yielded consistent and statistically significant results.

Intervention Design

The first self-affirmation activities were administered just after the transition to middle school. Maintaining initial condition assignments, students received up to eight additional administrations through 6th, 7th, and 8th grades. Each student was randomly assigned to either the affirmation or control condition using a within-classroom stratification procedure to ensure a roughly equal number of each gender, and of members of each ethnic group, in the affirmation and control conditions within each classroom. Students who did not provide consent were provided an alternative assignment to occupy their attention during class sessions in which the interventions were delivered. The study followed the schedule depicted in Table 2.

For logistical reasons, the course in which the control and intervention exercises were delivered differed each school year. The intervention was delivered in students' language arts course in 6th grade, their mathematics course in 7th grade, and their science course in 8th grade. The teacher for each course always led the intervention. An attempt was made to time the interventions to occur on stressful or evaluative days (e.g., days with a test in class) – that is, on days when threat, and the opportunity to address that threat, might be greatest. In each year, the

intervention was delivered across 15 classrooms, administered by three teachers who taught five periods each (three teachers each year, with nine teachers in total participating). These teachers comprised all the mainstream teachers at the school teaching their respective subject in each grade. In each school year, checks on differences between teachers who administered the intervention revealed no mean differences between teachers on any focal outcomes. We therefore collapsed across teachers.

Teacher training. Since the interventions were always delivered by students' teachers, teachers were extensively trained each year on how to administer the protocol while maintaining the fidelity of the double-blind research design. Our efforts were in line with theorizing on the importance of keeping social psychological interventions subtle to reduce reactance (Garcia & Cohen, 2013; Yeager & Walton, 2011). Thus we sought to ensure teachers would make each intervention session feel like a normal, engaging, unobtrusive exercise. Rather than alerting students that they were taking part in a research study that was designed to be helpful, teachers were taught to present the intervention as part of regular classroom activities. Teachers were not told the hypotheses of the study, and considerable effort went into reducing the likelihood that both the teachers and students would become aware of the different treatment and control conditions. For example, the affirmation and control materials were formatted similarly and designed to look as similar to each other as possible. This way, at a glance, teachers and students would not notice differences between them (as in Cohen et al., 2006, 2009; Sherman et al., 2013). Teachers were explicitly instructed not to tell students that the activities were designed to be helpful because awareness of affirmation's intended benefits can undermine its impact (Sherman et al., 2009). To ensure teachers were prepared to answer questions, they were provided a script that specified generic answers to questions that students might ask. If students

asked about the purpose of the activity, for instance, teachers were instructed to state that the activity was simply something that they and the school would like them to complete. Although we did not collect data on how closely teachers adhered to the protocol, our research staff met with teachers both before and after each intervention to verify that they followed the protocol and to address questions with them about its execution.

6th and 7th grade implementation. As shown in Table 2, there were three main variants of the affirmation procedure. These variants were intended to make the repeated affirmation activities feel novel and engaging. Over 6th and 7th grade, the timing and order of the materials were very similar, as students completed two standard affirmation procedures that have been widely used in past research (see McQueen & Klein, 2006), followed by two procedures that have also been used in prior research (e.g., Sherman et al., 2013, Study 1). For the most frequent affirmation manipulation (also the first affirmation manipulation given each school year), students were presented with the following list of values: Athletic Ability, Being Good at Art, Creativity, Independence, Living in the Moment, Membership in a Social Group, Music, Politics, Relationships with Friends and Family, Religious Values, Sense of Humor. Students in the affirmation condition were asked to select their three most important values and to write a brief essay "to explain why those values are important to you." Below is a representative example of a student's affirmation essay:

> My independence is important to me because I don't want to go through life with people telling me what to do. This is my life and I want to run it my way, and if I don't start now, I will never learn. Living in the moment is important to me because you shouldn't

> plan everything. I like it when whatever happens, happens. Relationship with friends & family is important because they are always there for me when I need them.

Students in the control condition were prompted to write about various non-affirming topics, such as their unimportant values (and why they might be important to others), their morning routine, or their typical afternoon. For example, for several exercises students in the control condition selected their three least important values and wrote an essay "to explain why those values might be important to someone else." Below is a representative control essay:

> Religious values aren't important to me because I don't have a religion. However, for someone with a religion, religious values are very important. Politics I find boring, and aren't a big thing in my life. However, If you're say…running for president, politics mean everything.

Materials were delivered on four occasions in each of the first two years (6th and 7th grade), and they followed the same time schedule for each year (see Table 2). Students were on a traditional nine-month calendar. They completed the first and second exercises in September and October, respectively. They completed additional exercises in early January and March.

8th grade implementation. In 8th grade, due to a funding gap, only one writing exercise was administered. As reflected in Table 2, the first survey in 8th grade took place in January and the lone affirmation exercise took place in March. This manipulation was intended to test whether a ninth affirmation would have an incremental impact above and beyond those who received only eight affirmations. Students who had been assigned to the affirmation condition

were randomly allocated to two different groups. In one condition, the "booster" affirmation condition, they received an additional, 9th affirmation. In the "no booster" affirmation condition, they received a control exercise. All students who had previously been in the control condition received another control exercise. However, the results of the booster analyses were inconclusive across both primary outcomes of interest in 8th grade. We conducted a variety of contrast analyses comparing the booster, no booster affirmed students, and control students. These results were in the direction suggesting that the booster had a slight but not statistically significant benefit. As such, we collapsed across the booster variable for the analyses below.

School Trust. As part of a larger battery of questions, school trust was assessed six times (twice each year, once at the beginning of the school year and once at the end) with five items that had end-points labeled 1 (*Very much disagree*) and 6 (*Very much agree*): "I am treated fairly by teachers and adults at (school name)," "When students at (school name) break the rules, their punishment is decided in a fair way," "Teachers give me the grades I think I deserve," "Teachers and other adults treat me with respect," "My teachers at (school name) have a fair and valid opinion of me." Alpha reliabilities within each year were acceptable (.73-.84). The distribution of trust was negatively skewed in each of the six measurement occasions, indicating that trust scores tended to be concentrated at the high end of the scale (skewness range -1.06 to -.75; kurtosis range .02 to 1.02).

Discipline incidents. Students' discipline incidents, which were the sum of office referrals and suspensions, were obtained from the school's administrative records. In addition to the yearly discipline count, the school provided a pre-intervention count of discipline incidents during the first two weeks of 6th grade. They also provided the count of all remaining (post-intervention) 6th grade discipline incidents, all 7th grade incidents, and all 8th grade incidents (*Ms*

= 0.07, 1.49, 2.36, and 2.36; *SDs* = 0.26, 2.99, 4.99, and 4.53 for pre-intervention, 6th grade, 7th grade, and 8th grade, respectively). As expected, discipline incidents were positively skewed with high kurtosis (from years 1-3, skewness range 2.56 to 3.98; kurtosis range 7.14 to 19.27). Given the strong violations of normality, high variance, and the counted nature of the data, we analyzed the discipline incidents with a multi-level analysis using Poisson regression model with overdispersion.

Tests of Baseline Equivalence. We conducted a number of *t*-tests and chi-square analyses to verify our assumption that random assignment was successful in creating baseline equivalence between conditions at the beginning of the study. As shown in Table 3, chi-square tests revealed there were no condition differences in the representation of gender or ethnic groups between conditions, and *t*-tests revealed there were no differences in pre-intervention trust (measured prior to random assignment), pre-intervention discipline incidents, or pre-intervention performance (calculated as the z-scored mean of students' 5th grade GPA and state standardized test scores; all *ps* >.503). As such, we retained the assumption that random assignment was successful in creating equivalent groups at the beginning of the study.

Unreported Outcome Measures. Several additional variables were collected and analyzed as part of this broader research project that are not reported here. In addition to those mentioned above, we collected data on student grades and a variety of self-report measures tapping threats to students' self and social identities (e.g., low belonging, feelings of stereotype threat). The results on student grades showed a similar pattern to the results on discipline (as described below), with a statistically significant, positive main effect of affirmation that emerged over time for the full sample. However, the discipline results held even after controlling for student grades, suggesting the treatment exerted discrete effects on these two outcomes.

Moreover, exploratory analyses found that the measure of trust did not load consistently with measures of identity threat, suggesting that the two psychological variables were distinct. The findings involving grades and threat will be presented in a forthcoming report focused specifically on the effects of affirmation on academic achievement.

Model Specification. Although random assignment was conducted within students' 6th grade classrooms, the fact that students were grouped into classrooms when they received the intervention raised the possibility of non-independence of observations. As such, we conducted an initial set of analyses to determine if it was necessary to account for non-independence at the classroom level by nesting students within classrooms for the analyses. We compared empty (unconditional) models on trust and discipline that had no predictors to models that included only intervention classroom dummy-codes at Level 2. The analyses consistently revealed no differences between classrooms and that adding these codes had no discernible impact on the total amount of variance explained by the models. For example, adding students' 6th grade teachers to the models produced no change in model fit for either trust or discipline (ΔR^2 =.00 for both trust and discipline). As such, in the analyses below we ignored the classroom groups.

Analyses on trust and discipline were conducted using restricted maximum likelihood estimation, with repeated observations (e.g., yearly discipline) at Level 1 nested within students at Level 2. All multi-level models were analyzed using HLM 7.0 software. All analyses included a consistent set of control variables. Namely, we always controlled for participant ethnicity, gender, and number of prior discipline incidents (i.e., incidents prior to the initial intervention in 6th grade). Whereas trust was analyzed with a OLS regression model, the counted and skewed nature of the discipline data led us to analyze discipline with a Poisson regression model with a log-link function, equal exposure assumptions (since all students had discipline reported over the

same school year), and over-dispersion to account for the inflated variance. Inferential tests for all multi-level regression analyses were based on robust standard errors.

Using Raudenbush and Bryk's (2002) notation, the Level 1 equation for the analysis of trust is represented as follows:

$$Y_{ti} = \pi_{0i} + \pi_{1i}a_{ti} + \pi_{2i}a_{ti}^2 + e_{t1}$$

In this equation, *Y* represents the observed status on an outcome variable (e.g., discipline incidents) at time *t* for individual *i*, π_{0i} represents the intercept for person *i* at the beginning of the intervention (initial status), $\pi_{.i}$ represents the rate of change for person *i*, *a* represents the point in time at which person *i*'s outcome is estimated, and *e* represents random error. The coefficients for the non-linear change were calculated by simply squaring the linear change term (a^2). The Poisson analysis on discipline incidents also included a log-link function at Level 1.

Analyses on trust and discipline used the same Level 2 model:

$$\pi_{pi} = \beta_{p0} + \beta_{p1}(Treatment_i) + \beta_{p2}(Gender_i) + \beta_{p3}(Ethnicity_i) + \beta_{p4}(PreDiscipline) + r_{0i}$$

For example, for $p = 0$ (initial status), the equation holds that students' outcomes are a function of the sample mean at the beginning of the intervention (i.e., the intercept, β_{00}), whether person *i* was in the affirmation condition (β_{01}), his or her gender (β_{02}), ethnic grouping (β_{03}), pre-intervention discipline incidents (β_{04}), and individual error (r_{1i}). This same equation was then applied to estimate variability in Level 1 slopes (linear and quadratic). Doing so allowed us to test our focal research questions, such as whether the linear effect of time (e.g., discipline

incidents increasing over time) was further moderated at Level 2 (e.g., with different rates of change in the affirmation versus control conditions).

Results

School Trust

Table 3 displays the output from the multi-level models used to estimate trust and discipline incidents. In support of Hypothesis 1, students' trust in their teachers was significantly bolstered by affirmation over time. The analyses used multi-level modeling and began by examining the trajectory of trust over time for the full sample. Trust decreased over time but this decrease was non-linear. That is, our examination of the unconditional model with no Level 2 predictors (Raudenbusch & Bryk, 2002) found both linear, B = -.23, SE = .05, $t(162)$ = -4.88, p < .001, and quadratic change in trust for the full sample over time, B = .02, SE = .01, $t(162)$, 2.60, p = .008. As suggested by the negative sign on the linear term and the positive sign on the quadratic term, the non-linearity was due to students' trust declining more quickly during the middle portion of middle school than at the beginning or end. We retained both the linear and quadratic terms at Level 1 and proceeded to analyze effects at Level 2.

There was a Condition × Time interaction on the linear slope, B = .08, SE = .03, $t(158)$ = 2.51, p = .013. A comparison of the trends over time in each condition revealed that in the control condition, there was a significant decrease in trust each measurement period, B = -.27, SE = .05, $t(158)$ = -5.25, p < .001. However, as indicated by the significant interaction, this decrease was tempered in the affirmation condition, B = -.18, SE = .05, $t(158)$ = -3.90, p < .001.

To decompose how affirmation affected trust during each school year and to test for possible ethnicity differences in the deterioration of trust over time, we conducted follow-up analyses on the change in trust within each school year (e.g., changes in trust from the beginning

of 6th grade to the end of 6th grade). That is, we were interested in how trust changed over the course of the school year (and less interested in how it changed between school years). To avoid the well-known problems with using simple difference scores (Cronbach & Furby, 1970), we calculated and saved unstandardized residual scores for each participant. These scores represented the difference between students' actual trust level at the end of each school year and the trust level they were predicted to have from a regression analysis on their trust level at the beginning of that school year. Thus, positive scores indicated higher than expected trust levels at the end of the year, while negative scores indicated lower than expected trust levels at the end of the year.

Multi-level modeling using the same controls as above found two main effects. First, White and Asian students reported significantly higher than expected trust levels compared to Black and Latinx students, averaging over all three years, $B = .26$, $SE = .08$, $t(158) = 3.23$, $p = .002$. Second, affirmation was associated with higher than expected trust levels averaging across all three years, $B = .20$, $SE = .08$, $t(158) = 2.55$, $p = .012$. However, inconsistent with Hypothesis 4, this effect was not moderated by racial group status (Treatment × Racial grouping; $B = -.09$, $SE = .16$, $t(157) = -0.58$, $p = .563$), indicating that affirmation predicted higher than expected trust, regardless of students' racial background. Further probing revealed the change within 7th and 8th grades to be the main drivers of this trend, as students in the affirmation condition had significantly higher than expected trust levels at the end of 7th grade compared to students in the control condition, $B = .18$, $SE = .08$, $t(158) = 2.29$, $p = .023$. They also had significantly higher than expected trust levels at the end of 8th grade, $B = .26$, $SE = .10$, $t(158) = 2.52$, $p = .013$. By contrast, affirmation did not predict changes in trust in 6th grade, $B = .08$, $SE = .12$, $t(158) = 0.78$, $p = .437$.

The lack of significant moderation by race was surprising, and as such we plotted the estimated changes in trust in 7th grade for both Asian/White and Black/Latinx students. This result is depicted in Figure 1. It reveals that although affirmation had an overall main effect, there was also a powerful main effect of race, $B = .30$, $SE = .08$, $t(155) = 3.82$, $p < .001$, and this effect was not moderated by affirmation, $B = -.10$, $SE = .15$, $t(155) = -0.63$, $p = .531$. This main effect replicates prior work showing that Black and Latinx students tend to show marked declines in trust during 7th grade (Yeager et al., 2017). However, there was no Ethnicity × Condition interaction, indicating that affirmation did nothing to address the ethnic group trust gap. This finding was inconsistent with Hypothesis 4.

Discipline Incidents

In support of Hypothesis 2, affirmation significantly reduced the rate of discipline incidents over time. However, Hypothesis 4 was again not supported, as there was no difference in the affirmation effect between ethnic groups. Analyses began by modeling the discipline trend over time for the full sample (Raudenbush & Bryk, 2002). Results found that discipline incidents increased linearly from 6th to 8th grade but the non-linear portion of the model also revealed a significant spike in incidents in 7th grade. That is, our examination of the unconditional model with no Level 2 predictors found both linear, $B = .64$, $SE = .11$, $t(162) = 3.13$, $p = .002$, and quadratic change in discipline incidents over time, $B = -.22$, $SE = .07$, $t(162) = -3.03$, $p = .003$, with the negative quadratic change slope capturing the spike in 7th grade. As such, we retained both terms at Level 1 and proceeded to analyze effects at Level 2.

There was a Condition × Time interaction on the linear slope, $B = -.44$, $SE = .11$, $t(158) = -4.09$, $p < .001$. Figure 2 depicts both the raw means and the trends predicted by the full model (including the quadratic term). A comparison of the trends over time in each condition revealed

that in the control condition, there was a significant increase in discipline incidents each year, *B* = 1.02, *SE* = .20, *t*(158) = 5.13, *p* < .001. This increase was tempered in the affirmation condition, *B* = .58, *SE* = .16, *t*(158) = 3.72, *p* < .001. Notably, the effect of affirmation emerged over time. In 6th grade, there was no effect of condition, *B* =-.28, *SE* = .27, *t*(158) = -1.02, *p* = .312, *Event ratio* (*ER)* = .76. But during 7th grade, affirmed students had significantly fewer discipline incidents, as the event ratio indicated that students in the affirmation condition had just .49 incidents for every one incident among control students, *B* = -0.72, *SE* = .28, *t*(158) = -2.51, *p* = .013, *ER* = .49. This effect grew in 8th grade, *B* = -1.15, *SE* = .33, *t*(158) = -3.46, *p* = .001, *ER* = .31. This indicated that students in the affirmation condition experienced discipline at just 31% of the rate seen among control students.

Again Hypothesis 4 was not supported. That is, ethnic group status did not moderate the significant Time x Condition interaction (*B* = -.28, *SE* = .20, *t*[157] = -1.37, *p* = .172). With the interaction term dropped from the model, however, there were two notable effects. First, there was a main effect of ethnic status in 6th grade (Year 1), indicating that White and Asian students were significantly less likely to have discipline incidents than were Black and Latinx students, *B* = -1.47, *SE* = .29, *t*(158) = -5.03, *p* < .001. This finding replicates previous research documenting ethnic group disparities in disciplinary incidents arising in the teenage years (Noguera, 2003; Okonofua, Paunesku, & Walton, 2016; Skiba, Michael, Nardo, & Peterson, 2002). There was also a significant Ethnicity × Time interaction on the linear slope, *B* = .29, *SE* = .11, *t*(158) = 2.55, *p* = .012. This effect indicated that the ethnic group gap in discipline incidents lessened in each subsequent year after 6th grade. However, none of these effects interacted with affirmation status.

Mediation Analysis: Affirmation, Trust, and School Discipline

If reductions in trust are a defensive response to protect the self-concept, then affirmations designed to bolster the self may ward off a deterioration in trust and result in more favorable behavioral conduct and fewer discipline incidents over time (Hypothesis 3). The analyses above indicated that affirmation altered the trajectory of discipline and trust over time, as students who were randomly assigned to repeat multiple values affirmation tasks during middle school reported significantly higher trust and received significantly fewer discipline incidents in the last two years of middle school. It did so, moreover, regardless of students' ethnic background.

Our next step was to test whether the school trust mediated the effect of affirmation on school discipline incidents. In particular, we tested how within-school year changes in trust related to the number of discipline incidents students had in each school year. Given the results above showing the largest affirmation effects in 7th and 8th grade, we were particularly focused on these two grades. First, as shown in the pattern of correlations in Table 1, only changes in trust during 6th and 7th grade were associated with discipline incidents. The change in trust students showed in 8th grade was unrelated to discipline that year (r = -.03, p =.703). Given that previous analyses indicated affirmation had little to no impact in 6th grade, and given the non-linear trends above pointing to 7th grade as a critical year, we focused our analyses on 7th grade. Results should be interpreted with caution.

First, analyses indicated that changes in trust in a given year tended to predict the *subsequent* year's discipline incidents. That is, declines in trust in 6th grade predicted increased discipline incidents in 7th grade, B = -.82, SE =.30, p = .007, while the change in trust within 7th grade was not a significant predictor of 7th grade discipline, B = -.23, SE = .32, p = .475. Similarly, declines in trust during 7th grade predicted increased discipline in 8th grade discipline,

B = -1.01, *SE* = .36, *p* = .005. Residual trust scores from 6th and 8th grade were not predictive of 8th grade discipline (*Bs* = -.51 and -.28, *ps* > .134).

The evidence above points to a plausible temporal sequence involving affirmation: affirmation mitigated the deterioration of trust in 7th grade which, in turn, contributed to fewer discipline incidents in 8th grade (see Figure 3). Analyses supported this model, in support of Hypothesis 3. Mediation analyses estimated the indirect effect of affirmation on discipline (Condition→Trust→Discipline) using a non-parametric bootstrapping procedure. It estimated the indirect effect 1000 times with 1000 random samples of the data with replacement. The indirect effect was estimated from the mean of these re-samples and confidence intervals were created based on the distribution of the estimates around the mean estimate (Hayes, 2014). An estimate in which the value zero is not included in the 95% confidence interval indicates a significant indirect effect. This non-parametric approach does not make assumptions about normality of the data, and thus it was preferred due to the non-normality of discipline incidents. With all three residual trust scores in the model, only the 7th grade trust variable yielded an indirect effect on 8th grade discipline with a confidence interval that did not include zero, *Indirect effect* = -.27, *Boot SE* = .17, *95% CI* = -.72 to -.03. This indicates 7th grade changes in trust significantly mediated the effect of affirmation on 8th grade discipline.

Notably, further analyses indicated that the reverse did not hold, as discipline did not predict subsequent changes in trust. Discipline incidents in 6th and 7th grade were not predictive of changes in trust in 7th grade (*Bs* = -.06 and .01, *ps* > .166). And discipline incidents in 6th, 7th, or 8th grade were not predictive of 8th grade trust (*Bs* < .02, *ps* > .605). Not surprisingly, then, a model testing the reverse mediational sequence, in which 7th grade discipline incidents predicted 8th grade residual trust, yielded a confidence interval that included zero and was therefore not

supported, *Indirect effect* = .001, *Boot SE* = .01, *95% CI* = -.02 to .04. The pattern of data suggests one plausible account of the effect of affirmation on discipline is that affirmation helped forestall a deterioration of school trust during 7th grade, which in turn contributed to a reduction in 8th grade discipline incidents (see Figure 3 for standardized path estimates).

General Discussion

Evidence from the three-year, double-blind field experiment found that giving students opportunities to repeatedly affirm their core personal values during middle school forestalled declines in teacher trust and improved students' behavioral conduct. During 7th grade, discipline incidents spiked and school trust sagged for the sample as a whole. However, both of these outcomes were muted among students assigned to the affirmation condition. In addition, mediation analyses found that the changes in trust that occurred during 7th grade predicted students' discipline incidents in 8th grade. Thus Hypotheses 1-3 all found support. However, unlike previous affirmation studies (e.g., Cohen et al., 2006; Sherman et al., 2013), in this study there was no significant moderation of the affirmation effects by ethnicity. Thus Hypothesis 4 was not supported.

No Moderation by Student Ethnicity or Gender

The fact that demographic categories did not moderate the results runs counter to our expectation that groups who consistently experience psychological threat (e.g., negatively stereotyped ethnic groups) should show the largest benefits of affirmation (see Borman, 2017). However, we note that self-affirmation theory asserts that affirmation will be moderated, not by an objective or demographic variable, but by a *psychological state*. Namely, a felt threat to adequacy should moderate the effects of affirmation, and the degree to which different people experience this threat may vary by context and time (Steele, 2011). As the results displayed in

Table 5 suggest, at least in the present research the intervention had a generally beneficial effect across student subgroups on both school trust and incidents of discipline.

Such general effects of affirmation, while uncommon in the education literature, have several precedents in the literature. Self-affirmation theory holds that threats to self-integrity can come from many sources beyond ethnicity, gender, or other demographic categories (Sherman & Cohen, 2006; Steele, 2011), and self-affirmation exercises have been shown to alleviate the effects of threats to self-integrity regardless of their source (Cohen & Sherman, 2014). For example, affirmation benefits have been seen among White men undergraduates who reported a relatively low sense of belonging at school (Layous et al., 2017), among participants asked to give an impromptu speech (Creswell et al., 2005), among smokers exposed to anti-smoking messages (Crocker, Niiya, & Mischkowski, 2008), and among women who learned that their coffee-drinking habits put their health at risk (Sherman et al., 2000). Thus affirmation can help people contend with threats to self-integrity in general, not just threats tied to ethnicity or other stereotype-threatened identities.

Nevertheless, we conducted additional exploratory analyses to attempt to understand the lack of moderation by ethnicity. For example, we wondered whether other features of participants' backgrounds might have been important. Perhaps White and Asian students from lower socio-economic status backgrounds or students with lower prior performance responded more positively to the intervention. Post-hoc analyses did not support these ideas, however. We also tested whether just the Black participants responded differently than just the White participants, given that these were the two largest groups on campus. However, again results were largely the same as with the full sample. Thus, the results suggest that, at least in the present sample, the effect of affirmation on discipline and trust was not different across ethnic

groups. Perhaps the processes involved tapped a psychology that is largely shared among middle school students.

Trust and Ego-Risk

Many scholars have argued that one of the essential features of trust is that it involves vulnerability to the self (Rousseau et al., 1998). By trusting, people put their fate partially in the hands of others. In the present research we argued that students' trust in their teachers carries the expectation that teachers will not harm their self-integrity (Binning & Huo, 2012). When trust is confirmed, such as through treatment that is perceived as fair, dignified, and respectful, it validates the decision to trust and affirms people's sense of self-integrity (cf., Tyler & Lind, 1992). However, relational trust violations can have dramatic, precipitous consequences (Binning & Huo, 2012; Burt & Knez, 1996). We argue these consequences can catalyze a negative recursive process that makes discipline self-defeating and further deteriorations of trust more likely.

The present research identifies students' concerns for their global self-integrity as an important factor in the decision to trust or distrust teachers. That is, the present research found that by providing students repeated opportunities to affirm important personal values, the intervention may have helped students overcome the psychological barriers to trust, in particular the threat to self that is inherent in making oneself vulnerable to the potentially biased judgment and treatment of an authority. These affirmations focused students on their core values, for example, by reflecting on why their friends and family are important, and thus they were theorized to satisfy students' need to maintain their global self-integrity (Steele, 1988; Sherman & Cohen, 2006). The finding that this manipulation forestalled declines in trust suggests that those declines occurred in the control condition, in part, as a means to protect global self-

integrity. The results therefore support the idea that trust involves not just vulnerability, but ego risk. Self-affirmation seems to be one means to mitigate that risk and improve students' behavioral conduct over time.

Teachers as a Catalyst of Affirmation Processes

Although speculative, the potentially significant role of teachers in enacting the effects of self-affirmation writing exercises is worth considering. Through a series of relatively brief writing exercises administered in class, teachers provided students with a venue to express what was important to them. When the teacher delivered an intervention, it may have conveyed a message to students that the teacher cares about the student's values and what they find important. Such concern could foster a positive recursive cycle (Yeager et al., 2014). For example, students' personal disclosure of important personal values may strengthen the bond students feel with the teacher (Collins & Miller, 1994), which may cause students to trust the teacher more, engage more in the class and perceive higher belonging in their school environment, which have been shown to lead to academic success (Walton & Cohen, 2011; Yeager & Walton, 2011). If an outside researcher delivered the intervention, by contrast, there would be no opportunity for this social-relational process to flourish. Indeed, one replication attempt that used members of the research team to deliver the affirmation intervention failed to find effects of the intervention (Protzko & Aronson, 2016). Thus one possibility is that having teachers repeatedly ask students about their core values signals to students that they are valued. Additional research should explore this possibility.

Limitations

Although main effects of affirmation have occurred before, it remains unclear why the present study did not find moderation by demographic variables, particularly ethnicity, when

other field studies have found such moderation. Although such studies focused primarily on performance, not discipline, several distinctive features of the present study may have contributed to the effects found here. For example, the longitudinal, intensive nature of the affirmation protocol (with up to nine affirmations over three years) was well-designed for casting a wide net across students. Although one study found that minority students showed a rapid deterioration of trust during middle school, school trust among White students also declined over time, albeit less steeply (e.g., Yeager et al., 2017). The early start of the intervention at the beginning of 6th grade and intensity of the intervention (with 8 writing assignments during the first two years of the study) may have created more opportunities for the affirmation to bolster trust over time for all students.

Another limitation pertains to the nature of the discipline outcome. The simple count of disciplinary incidents leaves unclear what types of incidents students had. For example, students were counted as having an incident whether they displayed aggressive behavior (e.g., for fighting) or avoidant behavior (e.g., drug possession), leaving us unable to differentiate them. Research suggests different types of conduct problems are related to different peer and self-perceptions. Internalizing behaviors such as withdrawal and avoidance tend to be associated with lower social competence (Lansford et al. 2006), while externalizing behaviors are often related to impulse control issues (Hymel, Bowker, & Woody, 1993). Additional research is needed to determine which types of problem behaviors are affected by affirmation.

Another limitation is that the present study focused on just one cohort in one school, meaning we were unable to determine how different contextual factors (e.g., different percentages of racial minorities on campuses) may have impacted the intervention effects (cf. Borman et al., 2015). Factors that might affect school contexts and the genesis of distrust include

features of the school climate, such as whether ethnic group stereotypes are salient (Borman, Grigg, Rozek, Hanselman, & Dewey, 2018) or whether students feel a sense of *identity safety* on campus (Purdie-Vaughns et al., 2008; Wanless, 2016). Understanding how contextual factors impact affirmation interventions may be key for understanding the necessary conditions for effective affirmation interventions. Several research studies have recently failed to find expected benefits of affirmation interventions (e.g., de Jong, Jellesma, Koomen, & de Jong, 2016; Dee, 2015; Hanselman, Rozek, Grigg, & Borman, 2017; Protzko & Aronson, 2016), which points to a pressing need to understand the possibility of contextual moderators of affirmation interventions.

Finally, an important limitation is that the present work focused solely on the perspective of students. That is, unlike work that has targeted teachers' mindsets about discipline as a means to affect their disciplinary decisions (Okonofua et al., 2016), we focused on students and their experiences as the units of analyses. We did not assess if affirmation changed teachers' perceptions of their students (cf. Thomaes et al., 2012). However, the delivery of discipline is a two-way street, where teachers and administrators who themselves may vary in their trust levels of students frequently have discretion in determining when discipline incidents occur. The present study showed that a student-level affirmation can affect the rate at which teachers and administers discipline students, and research has shown that affirmation interventions can have ecological benefits, whereby benefits among students can spread to affect others who share their social context (Powers et al., 2016). Future research may benefit from examining perspectives of teachers and students simultaneously for a more complete picture of intervention effects.

Conclusion

This report found evidence that an early, sustained delivery of self-affirmation exercises during the transition to middle school can have a gradual, powerful impact on students' trust and

behavioral conduct over time. While no evidence of the intervention was seen in 6th grade, by 7th grade students' in the affirmation condition received fewer discipline incidents and maintained relatively high trust in their teachers. By 8th grade, affirmed students continued to receive fewer discipline incidents, a tendency that was predicted by the maintenance of trust during the prior year. Surprisingly, and in contrast to much prior research on self-affirmation interventions in education, the results were not moderated by participant ethnicity. Instead, participants benefited similarly from the affirmation, regardless of their ethnic group background. Repeated administrations of self-affirmation manipulations may help maintain students' school trust and yield long-term benefits for students and their schools.

References

Amemiya, J., Fine, A., & Wang, M. T. (2019). Trust and discipline: Adolescents' institutional and teacher trust predict classroom behavioral engagement following teacher discipline. *Child Development*. doi.org/10.1111/cdev.13233

Belmi, P., Barragan, R. C., Neale, M. A., & Cohen, G. L. (2015). Threats to social identity can trigger social deviance. *Personality and Social Psychology Bulletin, 41*, 467–484. doi.org/10.1177/0146167215569493

Bies, R. J., & Tripp, T. M. (1996). Beyond distrust: Getting even and the need for revenge. In R. M. Kramer & T. Tyler (Eds.), *Trust in organizations* (pp. 246–260). Thousand Oaks, CA: Sage. dx.doi.org/10.4135/9781452243610.n12

Binning, K. R., & Huo, Y. J. (2012). Understanding status as a social resource. In K. Tornblom & A. Kazemi (Eds.), *Handbook of social resource theory* (pp. 133–147). New York: Springer. doi.org/10.1007/978-1-4614-4175-5_8

Borman, G. D. (2017). Advancing values affirmation as a scalable strategy for mitigating identity threats and narrowing national achievement gaps. *Proceedings of the National Academy of Sciences, 114*, 7486–7488. doi.org/10.1073/pnas.1708813114

Borman, G. D., Grigg, J., & Hanselman, P. (2016). An effort to close achievement gaps at scale through self-affirmation. *Educational Evaluation and Policy Analysis, 38*, 21–42. doi.org/10.3102/0162373715581709

Borman, G. D., Grigg, J., Rozek, C. S., Hanselman, P., & Dewey, N. A. (2018). Self-affirmation effects are produced by school context, student engagement with the intervention, and time: Lessons from a district-wide implementation. *Psychological Science, 29*, 1773–1784. doi.org/10.1177/0956797618784016

Brady, S. T., Reeves, S. L., Garcia, J., Purdie-Vaughns, V., Cook, J. E., Taborsky-Barba, S., Tomasetti, S., Davis, E. M., & Cohen, G. L. (2016). The psychology of the affirmed learner: Spontaneous self-affirmation in the face of stress. *Journal of Educational Psychology*, *108*, 353–373. dx.doi.org.pitt.idm.oclc.org/10.1037/edu0000091

Burt, R. S., & Knez, M. (1996). Trust and third-party gossip. In R. M. Kramer & T. R. Tyler (Eds.), *Trust in organizations* (pp. 68–89). Thousand Oaks, CA: Sage. dx.doi.org/10.4135/9781452243610.n5

Cohen, G. L., Steele, C. M., & Ross, L. D. (1999). The mentor's dilemma: Providing critical feedback across the racial divide. *Personality and Social Psychology Bulletin*, *25*, 1302–1318. doi.org/10.1177/0146167299258011

Cohen, G. L., & Garcia, J. (2014). Educational theory, practice, and policy and the wisdom of social psychology. *Policy Insights from the Behavioral and Brain Sciences*, *1*, 13–20. doi.org/10.1177/2372732214551559

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science*, *313*, 1307–1310. 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*, 400–403. doi.org/10.1126/science.1170769

Cohen, G. L., & Sherman, D. K. (2014). The psychology of change: Self-affirmation and social psychological intervention. *Annual Review of Psychology*, *65*, 333–371. doi.org/10.1146/annurev-psych-010213-115137

Cohen, G. L., & Steele, C. M. (2002). A barrier of mistrust: How negative stereotypes affect cross-race mentoring. In J. Aronson, *Improving academic achievement* (pp. 303-327). Academic press. doi.org/10.1016/B978-012064455-1/50018-X

Collins, N. L., & Miller, L. C. (1994). Self-disclosure and liking: a meta-analytic review. *Psychological Bulletin, 116*, 457–475. dx.doi.org/10.1037/0033-2909.116.3.457

Creswell, J. D., Welch, W. T., Taylor, S. E., Sherman, D. K., Gruenewald, T. L., & Mann, T. (2005). Affirmation of personal values buffers neuroendocrine and psychological stress responses. *Psychological Science*, *16*, 846–851. doi.org/10.1111/j.1467-9280.2005.01624.x

Critcher, C. R., & Dunning, D. (2015). Self-affirmations provide a broader perspective on self-threat. *Personality and Social Psychology Bulletin*, *41*, 3–18. doi.org/10.1177/0146167214554956

Crocker, J., Niiya, Y., & Mischkowski, D. (2008). Why does writing about important values reduce defensiveness? Self-affirmation and the role of positive other-directed feelings. *Psychological Science*, *19*, 740–747. doi.org/10.1111/j.1467-9280.2008.02150.x

Cronbach, L. J., & Furby, L. (1970). How we should measure" change": Or should we? *Psychological Bulletin, 74*, 68–80. dx.doi.org/10.1037/h0029382

De Jong, E. M., Jellesma, F. C., Koomen, H. M., & De Jong, P. F. (2016). A values affirmation intervention does not benefit negatively stereotyped immigrant students in the Netherlands. *Frontiers in Psychology*, *13.* doi.org/10.3389/fpsyg.2016.00691

Dee, T. S. (2015). Social identity and achievement gaps: Evidence from an affirmation intervention. *Journal of Research on Educational Effectiveness*, *8*, 149–168. doi.org/10.1080/19345747.2014.906009

Dodge, K. A., & Somberg, D. R. (1987). Hostile attributional biases among aggressive boys are exacerbated under conditions of threats to the self. *Child Development, 58,* 213–224. doi.org/10.2307/1130303

Eccles, J. S., Midgley, C., Wigfield, A., Buchanan, C. M., Reuman, D., Flanagan, C., & Mac Iver, D. (1993). Development during adolescence: The impact of stage-environment fit on young adolescents' experiences in schools and in families. *American Psychologist, 48*, 90–101. dx.doi.org.pitt.idm.oclc.org/10.1037/0003-066X.48.2.90

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Ferrer, R. A., & Cohen, G. L. (2018). Reconceptualizing self-affirmation with the trigger and channel framework: Lessons from the health domain. *Personality and Social Psychology Review*, 1–20. doi.org/10.1177/1088868318797036

Gambetta, D. (1988). *Trust: Making and breaking cooperative relations*. Oxford: Basil Blackwell.

Goyer, J. P., Garcia, J., Purdie-Vaughns, V., Binning, K. R., Cook, J. E., Reeves, S. J., Apfel, N., Taborsky-Barba, S., Sherman, D. K., & Cohen, G. L. (2017). Self-affirmation facilitates minority middle schoolers' progress along college trajectories. *Proceedings of the National Academy of Sciences, 114*, 7594–7599. doi.org/10.1073/pnas.1617923114

Hanselman, P., Rozek, C. S., Grigg, J., & Borman, G. D. (2017). New evidence on self-affirmation effects and theorized sources of heterogeneity from large-scale replications. *Journal of Educational Psychology*, *109*, 405–424. dx.doi.org/10.1037/edu0000141

Harackiewicz, J. M., Tibbetts, Y., Canning, E., & Hyde, J. S. (2014). Harnessing values to promote motivation in education, *18*, 71–105. doi.org/10.1108/S0749-742320140000018002

Hayes, A. F. (2012). *PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling.* Retrieved from http://www.afhayes.com/public/ process2012.pd

Hirschfield, P. (2009). Another way out: The impact of juvenile arrests on high school dropout. *Sociology of Education, 82,* 368–393. doi.org/10.1177/003804070908200404

Huo, Y. J., Binning, K. R., & Molina, L. E. (2010). Testing an integrative model of respect: Implications for social engagement and well-being. *Personality and Social Psychology Bulletin*, *36*, 200–212. doi.org/10.1177/0146167209356787

Hymel, S., Bowker, A., & Woody, E. (1993). Aggressive versus withdrawn unpopular children: Variations in peer and self-perceptions in multiple domains. *Child Development, 64*, 879-896. doi.org/10.1111/j.1467-8624.1993.tb02949.x

Kinias, Z., & Sim, J. (2016). Facilitating women's success in business: Interrupting the process of stereotype threat through affirmation of personal values. *Journal of Applied Psychology*, *101*, 1585–1597. dx.doi.org.pitt.idm.oclc.org/10.1037/apl0000139

Lansford, J. E., Malone, P. S., Stevens, K. I., Dodge, K. A., Bates, J. E., & Pettit, G. S. (2006). Developmental trajectories of externalizing and internalizing behaviors: Factors underlying resilience in physically abused children. *Development and Psychopathology*, *18*, 35-55. doi.org/10.1017/S0954579406060032

Layous, K., Davis, E. M., Garcia, J., Purdie-Vaughns, V., Cook, J. E., & Cohen, G. L. (2017). Feeling left out, but affirmed: Protecting against the negative effects of low belonging in college. *Journal of Experimental Social Psychology, 69*, 227–231. doi.org/10.1016/j.jesp.2016.09.008

Lind, E. A., & Tyler, T. R. (1988). *The social psychology of procedural justice*. Berlin, Germany: Springer Science & Business Media. doi.org/10.1007/978-1-4899-2115-4

Lyubomirsky, S., & Della Porta, M. D. (2010). Boosting happiness, buttressing resilience. In J. W. Reich, A. J. Zautra, & J. Hall (Eds.), *Handbook of adult resilience: Concepts, methods, and applications* (pp. 450–464). New York: Guilford.

Major, B., Spencer, S., Schmader, T., Wolfe, C., & Crocker, J. (1998). Coping with negative stereotypes about intellectual performance: The role of psychological disengagement. *Personality and Social Psychology Bulletin*, *24*, 34–50. doi.org/10.1177/0146167298241003

Martens, A., Johns, M., Greenberg, J., & Schimel, J. (2006). Combating stereotype threat: The effect of self-affirmation on women's intellectual performance. *Journal of Experimental Social Psychology*, *42*, 236–243. doi.org/10.1016/j.jesp.2005.04.010

McIntosh, K., Girvan, E. J., Horner, R. H., & Smolkowski, K. (2014). Education not incarceration: A conceptual model for reducing racial and ethnic disproportionality in school discipline. *Journal of Applied Research on Children: Informing Policy for Children at Risk, 5*, 1–22.

McQueen, A., & Klein, W. M. P. (2006). Experimental manipulations of self-affirmation: A systematic review. *Self and Identity*, *5*, 289–354. doi.org/10.1080/15298860600805325

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*, 1234–1237. doi.org/10.1126/science.1195996

Noguera, P. A. (2003). The trouble with black boys: The role and influence of environmental and cultural factors on the academic performance of African American males. *Urban Education, 38,* 431–459. doi.org/10.1177/0042085903038004005

Okonofua, J. A., Paunesku, D., & Walton, G. M. (2016). Brief intervention to encourage empathic discipline cuts suspension rates in half among adolescents. *Proceedings of the National Academy of Sciences, 113,* 5221–5226. doi.org/10.1073/pnas.1523698113

Okonofua, J. A., Walton, G. M., & Eberhardt, J. L. (2016). A vicious cycle: A social–psychological account of extreme racial disparities in school discipline. *Perspectives on Psychological Science, 11,* 381–398. doi.org/10.1177/1745691616635592

Pettit, B., & Western, B. (2004). Mass imprisonment and the life course: Race and class inequality in U.S. incarceration. *American Sociological Review, 69,* 151–169. doi.org/10.1177/000312240406900201

Powers, J. T., Cook, J. E., Purdie-Vaughns, V., Garcia, J., Apfel, N., & Cohen, G. L. (2016). Changing environments by changing individuals: The emergent effects of psychological intervention. *Psychological Science, 27,* 150–160. doi.org/10.1177/0956797615614591

Protzko, J., & Aronson, J. (2016). Context moderates affirmation effects on the ethnic achievement gap. *Social Psychological and Personality Science, 7,* 500–507. doi.org/10.1177/1948550616646426

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA: Sage.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review, 23,* 393–404. doi.org/10.5465/amr.1998.926617

Schmader, T., & Sedikides, C. (2018). State authenticity as fit to environment: The implications of social identity for fit, authenticity, and self-segregation. *Personality and Social Psychology Review*, *22*, 228–259. doi.org/10.1177/1088868317734080

Sherman, D. A., Nelson, L. D., & Steele, C. M. (2000). Do messages about health risks threaten the self? Increasing the acceptance of threatening health messages via self-affirmation. *Personality and Social Psychology Bulletin, 26*, 1046–1058. doi.org/10.1177/01461672002611003

Sherman, D. K., Hartson, K. A., Binning, K. R., Purdie-Vaughns, V., Garcia, J., Taborsky-Barba, S., Tomassetti, S., Nussbaum, A. D., & Cohen, G. L. (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,* 591-618. dx.doi.org/10.1037/a0031495

Sherman, D. K., Lokhande, M., Müller, T., Cohen, G. L. (in press). Self-affirmation interventions. In G. Walton & A. Crum, *Handbook of wise interventions: How social-psychological insights can help solve problems.*

Sherman, D. K., & Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 38, pp. 183–242). San Diego, CA: Academic Press. doi.org/10.1016/S0065-2601(06)38004-5

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,* 745–764. dx.doi.org/10.1037/a0015451

Skiba, R. J., Michael, R. S., Nardo, A. C., & Peterson, R. L. (2002). The color of discipline: Sources of racial and gender disproportionality in school punishment. *The Urban Review, 34*, 317–342. doi.org/10.1023/A:1021320817372

Steele, C. M. (2011). *Whistling Vivaldi: How stereotypes affect us and what we can do (issues of our time)*. New York: W. W. Norton.

Steele, C. M. (1988). The psychology of self-affirmation: sustaining the integrity of the self. *Advances in Experimental Social Psychology*, *21*, 261–302. doi.org/10.1016/S0065-2601(08)60229-4

Steele, C. M. (1992). Race and the schooling of Black Americans. *The Atlantic Monthly, 269*, 68–78.

Thomaes, S., Bushman, B. J., Castro, B. O. de, Cohen, G. L., & Denissen, J. J. (2009). Reducing narcissistic aggression by buttressing self-esteem: An experimental field study. *Psychological Science*, *20*, 1536–1542. doi.org/10.1111/j.1467-9280.2009.02478.x

Thomaes, S., Bushman, B. J., Orobio de Castro, B., & Reijntjes, A. (2012). Arousing "gentle passions" in young adolescents: Sustained experimental effects of value affirmations on prosocial feelings and behaviors. *Developmental Psychology*, *48*, 103–110. dx.doi.org.pitt.idm.oclc.org/10.1037/a0025677

Tibbetts, Y., Harackiewicz, J. M., Canning, E. A., Boston, J. S., Priniski, S. J., & Hyde, J. S. (2016). Affirming independence: Exploring mechanisms underlying a values affirmation intervention for first-generation students. *Journal of Personality and Social Psychology*, *110*, 635–659. dx.doi.org/10.1037/pspa00000

Twenge, J. M., Baumeister, R. F., Tice, D. M., & Stucke, T. S. (2001). If you can't join them, beat them: effects of social exclusion on aggressive behavior. *Journal of Personality and Social Psychology, 81*, 1058–1069. dx.doi.org/10.1037/0022-3514.81.6.1058

Tyler, T. R., & Blader, S. L. (2005). Can businesses effectively regulate employee conduct? The antecedents of rule following in work settings. *The Academy of Management Journal, 48*, 1143–1158. doi.org/10.5465/amj.2005.19573114

Tyler, T. R., & Degoey, P. (1996). Trust in organizational authorities. In R. M. Kramer & T. R. Tyler, *Trust in organizations: Frontiers of theory and research* (pp. 331-356). Thousand Oaks, CA: Sage.

Tyler, T. R., & Lind, E. A. (1992). A relational model of authority in groups. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 25, pp. 115–191). San Diego, CA: Academic Press. doi.org/10.1016/S0065-2601(08)60283-X

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science, 331*, 1447–1451. DOI: 10.1126/science.1198364

Wanless, S. B. (2016). The role of psychological safety in human development. *Research in Human Development, 13*, 6-14. doi.org/10.1080/15427609.2016.1141283

Way, S. M. (2011). School discipline and disruptive classroom behavior: The moderating effects of student perceptions. *The Sociological Quarterly*, *52*, 346–375. doi.org/10.1111/j.1533-8525.2011.01210.x

Yeager, D. S., Purdie-Vaughns, V., Garcia, J., Apfel, N., Brzustoski, P., Master, A., Hessert, W. T., Williams, M. E., & Cohen, G. L. (2014). Breaking the cycle of mistrust: Wise

interventions to provide critical feedback across the racial divide. *Journal of Experimental Psychology: General*, *143*, 804–824. doi: 10.1037/a0033906

Yeager, D. S., Purdie-Vaughns, V., Hooper, S. Y., & Cohen, G. L. (2017). Loss of institutional trust among racial and ethnic minority adolescents: A consequence of procedural injustice and a cause of life-span outcomes. *Child Development*, *88*, 658-676. doi.org/10.1111/cdev.12697

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

Table 1. Sample sizes after attrition for the full sample and focal subgroups for all three years of the longitudinal design.

	Year 1	Year 2	Year 3
Total	163	153	145
Black or Latinx/Asian or White	75/88	71/82	67/78
Female/Male	82/81	77/76	73/72

Table 2. Schedule of study activities across all three years of the longitudinal design.

Intervention Schedule									
	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May
Year 1 (6th grade)	1, a	a			b		c		2
Year 2 (7th grade)	1, a	a			b		c		2
Year 3 (8th grade)					1		d		2

a. Affirmation: Explain why values you selected are important to you. Control: Explain why other people might value your least important values.
b. Affirmation: Write about something that is important to you. Control: Write about your typical morning routine.
c. Affirmation: Write about a personalized value that was selected based on your past responses. Control: Write about your typical afternoon routine.
d. Half of the affirmed participants were assigned to receive the affirmation from (a), all other participants received the control condition from (a).

1. Beginning of year climate survey.
2. End of year climate survey.

Table 3. Tests of baseline equivalence between control and affirmation condition. Categorical variables (gender and ethnicity) were compared using a chi-square tests, and continuous variables (pre-intervention trust, discipline, and performance) were compared using t-tests.

	Control	**Affirmation**	X^2 **or** t	p
Male	50%	49%	0.01	.937
Black or Latinx	46%	46%	0.01	.932
Pre-Int. School Trust	4.99	4.93	0.41	.680
Pre-Int. Discipline Incidents	0.05	0.07	-0.67	.504
Prior Performance (Z-scores)	-0.02	0.04	-0.45	.652

Table 4. Raw means for disciplinary incidents, trust in teachers, and residualized trust scores, in which higher scores indicate higher than expected trust at the end of the school year. Results are broken down as a function of Ethnic group × Condition and Gender × Condition, however, neither ethnic group nor gender significantly moderated the effect of condition, which was most pronounced in 7th and 8th grade.

Discipline incidents										
	White/Asian		Black/Latinx		Females		Males			
	Control	Affirmation	Control	Affirmation	Control	Affirmation	Control	Affirmation	Grand *M*	Grand *SD*
6th grade	0.28	0.03	0.62	0.39	0.29	0.15	0.29	0.40	0.28	0.45
7th grade	0.24	0.08	0.50	0.46	0.42	0.18	0.30	0.33	0.31	0.46
8th grade	0.28	0.03	0.62	0.39	0.44	0.16	0.43	0.23	0.32	0.47
Trust in teachers										
	White/Asian		Black/Latinx		Females		Males			
	Control	Affirmation	Control	Affirmation	Control	Affirmation	Control	Affirmation	Grand *M*	Grand *SD*
Pre-Intervention	4.99	5.13	4.98	4.69	4.93	5.04	5.04	4.82	4.96	0.85
End 6th grade	4.90	4.98	4.29	4.27	4.58	4.92	4.63	4.41	4.64	0.92
Beginning 7th grade	5.04	4.99	4.46	4.31	4.73	4.76	4.81	4.60	4.73	0.87
End 7th Grade	4.66	4.92	3.86	4.05	4.31	4.76	4.24	4.32	4.41	0.96
Beginning 8th Grade	4.72	4.70	3.98	4.28	4.36	4.64	4.40	4.37	4.44	0.92
End 8th Grade	4.62	4.68	3.97	4.44	4.13	4.71	4.51	4.42	4.45	0.91
Residualized trust scores										
	White/Asian		Black/Latinx		Females		Males			
	Control	Affirmation	Control	Affirmation	Control	Affirmation	Control	Affirmation	Grand *M*	Grand *SD*
6th grade	0.26	0.26	-0.35	-0.25	-0.02	0.22	-0.05	-0.16	0.00	0.81
7th grade	0.01	0.33	-0.35	-0.05	-0.12	0.30	-0.21	0.04	0.00	0.72
8th grade	-0.05	0.03	-0.12	0.14	-0.26	0.10	0.11	0.05	0.00	0.60

Table 5. Final multi-level models used to estimate trust and discipline over time. Coefficients (and standard errors) in bold indicate significance at $p < .05$. Level 2 predictor variables were grand mean centered except for condition, which was coded (0 = *Control*; 1 = *Treatment*) for ease of interpretation.

	Trust in teachers	**Discipline incidents**
Number of measurements	6	3
Unconditional Models		
Linear	**-0.23 (.05)**	**0.64 (.16)**
Quadratic	**0.02 (.01)**	**-0.22 (.07)**
Final Models		
Outcome: Initial status		
Intercept	**4.94 (.09)**	0.13 (.22)
Female	0.15 (.13)	**-0.55 (.28)**
Treatment	-0.11(.12)	-0.28 (.27)
Pre-Discipline	0.45 (.24)	**1.55 (.56)**
White and Asian	**0.45 (.13)**	**-1.55 (.55)**
Outcome: Linear change		
Intercept	**-0.27 (.05)**	**1.02 (.20)**
Female	**-0.04 (.03)**	**0.29 (.11)**
Treatment	**0.08 (.03)**	**-0.44 (.11)**
Pre-Discipline	-0.07 (.07)	-0.03 (.16)
White and Asian	-0.01 (.03)	**0.29 (.11)**
Outcome: Quadratic change		
Intercept	**0.02 (.01)**	**-0.25 (.08)**

Table 6. Pearson correlations between yearly discipline incidents and residualized changes in school trust.

		1.	*2.*	*3.*	*4.*	*5.*	*6.*
1.	6th Grade Discipline	1.00					
2.	7th Grade Discipline	.54**	1.00				
3.	8th Grade Discipline	.56**	.66**	1.00			
4.	6th Grade Change in Trust	-.41**	-.41**	-.37**	1.00		
5.	7th Grade Change in Trust	-.44**	-.21*	-.33**	.35**	1.00	
6.	8th Grade Change in Trust	0.09	-0.02	-0.03	0.03	0.03	1.00

Figure 1. Residual school trust during 7th grade, representing the change in trust from the beginning to the end of the school year. Results show a main effect for affirmation condition, and a main effect for ethnicity, but no Ethnicity × Condition interaction. Error bars represent +/- 1 standard errors.

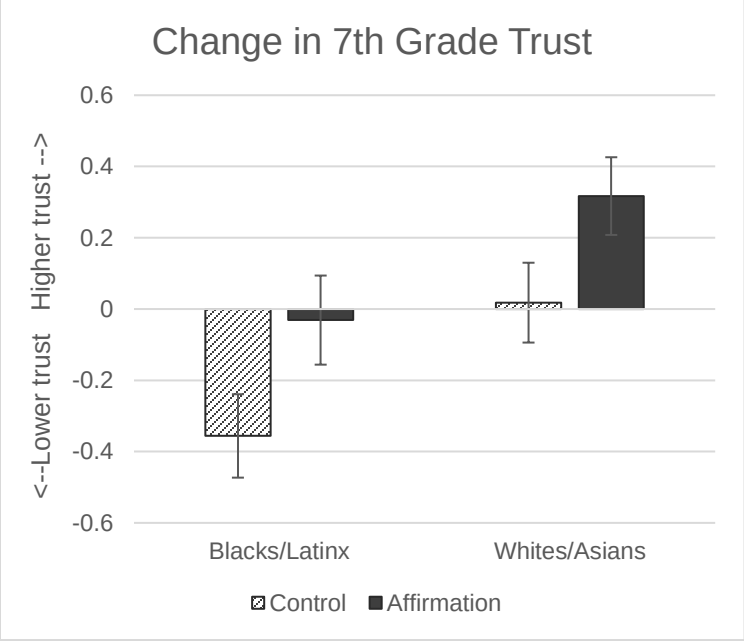
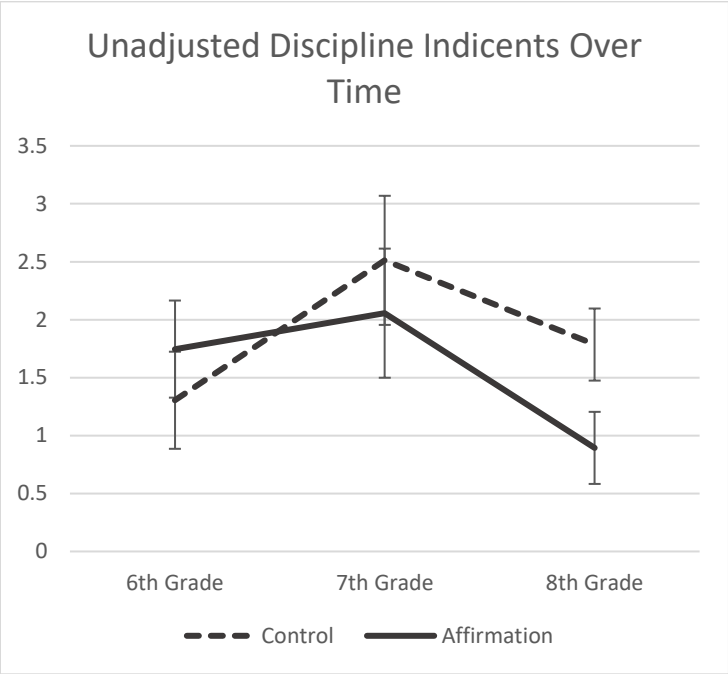


Figure 2. Unadjusted discipline incidents (top panel) and covariate-adjusted discipline incidents (bottom panel) over time as a function of condition, imputed from estimates of Poisson models with over-dispersion. Covariate-adjusted estimates control for pre-intervention discipline, gender, and ethnicity. Error bars represent +/- 1 standard errors.



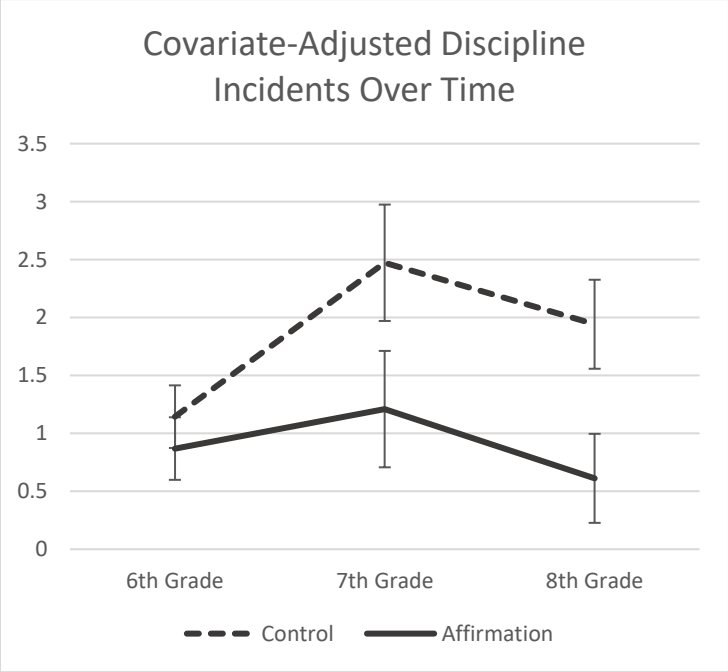


Figure 3. Standardized regression coefficients for mediation model testing whether the affirmation manipulation affected discipline incidents in 8th grade via a change in trust in 7th grade. Mediation model controls for pre-intervention discipline, gender, and ethnicity.