sarial growth would be reported most frequently among individuals exhibiting the classic recovery trajectory (i.e., people who struggle with elevated distress and disruption in functioning for several months or longer and then only gradually return to their pre-event baseline levels of functioning).

Perhaps most important, to the best of my knowledge reliable data to support the idea that trauma induces movement toward optimal functioning are not actually available. As far as I know, virtually all the available research on adversarial growth consists of retrospective self-reports obtained using cross-sectional designs. In the absence of more objective and prospective data, one could just as easily argue that reports of adversarial growth are simply the result of retrospective reattribution for the pain caused by the recovery process (e.g., "I am better now, so I must have grown"). Convincing evidence of authentic (rather than assumed or imagined) trauma-induced growth and movement toward optimal functioning will require actual pre-event data on functioning and objective demonstration that levels of health and well-being had improved after a designated traumatic event.

Ironically, the only data of this nature that I know of come from bereavement studies, including those that my colleagues and I have conducted. These studies used prospective data to identify a relatively small (approximately 10%) subset of participants who exhibited markedly improved functioning following the death of their spouse (Bonanno et al., 2002) or life partner (Bonanno, Rennicke, & Dekel, in press). Should these individuals be seen as moving toward optimal functioning? I would argue against such a conclusion; in each study the improved individuals had been highly depressed prior to their partner's death and then improved into the normal range of functioning. There was no evidence that they experienced optimal functioning, and there was some indication that over longer periods of time they eventually experienced renewed elevations in distress (Boerner, Wortman & Bonanno, in press).

Fourth, Litz (2005) and Roisman (2005) argued that I had overstated the proposition that resilience is common following extremely aversive events. They suggested instead that resilience should be markedly less prevalent following more extreme types of PTEs. I certainly agree that the proportion of individuals exhibiting a resilient outcome will tend to fluctuate across different types and durations of stressor events. Nonetheless, I hold firm to the prediction that resilience will almost always be the modal outcome among adults exposed to even the most pernicious of stressor events. Our most recent studies have borne this out in the context of such extreme events as the death of a life partner to AIDS (Bonanno, Moscowitz, et al., in press) and high levels of exposure to terrorist attack (Bonanno, Rennicke, & Dekel, in press). These are hardly a "trivial form of resilience" (Roisman, 2005, p. 264). This issue can only be truly resolved, of course, after researchers have amassed a large body of data on the relative prevalence of resilience across a range of aversive events. I firmly believe that such data, when they become available, will continue to robustly document the natural human ability to thrive under even the most adverse circumstances.

REFERENCES

Block, J. H., & Block, J. (1980). The role of ego-control and ego-resiliency in the organization of behavior. In W. A. Collins (Ed.), *The Minnesota Symposia on Child Psychology* (Vol. 13, pp. 39–101). Hillsdale, NJ: Erlbaum.

Boerner, K., Wortman, C. B., & Bonanno, G. A. (in press). Resilient or at risk?: A four-year study of older adults who initially showed high or low distress following conjugal loss. *Journal of Gerontology: Psychological Science.*

Bonanno, G. A. (2004). Loss, trauma, and human resilience: Have we underestimated the human capacity to thrive after extremely aversive events? *American Psychologist, 59,* 20–28.

Bonanno, G. A. (in press). Adult resilience to potential trauma. *Current Directions in Psychological Science.*

Bonanno, G. A., Field, N. P., Kovacevic, A., & Kaltman, S. (2002). Self-enhancement as a buffer against extreme adversity: Civil war in Bosnia and traumatic loss in the United States. *Personality and Social Psychology Bulletin, 28,* 184–196.

Bonanno, G. A., Moskowitz, J. T., Papa, A., & Folkman, S. (in press). Resilience to loss in bereaved spouses, bereaved parents, and bereaved gay men. *Journal of Personality and Social Psychology.*

Bonanno, G. A., Papa, A., LaLande, K., Westphal, M., & Coifman, K. (2004). The importance of being flexible: The ability to both enhance and suppress emotional expression predicts long-term adjustment. *Psychological Science, 15,* 482–487.

Bonanno, G. A., Rennicke, C., & Dekel, S. (in press). Self-enhancement among high-exposure survivors of the September 11th terrorist attack: Resilience or social maladjustment? *Journal of Personality and Social Psychology.*

Bonanno, G. A., Wortman, C. B., Lehman, D. R., Tweed, R. G., Haring, M., Sonnega, J., et al. (2002). Resilience to loss and chronic grief: A prospective study from pre-loss to 18 months post-loss. *Journal of Personality and Social Psychology, 83,* 1150–1164.

Garmezy, N. (1991). Resilience and vulnerability to adverse developmental outcomes associated with poverty. *American Behavioral Scientist, 34,* 416–430.

Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J., & Wheatley, T. (1998). Immune neglect: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology, 75,* 617–638.

John, O. P., & Robins, R. W. (1994). Accuracy and bias in self-perception: Individual differences in self-enhancement and role of narcissism. *Journal of Personality and Social Psychology, 66,* 206–219.

Kelley, T. M. (2005). Natural resilience and innate mental health. *American Psychologist, 60,* 265.

Linley, P. A., & Joseph, S. (2005). The human capacity for growth through adversity. *American Psychologist, 60,* 262–263.

Litz, B. T. (2005). Has resilience to severe trauma been underestimated? *American Psychologist, 60,* 262.

Luthar, S. (in press). Resilient adaptation: A synthesis of evidence spanning five decades. In D. Cicchetti & D. J. Cohen (Eds.), *Developmental psychopathology: Risk, disorder, and adaptation.* New York: Wiley.

Maddi, S. R. (2005). On hardiness and other pathways to resilience. *American Psychologist, 60,* 261–262.

Roisman, G. I. (2005). Conceptual clarifications in the study of resilience. *American Psychologist, 60,* 264–265.

Correspondence concerning this comment should be addressed to George A. Bonanno, Teachers College, Columbia University, 525 West 120th Street, Box 218, New York, NY 10027. E-mail: gab38@columbia.edu

DOI: 10.1037/0003-066X.60.3.267

Stereotype Threat Research and the Assumptions Underlying Analysis of Covariance

Jelte M. Wicherts
University of Amsterdam

Recently, Sackett, Hardison, and Cullen (January 2004) discussed the role of covariates in Steele and Aronson's (1995) seminal research on the effects of stereotype threat on scores of African American test takers. Besides highlighting some common misinterpretations that stem from the use of covariance-adjusted means in reporting Steele and Aronson's (Study 2) experimental results, Sackett et al. argued that these results indicate that Black–White test-score difference within the no-stereotype threat (i.e., nondiagnostic) condition actu-

ally reflects the test score difference on the SAT (i.e., the covariate). This implies that stereotype threat effects add to the often-found Black–White test score gap instead of partly accounting for it (Sackett et al., 2004). Here I comment on the use of analysis of covariance (ANCOVA) in stereotype threat (ST) experiments, because ST theory implies violations of the assumptions underlying ANCOVA. Such violations could result in incorrect Type I error rates and distortions in the adjustment of means. Because of this, ANCOVA appears inappropriate for analyzing (quasi-) experimental results of ST research. In addition, the interpretation proposed by Sackett et al. of Steele and Aronson's results may be due to distortions of mean adjustments caused by violations of model assumptions.

While avoiding technical detail, I will provide the assumptions underlying ANCOVA and discuss why these assumptions do not sit well with several aspects of ST theory. Besides the usual analysis of variance assumptions, the assumptions underlying ANCOVA are as follows (e.g., Wildt & Ahtola, 1978): (a) The relationship of the dependent variable and the covariate is linear, (b) the regression weights of the dependent variable on the covariate are equal for all design cells (i.e., regression weight homogeneity), (c) the variance of residuals is equal over cells (i.e., homogeneity of residual variance), and (d) the covariate is measured without error and is independent of the experimental manipulation. For theoretical reasons, the tenability of these assumptions within ST experiments is at least questionable.

In a typical ST experiment (e.g., Steele & Aronson, 1995, Study 2), the effects of an ST manipulation (e.g., nondiagnostic vs. diagnostic condition) on the test scores (i.e., dependent variable) of two groups (e.g., Blacks and Whites) are investigated. If a covariate (e.g., SAT scores) is used to adjust the dependent variable for preexisting group differences, a (2×2) ANCOVA appears suitable. However, the tenability of assumptions underlying this analysis appears unlikely, especially when one compares the ST cell (i.e., stigmatized group, diagnostic condition) with the other cells in the design.

Stereotype threat theory states that ST effects particularly influence test scores of people for whom the ability of interest is important or self-relevant (Steele, 1997). It is likely that within each cell there are individual differences in domain identification. Therefore, the manipulation triggering ST would result not only in mean effects (i.e., ST effects identical for each subject) but also in (co)variance effects (i.e., ST effects differing for subjects) on the dependent variable. Furthermore, if one supposes the presence of a positive correlation between (latent) ability and domain identification (see Steele, 1997, p. 617), this would result in an interaction between the covariate (i.e., ability as measured by the SAT) and the experimental manipulation (i.e., ST effects on the dependent variable). Higher SAT scores would imply higher domain identification and therefore stronger ST effects. This would result not only in a curvilinear relation between covariate and dependent variable in the affected cell (i.e., ST condition) but also in differences in regression weights over the cells. Admittedly, most ST research has used homogeneous samples, but even if individual differences in domain identification within cells are absent, there are other reasons to expect a violation of homogeneity of regression weights.

If mediators such as heightened anxiety or lowered motivation are the causes of lowered test scores within ST conditions, it is likely that these mediators will also affect the regression of the dependent variable on the covariate. Again, such mediators can result in a violation of homogeneity of regression weights. In addition, added mediator variance (e.g., anxiety variance) could result in differences in error variances between design cells, which would violate the homogeneity of variance assumption.

Finally, the assumption that the covariate is error free seems to be untenable because such measures are not perfectly reliable. Error in the covariate lowers the precision of the analysis. More important, the covariates themselves (e.g., the SAT) are possibly affected by ST. It could be argued that for the high-ability participants involved in most ST studies, the SAT is fairly easy and hence would not be stereotype threatening (Spencer, Steele, & Quinn, 1999). However, there are several reasons (e.g., the SAT is by definition self-relevant and diagnostic of ability) to expect that the SAT scores are affected by ST. Either way, from a theoretical point of view, use of a covariate that may already be affected by the phenomenon under investigation is potentially tautological. Technically, if the covariate is affected by ST, then this implies that the covariate and the manipulation are not independent, which may obscure the effects of the manipulation or even produce effects that are spurious (Wildt & Ahtola, 1978, p. 90).

In conclusion, ST theory *explicitly* predicts violations of practically all assumptions underlying ANCOVA. Therefore, ANCOVA appears to be unsuitable for investigating ST effects in quasi-experimental settings. At the very least, the use of ANCOVA should be accompanied by information concerning the tenability of model assumptions. Unfortunately, such information is not provided in the articles that have used covariates in investigating ST (e.g., Gonzales, Blanton, & Williams, 2002; Steele & Aronson, 1995). Testing ANCOVA assumptions is possible with most software packages and can provide important information on experimental effects. If model assumptions are not met, observed Type I errors are not at the supposed (nominal) significance level (i.e., $\alpha \neq .05$). More important, differences in regression weights will result in bias in adjusted means reported in several articles. This could distort the interpretation of Sackett et al. (2004) that, absent stereotype threat, Black–White differences are of the same degree as Black–White differences found on the SAT.

Although one could argue that ANCOVA is robust against violations of model assumptions, adjusted means should be interpreted with caution. I should stress that my critique is not particularly aimed at Steele and Aronson's 1995 article, because several points raised have only recently been investigated in the literature. My main concern is that the consequences of ST theory render ANCOVA unsuitable, and yet ANCOVA is still used quite often (e.g., Gonzales et al., 2002; Keller, 2002). In light of ST theory's emphasis on individual differences, it seems unlikely that ST only affects the means of the dependent variable (i.e., effects are identical for each subject within a cell) and leaves the covariance structure unaffected. Therefore, measurement models in which such effects are explicitly modeled (e.g., structural equation modeling) appear more suitable in analyzing ST effects.

REFERENCES

Gonzales, P. M., Blanton, H., & Williams, K. J. (2002). The effects of stereotype threat and double-minority status on the test performance of Latino women. *Personality and Social Psychology Bulletin, 28,* 659–670.

Keller, J. (2002). Blatant stereotype threat and women's math performance: Self-handicapping as a strategic means to cope with obtrusive negative performance expectations. *Sex Roles, 47,* 193–198.

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for African American–White differences on cognitive tests. *American Psychologist, 59,* 7–13.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35,* 4–28.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613–629.
Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797–811.
Wildt, A. R., & Ahtola, O. (1978). *Analysis of covariance.* Thousand Oaks, CA: Sage.

Correspondence concerning this comment should be addressed to Jelte M. Wicherts, Psychological Methods, Department of Psychology, University of Amsterdam, Roetersstraat 15, 1018 WB, Amsterdam, the Netherlands. E-mail: j.m.wicherts@uva.nl

DOI: 10.1037/0003-066X.60.3.269

Stereotype Threat Might Explain the Black–White Test-Score Difference

Janet E. Helms
Boston College

In their correction of the ostensibly widespread misinterpretation of Steele and Aronson's (1995) seminal study of the effects of stereotype threat on intellectual test scores, Sackett, Hardison, and Cullen (January 2004) expressed dismay and puzzlement that so many erudite people consistently have gone so far astray in their understanding of this matter. The gist of Sackett et al.'s correction was that interpreters of Steele and Aronson's (1995) results have ignored the researchers' statistical adjustment of their dependent measure for SAT scores and, consequently, have wrongly concluded that racial-group (i.e., Black–White) differences in test scores disappear when stereotype threat is removed.

In their justification for this much-needed clarification, Sackett et al. (2004) implied that the interpretation that stereotype threat explains the Black–White test-score disparity is not plausible. Yet whether or not the construct of stereotype threat generally can account for the Black–White test-score disparity was not the question that was directly addressed by either Steele and Aronson's (1995) original study or Sackett et al.'s (2004) critique of it.

It appears that Sackett et al. (2004), as well as the many people responsible for the allegedly faulty interpretation, essentially want an answer to the question, What *causes* or explains racial-group difference(s) in Black–White test scores? This question logically flows from (a) reviews demonstrating the chronic resistance of these differences to psychoeducational interventions (e.g., Sackett, Schmitt, Ellingson, & Kabin, 2001), (b) general recognition that racial-group membership cannot cause behavior (e.g., differences in test scores), as well as (c) acknowledgment that use of test scores for high-stakes decision making under prevailing circumstances amounts to "racial profiling" condoned by society and the law (Schmitt, Sackett, & Ellingson, 2002, p. 305). Therefore, if stereotype threat or analogous race or culture-related psychological constructs could be shown to account for the Black–White test-score disparity, then society would be relieved of the burden of unfair testing practices, and Sackett et al. (2004) would be relieved of the burden of "heading off future interpretive errors" (p. 11) regarding Steele and Aronson's (1995) results.

Fortunately, implied in Sackett et al.'s (2004) critique is the skeleton of a methodology for addressing the question of whether stereotype threat can account for between-groups differences in test scores. Accordingly, they argued that "if group differences in scores on the SAT and other tests were largely explainable by the mind-set [e.g., stereotype threat] with which examinees approach the testing situation, it would then follow that differences in factors such as quality of instruction or per-pupil educational expenditure do not matter much in terms of achievement in the domains measured by high-stakes tests" (Sackett et al., 2004, p. 11).

Yet "mind-sets" are well within the domain of behaviors that psychologists intend to measure, manipulate, or predict, whereas racial groups are not. Consequently, if stereotype-threat mind-set could replace racial group as the explanatory mechanism for the test-score disparity, then psychologists, educators, and other test users could utilize theory-driven interventions for increasing the test scores of Black test takers rather than use the arguably ineffectual approaches that Sackett et al. (2004) wanted to preserve but that Sackett et al. (2001) devalued. Perhaps the misguided interpreters of Steele and Aronson's (1995) study are inspired by this prospect.

There are two reasons why it was not possible to determine whether stereotype threat accounted for the variance attributable to racial group in Steele and Aronson's (1995) original study. The reasons are that the researchers (a) used self-reported SAT scores as covariates in their design (Steele & Aronson, 1995, p. 799), and (b) as seems to be typical of most subsequent race-related studies in this genre, they did not report enough descriptive information to permit computation of effect sizes of their findings.

With respect to the first reason, it is likely that the same race-related "mind-set" that triggered stereotype threat in Steele and Aronson's (1995) laboratory settings also triggered it when a telephone stranger asked study participants to recall their SAT scores. Thus, the SAT covariate in the original study may have been contaminated by stereotype threat, thereby contributing to incorrect adjustments of the dependent variable. Therefore, it is not clear whether it is Sackett et al.'s (2004) illustration of the "misinterpretation" of Steele and Aronson's results (Figure 1B, p. 9) or the "correct" interpretation of their results (Figure 1C, p. 9) that is truly accurate.

Although effect sizes could not be calculated from Steele and Aronson's (1995) summaries of their data, it was possible to reanalyze McKay, Doverspike, Bowen-Hilton, and McKay's (2003) descriptive summary data to show that the idea that stereotype threat might account for Black–White test-score disparities is not far-fetched. For stereotype threat to account for racial-group differences in test scores, measures or manipulations of it only have to account for at least as much variance in racial group as racial group explains in test scores. That is, racial group and stereotype threat should be redundant in a given testing situation. If such is the case, the researcher will have demonstrated that stereotype threat explained (i.e., mediated) the putative effects on test scores of racial group (Baron & Kenny, 1986).

The descriptive statistics summarized in Table 1 may be used to illustrate this principle of mediation. The correlation between stereotype threat and racial group ($r = -.56$) exceeded the correlation between racial-group and performance on a test of cognitive abilities ($r = .31$). Thus, when I conducted a hierarchical regression analysis in which stereotype threat and racial group were successively entered in Steps 1 and 2 of the analysis predicting Black–White test scores, stereotype threat accounted for a significant amount of variance in test scores ($R^2 = .10$), $F(1, 85) = 9.7$, $p < .003$ ($d = .66$), but racial group did not ($R^2_\Delta = .03$), $F(1, 84) = 2.51$, $p < .12$ ($d = .35$). Thus, stereotype explained the between-groups difference in McKay et al.'s (2003) study, which was approximately two thirds of a standard deviation.

The results would have been essentially reversed if I had reversed the order of entry of variables, although stereotype threat still accounted for slightly more variance in the second step than racial group did. Perhaps it is obvious, but the reason the first order of entry is preferable is that one wants to replace racial group, a con-