Running Head: IMPACT OF FACE VALIDITY ON GENDER DIFFERENCES

How Far Does Stereotype Threat Reach?

The Potential Detriment of Face Validity in Cognitive Ability Testing

James A. Grand, Ann Marie Ryan, Neal Schmitt, and Jillian Hmurovic

Michigan State University

Citation:
Grand, J.A., Ryan, A.M., Schmitt, N., & Hmurovic, J. (2010). How far does stereotype threat reach? The potential detriment of face validity in cognitive ability testing. *Human Performance, 24*, 1-28.

Abstract

A common belief is that improvements in a test's face validity have a positive impact on the performance and perceptions of test takers. However, the stereotype threat literature suggests adding job relevant context to tests could negatively impact the performance of women if that context is traditionally male-stereotyped. 345 participants (*n* = 236 females) completed either a face valid or generic version of a mathematical and mechanical ability test under conditions of explicit or no explicit stereotype threat. Contrary to stereotype threat theory predictions, face validity had either beneficial or non-significant effects on test performance and test perceptions, and did not affect the psychometric properties of either test. Implications and recommendations for future research regarding the study of face validity and stereotype threat are discussed.

Keywords: face validity, stereotype threat, gender differences, ability testing, test perceptions

Face validity holds a tenuous position among the pantheon of test development “desirables” in psychological and educational testing. Broadly defined, face validity refers to the degree to which an assessment tool (i.e., paper-and-pencil test, interview, work sample, etc.) appears practical, valid or relevant to examinees or other administrators who decide on its use in relation to the test’s intended purpose (Anastasi, 1988; Mosier, 1947; Nevo, 1985; Shotland, Alliger, & Sales, 1998; Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993; Wiggins, 1973). Such evaluations require no explicit expertise, and generally involve only surface-level judgments regarding the perceived relevance of a test’s content (Elkins & Phillips, 2000; Hausknecht, Day, & Thomas, 2004). As a result, face validity is typically treated as more of a pleasant afterthought than a quality of dependable measurement (American Educational Research Association, 1999).

Nevertheless, test developers understand that there are substantial benefits to making selection tools more job relevant. Shotland et al. (1998) highlight five specific advantages to improving the face validity of a test. First, face validity has been shown to be positively correlated with test-taking motivation, which in turn has been reliably linked to greater test performance (Chan, Schmitt, DeShon, Clause, & Delbridge, 1997; Hausknecht et al., 2004; Ployhart, Ziegart, & McFarland, 2003). Second, face validity is positively related to organizational attractiveness. A test whose content is transparent and appears job relevant to respondents can act as a signal to the test taker that the employer is not attempting to hide the purpose of the testing instrument in any way. Third, face valid assessments can also serve as realistic job previews in the selection process. A fourth, often overlooked, advantage to using face valid assessment techniques is that managers within the organization typically report greater levels of comfort with and support for using more contextually relevant testing tools (Shotland et al., 1998). Finally, face valid tests are typically less susceptible to legal challenge and are often easier to defend should they be brought to court (Seymour, 1988).

In addition to Shotland et al.'s (1998) list, empirical efforts have also revealed the value of face valid tests. A meta-analysis performed by Hausknecht et al. (2004) summarizes findings that show perceptions of face validity tend to correlate with more positive applicant perceptions, reporting estimated population correlations of .58 with procedural justice perceptions, .33 with distributive justice perceptions, .35 with test motivation, .54 with positive attitude toward tests, and .30 with self-assessed performance. Additionally, the meta-analysis reported positive correlations between enhanced perceptions of face validity and a number of important outcome variables, including recommendation intentions (ρ = .37), self-efficacy (.28) and actual test performance (.13).

Overall, there appears to be consensus from previous research to support the commonly held belief that face validity is a desirable feature of the testing situation across nearly all types of evaluation techniques (Cascio, 1987). By better understanding how, why and what sorts of test-related cognitions are impacted by face validity, selection, training and educational practitioners could effectively adapt their test development strategies to improve their measurement instruments (Bornstein, 1996; Lievens, De Corte, & Schollaert, 2008). It should come as no great surprise, then, that the demand for increasingly specific face valid testing instruments has increased dramatically in recent years (Shotland et al., 1998).

Hidden Costs of Face Validity?

Despite these sentiments, there may be reason to believe that enhancing the face validity of a test is not always so beneficial (Smither et al., 1993). Linn, Baker, and Dunbar (1991) note that while directedness of assessments and transparency may seem desirable, there is no evidence that such test characteristics do not simultaneously produce other unintended—and undesirable—effects. The present research investigates one such possible scenario where improving a test's face validity may result in *negative* collateral effects. Specifically, it is posited that under relatively common circumstances, enhancing face validity can worsen the psychometric properties of a test while introducing stereotype threat-like conditions for particular subgroups.

With respect to its impact on psychometrics, Bornstein (1996) suggests that one potential risk of enhancing face validity lies in the introduction of *construct-irrelevant variance*. Messick (1995) defines construct-irrelevant variance as any methodological quality of a test that affects test takers' responses in a manner that is irrelevant to the measurement of the test's intended construct. Messick (1995) further distinguishes this variance into two general categories—construct-irrelevant difficulty and construct-irrelevant easiness (p. 742). Construct-irrelevant difficulty describes reliable variance in an assessment tool that is extraneous to the focal construct and which makes the test unnecessarily difficult for certain respondents; construct-irrelevant easiness, on the other hand, manifests when cues in item content, context or format permit some individuals to achieve scores that are invalidly high with respect to the intended focal construct.

While Messick (1995) argues that both forms of irrelevant variance likely occur in all measurement tools to some extent, there is reason to believe that face validity manipulations may facilitate their emergence. In the case of construct-irrelevant difficulty, for example, a mathematical ability test in which items are presented in a job-relevant versus generic format (i.e., asking “Two boxes of Product A plus two boxes of Product B” rather than “2 + 2”) likely imposes greater verbal comprehension requirements. As a result, individuals with poorer verbal comprehension skills may achieve test scores that are artificially low and not wholly indicative of their underlying mathematical ability. On the other hand, construct-irrelevant easiness can emerge when improvements to face validity enable some participants to better “guess” what is being assessed (e.g., respond in more socially desirable ways, engage in impression management, etc.; Bornstein, Rossner, Hill, & Stepanian, 1994; Tett, Anderson, Ho, Yang, Huang, & Hanvongse, 2006) or makes the material more readily grasped by certain test takers (e.g., using a reading comprehension passage that is well-known to some readers, Messick, 1995). To the extent that face validity thus increases susceptibility to “faking,” self-presentation biases, or differential familiarity, construct scores on more face valid measures may be inflated.

Another possible effect of increasing face validity may be unwittingly introducing negative contextual stereotypes that exaggerate differences in test perceptions and performance across subgroups. In general, greater test contextualization is often credited with improvements to test taker perceptions (Shotland et al., 1998; Hausknecht et al., 2004) and as a promising means for reducing subgroup differences in test performance (e.g., Chan & Schmitt, 1997; Chan et al., 1997; Hough, Oswald, & Ployhart, 2001; Ployhart et al., 2003). However, if particular contextual elements of a job can implicitly activate stereotype threat (Devine, 1989; Lepore & Brown, 2000; Levy, Stroessner, & Dweck, 1998; Steele & Aronson, 1995), transforming a test item into such a context may inadvertently result in more negative reactions among those in the stereotyped group and potentially reduce those members’ likelihood of correctly answering questions on the test.

In the present study, the effect of enhancing face validity on test taker perceptions and performance in a simulated application procedure is examined under conditions in which greater contextualization adds job-relevant material that is traditionally male-stereotyped. Two specific cognitive ability tests assessing mechanical and mathematical ability were used for these purposes. Evaluations tapping both of these constructs have long been shown to exhibit significant and reliable gender differences in favor of males (e.g., Bennett & Cruikshank, 1942; Feingold, 1988) and are often employed in occupations where masculine stereotypes and disproportionate gender ratios in favor of males traditionally exist in the employee population (e.g., mechanics, engineers). Thus, the use of these tests in the context of applying for a male-dominated occupation contributes both to a sense of external realism that is crucial to maximizing the effectiveness of face validity improvements (e.g., Anastasi, 1988; Lievens et al., 2008) as well as the potential for observing possible stereotype threat effects.

Study Rationale

Enhancing the face validity of a measure requires that one transform the content or context of the test; however, if by enhancing face validity one simultaneously introduces a context that invokes strongly-held stereotypes about a particular group, the end result may be problematic. Consider a situation in which a manufacturing company wishes to enhance the face validity of a perceptual acuity test. A common and reasonable approach might be to make the content of each item reflect the mechanistic, technical context in which the plant operates (for example, by asking test takers to count the number of gears with five teeth on a page rather than the number of stars with five points). In doing so, though, the item context of the question (gears versus stars) could invoke setting and task characteristics that subtly remind the respondent of the male-typing of jobs in this domain (Spence, Helmreich, & Stapp, 1975). In turn, the face valid context has the potential to invoke stereotypic reactions concerning females' poor performance on mechanical tasks and thereby negatively influence the performance of women on the test relative to how they would have performed had the additional context not been introduced.

This form of stereotype activation is considered a subtle form of stereotype threat (Steele & Aronson, 1995). Stereotype threat theory's primary prediction is that stereotyped individuals perform more poorly on an evaluative task (e.g., female applicants taking a mathematics test) when the presence of a threatening performance-stereotype is made salient versus non-threatening conditions where the evaluative stereotype is not salient (see Steele, 1997; Steele, Spencer, & Aronson, 2002). Note that stereotype threat is thus a within-group effect, even though researchers and the popular media often present it as an explanation for between-group differences (Sackett, 2003; Sackett, Hardison, & Cullen, 2005).

Stereotype threat can be activated by either indirectly cuing test takers to the link between a stereotype and performance on a test (*implicit stereotype threat activation*) or directly declaring that members of a social group tend to perform worse on the test than a comparison group (*explicit stereotype threat activation*). By this definition, a change in face validity that simultaneously introduces a negative stereotype into the testing scenario would be considered an implicit form of stereotype threat activation. Recall the earlier example of the manufacturing plant’s perceptual acuity test; if the use of face valid gears rather than generic stars as item context were enough to subtly cue women to the stereotype of male superiority in mechanically-related domains, then the necessary conditions for stereotype threat are induced without participants’ immediate awareness.

However, researchers who have examined the impact of face validity (cf., Dwight & Alliger, 1997; Holtz, Ployhart, & Dominguez, 2005; Smither et al., 1993) have generally found support for the popular notion that greater face validity positively influences test taker perceptions (Elkins & Phillips, 2000; Hausknecht et al., 2004; Whitney, Diaz, Mineghino, & Powers, 1999) and improves certain psychometric properties of the test (e.g., internal consistency, error variances; Holtz et al., 2005; Lievens et al., 2008). Given this evidence, how might stereotype threat-inducing face validity operate differently than more conventional face validity changes? Perhaps the most direct explanation can be gleaned from the concepts summarized previously from Messick (1995)—threat-inducing face validity has the potential to increase the construct-irrelevant difficulty of the test for the negatively stereotyped group. However, rather than the construct-irrelevant variance stemming from deficiency in an unrelated construct (i.e., reading comprehension on a mathematical ability test), the performance decrement is attributed to the self-evaluative apprehension elicited by stereotype threat (Steele & Aronson, 1995). Thus, the end result is a less accurate assessment of true scores for the stereotyped group.

In sum, there is reason to believe that introducing stereotyped job context to the mechanical and mathematical ability tests could negatively impact their construct validity for females as a result of added construct-irrelevant difficulty. In the present experimental design, this would be indicated by a pattern of nonequivalent measurement such that the factor loadings and means of the latent measurement models for women in a face valid versus non-face valid (i.e., generic) testing condition would differ significantly.

> *Hypothesis 1*: The latent factor loadings and means for women will be nonequivalent across conditions of face validity on the math and mechanical comprehension tests. Specifically, factor loadings and factor means will be smaller for women who take the face valid version of the test versus those who take the generic version of the test.

In addition to examining implicitly induced stereotype threat vis-à-vis face validity, the inclusion of a condition in which stereotype threat is explicitly activated is desirable for comparison purposes. Interestingly, a recent meta-analysis by Nguyen and Ryan (2008) on stereotype threat effects found that for women in math testing contexts, implicit threat-activating cues produced the largest performance effect sizes followed by blatant and moderately explicit cues (*d*s = |.24|, |.18|, and |.17|, respectively). The authors note that this pattern of findings may be attributable to explicitly threatening cues prompting test takers to “overperform,” a phenomenon identified in the literature as stereotype reactance (e.g., Kray, Thompson, & Galinsky, 2001; McFarland et al., 2003). Levy (1996) argues that explicit priming of a negative stereotype might produce a weaker effect than subtle priming because the latter bypasses individuals’ conscious coping mechanisms and other psychological resources used to minimize debilitating self-evaluative tendencies and thus can directly affect task performance. In keeping with meta-analytic findings on stereotype threat activation for women, we expect:

Hypothesis 2a: Women will perform worse on the face valid version of the math and mechanical ability tests than on the generic version.

Hypothesis 2b: The implicit manipulation of stereotype activation (i.e., face validity enhancement) will produce a greater negative effect on performance for women than explicit stereotype activation (i.e., direct statement of the performance stereotype).

Nguyen and Ryan (2008) report that in conditions in which no mention of a negative stereotype is made, women were still typically outperformed by men on mathematical ability tests (mean effect size $d = |.26|$), a finding consistent with the broader literature on sex differences in ability testing (Halpern et al., 2007). In threat-activated conditions, though, this effect size increases to $d = |.39|$, denoting a larger gap in test performance between males and females when stereotypes are made salient. Subsequently, a common conclusion drawn from these findings is that stereotype threat may account for a significant proportion of between-group testing differences in applicant hiring practices. However, as the baseline performance differences from Nguyen and Ryan (2008) would suggest, stereotype threat is only a sufficient—but not necessary—condition for such discrepancies. Given that the requirements for explicit stereotype threat are usually never achieved in typical employment testing contexts, critics of stereotype threat often argue that its impact on between-group performance differences is likely minimal, if not nonexistent, in real world applications (Sackett, 2003; Sackett et al., 2005).

Of note, though, this conclusion does not rule out the possibility that more common, subtle manipulations of stereotype activation—such as using face valid items in a stereotyped domain—could still lead to enhanced group differences. For example, in addition to the possibility of adding construct-irrelevant difficulty to the test for women, threat-inducing face validity may also contribute to construct-irrelevant easiness for men (Messick, 1995). This occurrence is captured in the notion of *stereotype lift*, a phenomenon in which the test performance of non-threatened group members is enhanced when a negative stereotype about another group's performance is made salient relative to when no stereotype is mentioned (see Walton & Cohen, 2003, for a meta-analytic review). With respect to the manipulation of face validity in the present study, if the addition of the traditionally male-oriented job context to the test items makes those items more familiar, accessible, or readily grasped by males, one might expect an "artificial" performance increase for male respondents *irrespective* of females' performance on the test (Messick, 1995). Regardless of whether male performance improves on the test or female performance diminishes, though, an increase in the performance gap between genders should be expected. Hence, it is hypothesized that:

> *Hypothesis 3a*: Males will outperform females by a greater margin on the face valid form of the ability tests than the generic form.
>
> *Hypothesis 3b*: Males will outperform females on both ability tests, but the male advantage will be larger in the face valid-generic comparison than in the explicit stereotype activation-no activation comparison.

The evidence reviewed earlier indicates that face validity tends to positively influence test taker perceptions (cf. Hausknecht et al., 2004); yet, given the description of face validity as a form of implicit stereotype threat in the current study, might one expect such an increased level of stereotype threat to negatively influence test taker perceptions? In short, previous research and theory does not offer a clear answer. For example, although Ployhart et al. (2003) found that stereotyped respondents' test-taking motivation was negatively correlated with levels of perceived stereotype threat, Nguyen, O'Neal, and Ryan (2003) failed to find evidence to support relationships between threat and similar test taker perceptions. Additionally, Steele and Aronson (1995) posit that the mechanisms through which individuals experience threat are not necessarily conscious; thus it is possible that an individual under such conditions might not even be aware of its effects. This would seem even more applicable to situations in which the stereotype is activated implicitly (Kray et al., 2001; Levy, 1996; McFarland et al., 2003). In sum, there is little empirical or theoretical evidence to suggest that enhanced levels of stereotype threat negatively affect the conscious perceptions of test takers.

As such, the present study captures a variety of test taker perceptions previously examined in the applicant reactions literature in an attempt to replicate the finding that greater face validity correlates with more positive perceptions (cf. Hausknecht et al., 2004). Support for this hypothesis would be especially intriguing if support for the previously proposed hypotheses is found as well, as it would indicate that although the proposed enhancements in face validity could improve test taker perceptions they come at the cost of increased construct contamination and performance discrepancies.

> *Hypothesis 4*: Participants will report higher ratings of self-assessed performance, test ease, pursuit intentions, recommendation intentions, job attractiveness, procedural fairness, and perceived predictive validity in the face valid versus generic testing condition.

While stereotype threat theory holds that one need not be actively aware of a stereotype to experience its negative effects (Steele, 1997; Steele & Aronson, 1995), this does not preclude that individuals may nonetheless *be* more or less conscious of threat. If this were the case, one would expect the effects of perceived threat to operate similarly to other perceptual variables in the context of test taking (e.g., Chan et al., 1997). Indeed, Ployhart et al. (2003) report that those with greater levels of perceived threat indicated lower perceptions of face validity (β = -.16) and decreased test-taking motivation (β = -.18). Furthermore, perceptions of stereotype threat exhibited a positive relationship with test anxiety (β = .14), which subsequently exhibited a negative relationship with cognitive ability test performance (β = -.25).

Given this rationale, an exploratory effort to extend the work of Ployhart et al. (2003) was conducted to examine the role of perceived stereotype threat in the experimental manipulations. We expect that perceived stereotype threat will be negatively correlated with performance, as those who are consciously thinking about stereotypes would be expected to have greater off-task thinking and therefore larger performance decrements. Further, those who are conscious of stereotype threat in the testing context should be more likely to view the test negatively.

Method

Design and Participants

This study used a 2 (test format: face valid versus generic) x 2 (threat activation: explicit threat versus no explicit threat) between subjects design, with test type (mechanical versus math) as a within subjects factor. Many stereotype threat research designs often include a “non-evaluative” control condition in which participants are informed that a test is used only as a problem-solving exercise and is not scored (cf., Nguyen & Ryan, 2008); however, given that the purpose of this experiment was to examine the effects of potentially threat-inducing face validity in a selection context—a situation that is always predicated on evaluative assessment and thus necessarily meets the minimal requirements for evoking stereotype threat (Steele & Davies, 2003)—a “threat-present” versus “threat-devoid” comparison does not contribute beyond what has been demonstrated in the literature previously and, further, does not accurately reflect an ecologically valid applicant testing procedure.

Participants were undergraduate students (*n* = 358) recruited from psychology courses at a large Midwestern university who completed the experiment for course credit. The sample was primarily composed of young (*M* = 19.44, *SD* = 1.48), White (79%; Black = 10%; Asian = 6%; Hispanic = 1%) females (68%). Assignment to the between-subjects conditions was random, though care was taken during recruitment to ensure the number of participants and the proportion of males and females within each cell of the 2 x 2 design were approximately equal.

Procedure

Participants were tested in groups of 20 to 50 individuals per test session and were assigned to either the explicit threat or no explicit threat condition. Prior to enrolling in the experiment, participants were informed that they would be adopting the role of a job applicant in which they would be taking a test of mechanical and mathematical ability as part of a company's hiring procedure. Participants were given a packet of materials that contained an informed consent, the experimental instructions, a face valid or generic version of the mathematical and mechanical ability tests, the perceptions measures and a demographics questionnaire. Each packet also included a detailed description of the job for which the participants were applying; namely, that of a manager in charge of supervising maintenance and repair workers for a real estate company. The job description listed the relevant tasks, knowledge, skills and abilities of the position, making clear that mechanical and mathematical ability were essential to performance on the job. Following the task and KSA description, the income/benefits of the position were provided that described the job as reasonably desirable (e.g., good pay, benefits and opportunities for advancement). Furthermore, to encourage participants to take the tests seriously, individuals were informed that the top performers on the ability tests would receive 20 dollars.

The job of manager of maintenance and repair workers was selected for use in the study for two specific reasons. First, it was important that the domain of the job in question be negatively stereotyped against women. Past research has demonstrated that domains in which mathematical and mechanical knowledge, ability and skill are important (such as with maintenance and repair workers, O*NET, 2004) are stereotypically regarded as male-advantaged (Spence et al., 1975). Thus this particular job title serves as an implicit indicator to females that they are at a disadvantage relative to males during the application process. Second, the managerial ranking was included in order to make the job relatively more appealing to the college-educated participants of the study (Muchinsky, 2004). It was emphasized to participants on two different occasions (once in the job description, once in the experimental instructions) that mathematical and mechanical ability were important to the job.

After all materials had been distributed, instructions were read aloud to participants. Individuals then completed a pretest measure of motivation, followed by both ability measures, and finished the experiment by filling out the test taker perceptions and demographic measures. The order of the ability tests was counterbalanced across all participants. Average time to complete was 50-60 minutes.

Manipulations

Face validity. Face validity of the mathematical and mechanical ability measures was manipulated using a strategy similar to that employed in Smither et al. (1993) and Holtz et al. (2005) in which the context of all test items was carefully altered to more appropriately match the job in question. A generic version of each test was first developed devoid of any context directly relevant to the job domain. The face valid version was then developed by adding job-relevant terms to the exact same items. Graphics were also altered so as to be more or less representative of the job domain (see Figure 1 for example items). Thus, both the face valid and generic versions contained parallel item *content*, though item *context* was appreciably different.

Stereotype threat. Explicit stereotype threat was manipulated by either informing or not informing participants of the male-female performance differences on the mathematical and mechanical ability tests. Specifically, the instructions to participants in the explicit threat condition contained the following line which was read aloud to participants after the normal instructions: “It is quite typical and expected that males will do significantly better than females on these tests of mechanical and mathematical ability." In the condition with no explicit threat, this line was not included in the instructions nor was it verbalized to participants (cf. Spencer, Steele, & Quinn, 1999).

Measures

Unless otherwise noted, all items were measured using a five-point, *strongly disagree* to *strongly agree* format, where higher numbers indicated more of the measured construct.

Manipulation check. A manipulation check was included to assess whether participants were aware of the face validity alterations to the ability tests. This scale consisted of seven items from Smither et al.'s (1993) job related-face validity scale and two additional items created by the authors that explicitly asked participants about the face validity of the math and mechanical tests in relation to the job domain of maintenance and repair workers. Sample items from this scale included "The items on the tests made direct reference to tasks that people doing maintenance or repair work might perform" and "The items on the test did not appear relevant to any of the job duties a maintenance or repair worker might perform" (reverse scored). The reliability estimate for the full nine-item scale was $\alpha = .87$.

Mathematical ability. The mathematical ability test consisted of 30 multiple-choice questions. The items were adapted from a test preparation book for the math section of a popular college entrance exam and covered a variety of topic areas including basic geometry, algebra and trigonometry. Questions on the math test were scored dichotomously (correct/incorrect) and summed together to obtain a total scale score. The internal consistency estimate for both the face valid and generic version of the math exam was $\alpha = .83$.

Mechanical ability. The mechanical ability test initially consisted of 30 multiple-choice questions selected and adapted from a pool of 218 mechanical comprehension items generated by the authors as part of a separate test development project. Prior research with this item pool revealed, on average, item characteristics comparable to those reported for a popular, commercially-available mechanical aptitude test (Bennett, 2006). Questions on the mechanical ability test were scored dichotomously (correct/incorrect) and summed together to obtain a total scale score. Item-level analyses indicated that five questions with near-zero/negative item-total correlations should be removed to improve overall scale reliability. The reliability coefficients for the 25-item face valid and generic versions of the exam were $\alpha = .79$ and .72, respectively.

Test-taking motivation. The 10-item pretest motivation scale was adapted from the Test Attitude Survey developed by Arvey, Strickland, Drauden, and Martin (1990). Reliability of the scale was high, with $\alpha = .96$. A sample item is, “I will try my best on this test.”

Self-assessed performance. Self-assessed performance was measured with a five-item scale developed by Brutus and Ryan (1996). Although the full scale has been used in previous research assessing test taker perceptions (cf. Wiechmann & Ryan, 2003), two items were identified as problematic in the present sample due to low item-total correlations. Reliability of the modified three-item version of the self-assessed performance scale was $\alpha = .82$. A sample item from the measure reads, “I did well on the test.”

Test ease. Test ease was measured with a five-item scale developed by Wiechmann and Ryan (2003). The reliability of the scale was α = .89. An example item from the measure is, “I found this test too simple.”

Intentions to pursue. Pursuit intentions were captured by a five-item measure (α = .87) taken from Highhouse, Lievens, and Sinar (2003). A sample item is, “I would accept this job if offered.”

Recommendation intentions. Recommendation intentions were measured with a four-item scale adapted from Gilliland’s (1994) applicant reaction scales. Reliability for the measure was α = .88. An item from the measure asks, “I would recommend jobs with similar application tests to my friends.”

Job attractiveness. Test takers’ perceptions of job attractiveness were measured using a five-item scale developed by Highhouse et al. (2003). The internal consistency of the scale was α = .92. A sample item is, “This job is very appealing to me.”

Procedural fairness. Procedural fairness was captured with four items adopted from Gilliland’s (1994) procedural justice scales. The reliability of the measure was α = .89; an example item from the measure is, “Whether or not I would get the job, I feel the selection *process* was fair” (emphasis in the original).

Perceived predictive validity. A five-item scale developed by Smither et al. (1993) was used to capture perceived predictive validity (α = .86). An item from the measure reads, “I am confident that the test can predict how well an applicant will perform on the job.”

Perceived stereotype threat. Two eight-item scales developed by Ployhart et al. (2003) were initially used to measure participants’ perceptions of stereotype threat.

Although both scales contained the same items, the first referenced individuals' perceptions of threat in relation to the domain of mathematical ability, while the second referenced perceptions of threat in relation to the domain of mechanical ability. Based on the factor structure of the scales presented in Ployhart et al. and preliminary analyses with the present sample, the same single item was dropped from both measures due to poor item-total and inter-item correlations. The scale reliability of both the seven-item perceived stereotype threat-math measure (α = .85) and the seven-item perceived stereotype threat-mechanical measure (α = .77) were acceptable. An example question from both measures is, "Some people feel that I have less (math/mechanical) ability because of my gender."

Demographics/background. A brief background questionnaire asking information about participants' race, gender and age was also included in the survey packets. In addition, participants were asked to report their current overall GPA and their ACT or SAT scores.

Results

Preliminary analyses

Thirteen participants were removed from the dataset who either filled out all measures in the experimental session in less than 20 minutes or who provided long strings of a single response option, both indicators that the individuals did not attend to the task seriously. Table 1 presents the means, standard deviations and correlation matrix of the study variables for the final dataset (*n* = 345). In addition, the means and standard deviations for males' and females' performance on the ability tests in each of the study's cells are given in Table 2.

The order of ability test presentation (math test first/mechanical test second versus mechanical test first/mechanical test second), time taken to complete all measures (less than 30 minutes versus 30 minutes or longer), and general cognitive ability (as measured by ACT/SAT performance) were included as control variables in all analyses.

Manipulation Check

One-way between-subjects ANOVA were used to evaluate whether individuals' perceptions of face validity differed across the face valid and generic testing conditions. For the entire sample, results of the ANOVA revealed that those taking the face valid version of the test (M = 3.63, SD = .67) perceived the ability tests as more face valid than those who took the generic version of the test (M = 3.40, SD = .67), F(1, 328) = 9.95, p < .01, corresponding to a small-to-moderate effect size of d = .36. Although a small mean difference between conditions may be taken as evidence that the manipulation of face validity was not potent, previous research (cf., Holtz et al., 2003; Smither et al., 1993) using similar face validity manipulations has failed to find any significant differences in perceptions of face validity or job relatedness across generic and face valid test formats. Additionally, gender-specific analyses revealed that the face validity manipulation was significantly more potent for women (d = .42) in the sample than men (d = .22). This is particularly noteworthy given that the assessment of stereotype threat is a within-group phenomenon and thus only women should be negatively affected by the presence of threat induced by test contextualization. As such, the fact that all individuals in the sample perceived greater face validity in the appropriate condition and that females' between-condition means were the largest suggests that the necessary information needed to make face validity judgments was made salient by the experimental manipulation, thus enabling an examination of implicit threat effects should they be present.

Hypothesis Testing

Hypothesis 1 predicted that the latent factor loadings and means for both the mathematical and mechanical ability tests would differ significantly for women who took the face valid versions of the tests versus those who took the generic versions. The multiple-group confirmatory factor analysis (MGCFA) procedure outlined by Vandenberg and Lance (2000) was used to examine this prediction. This approach entails systematically assessing model fit at each step in a series of increasingly restrictive tests to determine whether the factor structure of a measure is equivalent across two (or more) unique groups of individuals. To reduce the number of estimated parameters in the CFA model, items from the math and mechanical tests were separately clustered into five parcels prior to analysis (Baggozzi & Edwards, 1998; Baggozzi & Heatherton, 1994), resulting in a 10 x 10 covariance matrix for each group (face valid vs. generic). Based on the recommendations of Vandenberg (2002), the decision on which factor loadings to fix in order to scale the variance of the latent factors was made by conducting exploratory factor analyses and then identifying the most invariant parcels for each factor across the face valid and generic groups of women. The MGCFA analyses were conducted in Amos 5.0 (Arbuckle, 2003) using maximum likelihood estimation.

The first step in assessing measurement invariance involves determining whether a common factor model can be fit to the data reasonably well for both groups. Model M1 in Table 3 presents the results of the configural invariance test in which the two latent factors, mathematical and mechanical ability, predicted observed performance on their respective item parcels. This model fit the data well for women in both the face valid and generic testing conditions, thus supporting the decision to retain this structure as the baseline model against which the remaining tests of invariance could be compared.

Following this step, invariance of factor loadings for women across both conditions was assessed by constraining the factor loadings for the face valid testing group to be equal to those in the generic testing group (Model M2, Table 3). Contrary to Hypothesis 1, equating the factor loadings across groups did not result in a significantly worse model fit (M2 – M1: $\Delta\chi^2(8) = 6.81$, *n.s.*), indicating that the item parcel loadings were not significantly different for women in the face valid versus generic conditions on either the math or mechanical ability test.

The next step in the MGCFA procedure is to assess whether the intercepts (i.e., mean scores) from the regression equations for each of the observed variables are equal across groups (Vandenberg & Lance, 2000). As Model M3 in Table 3 shows, equating the item parcel means resulted in a significant decrease in fit from the model of metric equivalence (M3 – M2: $\Delta\chi^2(10) = 46.66$, $p < .001$). An examination of the parameter estimates indicated that the decrement in fit could be attributed to significant group differences in performance on two mechanical item parcels where women in the face valid condition achieved significantly higher mean scores on these items than women in the generic condition. A post hoc investigation of the content of these parcels at the item level revealed that the most likely explanation for the performance differences could be attributed to confusion regarding the graphics of items on the generic version of the mechanical test, making it potentially more difficult for participants in the generic condition to answer these items correctly. A partially invariant model was therefore estimated that allowed the intercepts for these two mechanical parcels to vary freely across the groups. The decrease in fit between this model (M4, Table 3) and Model M2 was not significant (M4 – M2: $\Delta\chi^2(8) = 12.42$, *n.s.*), indicating that women in both the face valid and generic conditions achieved equivalent mean scores on the remaining math and mechanical item parcels.

The results of the preceding analyses provide strong evidence in support of invariance in the measurement portion of the multigroup factor model[1]. To assess the structural components of the model, a *t*-test was conducted that compared the latent means for math and mechanical ability from the face valid versions of these tests with their respective means from the generic versions of the tests. The analyses revealed no significant differences for either math ($t(234) = .75$, *n.s.*) or mechanical ability ($t(234) = .90$, *n.s.*), indicating that the latent means for women on the math and mechanical ability tests were similar in both the face valid and generic conditions. In sum, the results of the MGCFA procedure failed to support Hypothesis 1; neither the factor loadings nor latent means for women as measured by the mechanical and mathematical ability tests differed significantly across conditions of face validity.

Hypotheses 2a and 2b were concerned with females' performance on the ability tests in relation to the stereotype threat manipulations present in the study. To examine these predictions, performance on the math and mechanical ability test were separately regressed onto the face validity and explicit threat manipulation conditions. As shown in Table 4, the face validity manipulation had no effect on performance for women on the math test above and beyond the control variables ($\beta = .03$, *n.s.*), though it did explain significant variance on the mechanical test ($\beta = -.14$, $p < .05$). However, this effect was opposite to that predicted by the hypothesis—females who took the face valid version of the mechanical ability test tended to perform better than those who took the generic version of the test (*d* = .41)—therefore failing to support Hypothesis 2a.

Hypothesis 2b proposed that the implicit manipulation of stereotype threat through changes in face validity would produce a greater negative effect on women's performance than the explicit activation of stereotype threat. Since the test for Hypothesis 2a failed to provide evidence of a negative effect for face validity on women's performance on either ability test, only the performance difference attributed to the explicit activation of stereotype threat was left to be examined. However, explicit threat did not produce a significant main effect on female's performance for either the math or mechanical ability test (Table 4). Therefore, the non-significant main effects of the stereotype activation manipulations failed to support Hypothesis 2b.

Hypotheses 3a and 3b examined the male-female differences in performance on the ability tests in relation to the stereotype threat manipulations present in the study. To test these hypotheses, a regression procedure was used in which math and mechanical test performance was regressed separately onto the dummy variables for gender, face validity, and explicit threat (Table 5). Hypothesis 3a predicted a significant interaction between gender and face validity such that males would outperform females by a greater margin on the face valid form of the ability tests than the generic form. As Table 5 demonstrates, this interaction effect was only significant for the math test (β = .15, $p < .05$). A closer examination of this interaction in Figure 2A reveals that the male-female performance difference was substantially larger in the generic testing condition (*d* = .57 versus *d* = .10); furthermore, the interaction was such that although female performance remained relatively similar across test versions, male performance appeared to decrease somewhat on the face valid version of the test (though this difference was only marginally significant, $t(107) = 1.67$, $p = .10$). These results are contrary to what was predicted and do not support Hypothesis 3a.

Hypothesis 3b predicted that males would outperform females on both ability tests, but the size of the male advantage would be greater in the face valid-generic comparison versus the explicit stereotype activation-no activation comparison. Table 5 shows that males did significantly outperform females on both the math ($\beta = .14$, $p < .05$) and mechanical ability tests ($\beta = .34$, $p < .05$). Thus to determine the degree to which males outperformed females across the stereotype activation methods, effects sizes were first calculated comparing male and female performance for each level of the implicit (face valid, generic) and explicit (explicit threat, no explicit threat) conditions (see Figure 2). Within-condition effect sizes were then subtracted from one another (i.e., face valid – generic effect sizes, explicit threat – no explicit threat effect sizes) to obtain an estimate of the absolute difference in male-female performance for each activation method (d_{diff}). These were then compared across activation method (i.e., implicit versus explicit) for each ability test to examine differences in male-female performance discrepancies.

On the math test, the male-female performance difference in the face validity-generic comparison was $d_{diff} = .47$ compared to a male-female performance difference of $d_{diff} = .18$ in the explicit threat-no explicit threat comparison. However, the direction of the performance difference in the face validity-generic comparison was opposite to that predicted; namely, the male-female performance difference was greater on the generic version of the test than the face valid version. On the mechanical test, the male-female performance difference in the face validity-generic comparison was d_{diff} = .12 compared to a male-female performance difference of d_{diff} = .09 in the explicit threat-no explicit threat comparison. Once again, although the male-female performance difference in mechanical test performance attributed to the face validity manipulation was slightly larger, the direction of the performance difference was not in the predicted direction. Furthermore, the direction of the explicit threat-no explicit threat comparison was also opposite to that predicted, such that the male-female performance difference was larger when there was no explicit threat relative to when explicit threat was present. In sum, the data failed to support Hypothesis 3b for either the math or mechanical ability tests[2].

Hypothesis 4 proposed that participants would report higher ratings of self-assessed performance, test ease, pursuit intentions, recommendation intentions, job attractiveness, procedural fairness, and perceived predictive validity in the face valid versus generic testing condition. Independent sample *t*-tests revealed that this prediction was only supported for perceived predictive validity ($t(328) = 3.13$, $p < .05$); no significant mean differences were found for any of the remaining applicant reactions. Of note, though, participants' *perceptions* of face validity (indicated by responses to the nine-item face validity manipulation check measure) were positively correlated with self-assessed performance ($r = .22$, $p < .05$), recommendation intentions (r =. 24, $p < .05$), procedural fairness ($r = .50$, $p < .05$), and perceived predictive validity ($r = .49$, $p < .05$). Furthermore, these correlations remained significant even when controlling for actual performance on the mathematical and mechanical ability tests. In sum, Hypothesis 4 was only partially supported.

As a final exploratory effort and extension of Ployhart et al. (2003), the effects of perceived stereotype were also examined. These analyses revealed three findings of particular interest. First, females in the explicit stereotype activation conditions did appear to perceive greater levels of threat in relation to the math (r = -.29, p < .05) and mechanical (r = -.39, p < .05) performance domains compared to those in the condition of no explicit threat, even when controlling for performance on the tests. Furthermore, neither of the perceived threat measures was significantly correlated with the face validity manipulation. Second, across the entire sample, perceived threat in both performance domains was negatively related to perceptions of self-assessed performance and test ease, though no other significant correlations were observed for the remaining reactions measures (see Table 1). Lastly, participants' perceived threat in the math (r = -.25, p < .05) and mechanical performance domains (r = -.22, p < .05) was negatively correlated with performance on the mechanical ability test, though not the mathematical ability test.

Discussion

The present study posited that there are situations in which improvements to face validity can only be achieved by introducing contextual information that may be detrimental to the performance of certain subgroups. Contrary to our predictions, the results revealed that even such potentially threatening face validity enhancements tended to have beneficial (or, at the very least, non-negative) effects. The introduction of the face valid context did not add construct-irrelevant variance to the measurement of the math and mechanical ability constructs for female respondents (Messick, 1995) and, on average, seemed to improve their performance on the mechanical ability test.

Furthermore, the male-female performance discrepancy was somewhat reduced across both ability domains when the tests were more versus less face valid.

The design of the present experiment also attempted to compare the effects of a more ecologically valid induction of stereotype threat with the traditional but less realistic activation method commonly used in laboratory studies (Sackett, 2003; Sackett et al., 2005). Of note, implicit threat elicited through changes in face validity demonstrated approximately the same mean effect on male-female performance differences as the activation of explicit threat through the verbal instruction protocol (d_{avg} in Figure 2). Additionally, the experimental manipulations of implicit and explicit threat were only minimally related to participants' reactions to the hiring process, although perceptions of stereotype threat were positively correlated with these outcomes.

Theoretical Implications

The degree of face validity exhibited by a testing instrument resides "in the eye of the beholder" as a judgment of the perceived relevance of a test in relation to its intended purpose (e.g., Anastasi, 1988; Elkins & Phillips, 2000; Nevo, 1985). Following this definition, most previous research has operationalized and measured face validity by asking test takers to provide responses to perceptual measures of face validity and then correlating these ratings with other perceptual outcomes or actual performance (e.g., Hausknecht et al., 2004). Unfortunately, such measurement approaches often suffer from a number of confounding effects (e.g., halo, common method bias, etc.) that can artificially inflate the reported relationships between face validity and test taker reactions/performance. The present study presents one of the few attempts to avoid these issues by explicitly manipulating item context to examine the role of face validity in ability testing.

To this end, it is interesting to note that the observed relationships between *perceptions* of face validity and test taker reactions/performance were remarkably similar to those reported in the Hausknecht et al. (2004) meta-analysis. However, the mean differences found for these variables across the face valid versus generic test versions demonstrates that the impact of face validity may not be as substantial as previously thought. This pattern of results is particularly meaningful for future research in the area of face validity and similar perceptual variables related to test taking (i.e., test ease, perceived predictive validity, etc.), as well as interpreting the ecological validity of published results regarding the effects of face validity that have almost exclusively employed correlational or survey research. While such methodologies are invaluable for revealing interesting phenomenon, the control offered by laboratory simulations enables researchers to manipulate and assess their observed effects much more precisely (McGrath, 1986). Thus, we support Smither et al.'s (1993) recommendation and view the incremental approach (i.e., survey research followed by careful experimental manipulation) to investigating the perceptions of test takers as crucial to improving our overall understanding of these variables' role in the arena of selection and assessment.

For example, the current study revealed new considerations regarding the utility of face validity that have previously been unanswered or left to conventional wisdom (cf. Cascio, 1987). First, enhancing face validity by introducing job relevant context at the item level may not be a potent enough manipulation to reliably influence test takers' performance and reactions (negatively or positively) across *all* domains of ability testing.

Although Anatasi (1988, p.45) suggests that “face validity can often be improved by merely reformulating test items in terms that appear relevant and plausible in the particular setting in which they will be used,” there is often little rationale to guide when or why such changes should be pursued. The present results indicate that changes in face validity made little difference in performance on the mathematical ability test; however, performance on the mechanical ability test was greatly improved for both women (*d* = .41) and men (*d* = .31) on the more job relevant version relative to the generic test.

While this study does not provide a clear explanation as to why this discrepancy across content area was observed, a number of possible explanations could be hypothesized. Perhaps instruments that assess more abstract reasoning or visuospatial concepts/principles (e.g., content based on physics, spatial acuity, etc.) are better aided by greater context specificity than instruments that tap more computationally-based concepts (e.g., content based on math, economics, etc.). Alternatively, greater face validity may only improve performance on measures in which the subject matter is less familiar to the test taker population; presumably, the college student sample employed in the present study had likely been exposed to more experiences in which mathematical ability was required than situations that required active processing of mechanically-related principles. While speculation, these possibilities suggest that greater consideration may be warranted before continuing to promote the blanket effectiveness of face validity.

A second implication concerns the recommendation that face validity can serve as a useful strategy for reducing subgroup performance differences on ability tests (e.g., Chan & Schmitt, 1997; Chan et al., 1997; Hough et al., 2001; Ployhart et al., 2003). Although our results would seem to support this conclusion, they also revealed that minimizing the performance gap in this manner may *still* come at an unexpected cost to certain test takers. As shown in Figures 2A and 2C, male-female differences in performance were reduced on the face valid versions of the math (dropping from *d* = .57 to *d* = .10) and mechanical (dropping from *d* = .93 to *d* = .81) ability tests, a relatively significant finding considering the ability measures employed in this study have traditionally demonstrated large gender differences (e.g., Bennett & Cruikshank, 1942; Feingold, 1988). However, while the performance gap on the mechanical ability test narrowed because women's performance responded more positively to the increased face validity, the performance gap on the math test diminished primarily because males seemed to perform slightly worse on the face valid version. Although this performance drop for males failed to achieve statistical significance, future research and applications in which face validity is enhanced in an attempt to minimize subgroup performance differences would nevertheless be well served to anticipate whether and how changes to the contextual components of a test could affect all relevant subgroups.

Overall, the results of this study indicate that although previous interpretations of the effects of face validity on test taker perceptions and performance may be slightly exaggerated, the practice is still generally beneficial. However, it should be noted that these findings are only relevant to the effects of face validity with respect to *cognitive* ability testing. Previous studies convincingly demonstrate that noncognitive instruments may be substantially more sensitive—for better or for worse—to face validity alterations with respect to changes in their psychometric properties and relationships with performance/test taker reactions (Chan & Schmitt, 1997; Elkins & Phillips, 2000; Holtz et al., 2005; Lievens et al., 2008; Robie, Schmit, Ryan, & Zickar, 2000; Whitney et al., 1999).

Practical Implications

Findings from this study suggest that the oft-noted relationship between face validity and improved perceptions may actually be smaller than previously suggested. However, this is not necessarily a troubling finding for test developers and organizations; positive spillover emanating from test takers' perceptions of face validity to ratings of job attractiveness, procedural fairness, pursuit intentions, etc. would seem welcome news for organizations seeking a practical way to improve their image to applicants. Additionally, evidence of a possible halo or common method bias does not negate the benefits of face validity in regards to helping an organization's legal defensibility of its selection instruments (Seymour, 1988) or face validity's relationship with other organizationally relevant variables not captured by this study (manager's acceptance/preference for a selection technique, etc.; cf. Shotland et al., 1998). Thus, while the research community should further question exactly how much face validity matters to performance (Smither et al., 1993), the present research indicates that face validity influences test taker perceptions in the desired direction.

Our results also revealed no evidence that test contextualization harmed the psychometric properties of either ability test. Although our theoretical rationale implied that face validity could potentially lead to a number of undesirable effects (cf. Bornstein, 1996; Linn et al., 1991; Messick, 1995), no significant differences in the latent measurement characteristics between the face valid and generic versions of either assessment were revealed. In fact, estimates of internal consistency on the face valid mechanical ability test even saw a marked improvement over the generic version, a novel and as of yet inadequately explained finding in the domain of cognitive ability testing (see Lievens et al., 2008, for a description of this finding with noncognitive tests).

Altogether, our results suggest that the potential risk for inducing implicit threat through job related contextual information on cognitive ability tests is likely very low and thus test developers and administrators are not likely to see any ill effects from similar test alterations. However, previous research by Wicherts, Dolan, and Hessen (2005) has demonstrated that explicit stereotype threat cues do have the potential to negatively influence the measurement characteristics of ability exams across subgroups. Thus, while the results from our research provide one example in which a specific (though more common and ecologically valid) manipulation of stereotype threat on two separate ability tests did not result in a noticeable attenuation in construct validity, practitioners are nevertheless encouraged to carefully consider the contextual characteristics of their applicant hiring systems to ensure that potentially biasing conditions are minimized.

Implications for Stereotype Threat Research

Owing to its elusiveness across empirical studies, the restrictive study designs required to elicit its effects, and the situational-/sample-specific nature of its outcomes, the generalizability of stereotype threat to areas of applicant testing has frequently been called into question (cf., Sackett, 2003; Sackett, Schmitt, Ellingson, & Kabin, 2001; Sackett et al., 2005). However, Nguyen and Ryan (2008) emphasize that the important question for future research in addressing this issue is not whether the results of the theory can be replicated consistently, as meta-analytic evidence across multiple comparison groups have clearly demonstrated its robustness. Rather, the task laid before the next generation of research lies in identifying 1) the *boundary conditions* and limitations of the theory, 2) the *moderators* of its effects, and 3) the changes in *psychological processes* that individuals experience when threat is induced.

With respect to theoretical boundary conditions, the present research suggests that the relative effects of stereotype threat on subgroup performance differences is not as salient when comparing individuals under “high-threat” (e.g., the face valid/explicit threat conditions in the present research) versus “low-threat” (e.g., the generic/no explicit threat conditions). To appreciate this subtlety, one must recall that the primary tenet of stereotype threat theory predicts that stereotyped individuals will perform more poorly on an evaluative task in a threatening context *than they would in a non-threatening/non-evaluative context* (Steele, 1997; Steele & Aronson, 1995; Steele, Spencer, & Aronson, 2002). In other words, the theory states that performance differences produced by stereotype threat are only observable across “threat-present” versus “threat-devoid” comparisons (Steele & Davies, 2003).

However, not only are completely non-threatening/-evaluative conditions difficult to produce experimentally (p. 318, Steele & Davies, 2003), they are highly unlikely to exist in realistic testing and selection contexts because such situations are *necessarily* evaluative in nature (cf., Sackett, 2003; Sackett et al., 2001; Sackett et al.,2005). Given that standards of ethics, legal defensibility, and professional conduct dictate that organizations which administer tests for purposes of legitimate assessment (e.g., selection, training, promotion, etc.) not misinform test takers about the diagnostic purpose of their scores (Ployhart, Schneider, & Schmitt, 2006), the removal of evaluative performance threats from realistic assessment practices is likely not a feasible alternative. In our opinion, then, the more relevant question for stereotype threat researchers is “What should we expect when groups are exposed to stereotype threat at varying degrees of severity?”

To the extent that even fictitious testing contexts heighten an individuals’ sensitivity to performance stereotypes, it is possible (if not probable) that all conditions in the current experiment were threatening—though to differing extents (Steele & Davies, 2003). Based on previous findings (cf., Nguyen & Ryan, 2008), the present research assumed that implicitly manipulating stereotype threat through face validity would generate a more threatening situation than is normally present in a generic testing context and thus should exert a more noticeable effect on women’s performance. However, no evidence was found to support this prediction in the high- to low-threat comparison, thus, suggesting that even potentially threatening face validity alterations are not potent enough to induce stereotype threat effects above and beyond normal testing conditions.

One of the more intriguing findings to emerge from the Nguyen and Ryan (2008) meta-analysis was limited support for certain moderators to the relationship between stereotype threat and performance. However, the authors are also quick to point out that our understanding of the conditional variables that influence the threat-to-performance correlation is limited and a substantial portion of variance in this relationship remains unexplained. As one example, Nguyen and Ryan suggest that specific domains of cognitive ability may exhibit performance discrepancies under conditions of threat differently than general cognitive ability tests, though there is currently not enough data to test this claim. While the present study certainly does not provide a substantive test of this particular hypothesis, our results would appear to offer at least some support for this proposition as no evidence was found of a stereotype threat effect on either of the specific ability tests used in this study. Although a more directed investigation would be needed to explicitly examine this possibility, the results observed here do provide an entrée to future researchers interested in more explicitly investigating test specificity as a potential moderator.

Lastly, in relation to better understanding the psychological processes engendered by stereotype threat, we again note the findings obtained with Ployhart et al.'s (2003) perceptions of threat measure. Previous conceptualizations of stereotype threat (Davies, Spencer, Quinn, & Gerhardstein, 2002; Spencer et al., 1999; Steele, 1997; Steele & Aronson, 1995) state that individuals need not be consciously aware of threat in order to experience negative performance effects; however, such a treatment does not indicate how conscious appraisal of the threat might affect meaningful outcomes. Women in our sample did in fact report greater levels of perceived threat in the math and mechanical performance domains across conditions of explicit threat relative to women not explicitly informed of the negative performance stereotype. Furthermore, these threat perceptions were negatively related to certain domains of test performance and performance-related perceptions (i.e., test ease and self-assessed performance). While we do not argue against the traditional position that stereotype threat need not be consciously experienced to influence an individual's performance in an evaluative situation, these results and those of Ployhart et al. (2003) seem to indicate that at least some individuals *are* actively aware of threat under certain conditions, that such perceptions are related to meaningful outcomes, and that this relationship can be measured. Given this possibility, we agree with Steele and Davies' (2003) recommendation that future research would benefit from examining test taker perceptions as a mechanism through which members from negatively stereotyped subgroups experience emotional and psychological stressors (e.g., test anxiety, decreased motivation, decreased self-efficacy) that could affect performance-related outcomes.

Limitations

Given that this study did not produce the pattern of findings predicted by stereotype threat theory, it is important to consider possible limitations and/or alternative explanations that may explain the observed results. In large part, these considerations primarily stem from concerns around the use of college-aged students as applicants for a fictitious job. First, as the sample may not have produced as realistically high levels of motivation and/or desire to perform as well as would be expected with actual job applicants, one might argue that the null findings related to stereotype were found because the stakes associated with the scenario were not consequential enough for test takers to feel at risk of confirming a self-relevant negative stereotype (Steele & Aronson, 1995; Steele & Davies, 2003). Apart from the obvious ethical restrictions that prevent manipulations of stereotype threat in a true selection procedure, we offer two counterarguments to this potential limitation. First, the reported level of test-taking motivation in our sample prior to the experiment was high (*M* = 4.21, *SD* = .68) and no significant differences in motivation were observed across experimental conditions. We take this as evidence that participants were attending to the task relatively seriously. Second, with respect to the supposed need for high stakes to produce threat effects, one should simply note that the vast majority of studies in *support* of stereotype threat have *also* been conducted in laboratory setting with stakes of similar (or lesser) consequence.

Of the over 100 studies examined in the Nguyen and Ryan (2008) meta-analysis—which does find evidence to support stereotype threat—only one did not use a student sample and none were conducted in operational/high stakes testing scenarios. Thus, although high stakes testing may be a sufficient condition for producing stereotype threat effects, it is not a necessary one and therefore not a strong account of the lack of significant results.

A second related limitation concerns the role of moderators as variables of importance in revealing significant stereotype threat effects. Specifically, it could be argued that an alternative explanation for the non-significant findings is that the student sample did not contain enough individuals susceptible to experiencing threat. For example, Steele, Spencer and Aronson (2002) contend that one's identification with a given performance domain is a significant moderator of stereotype threat such that only highly domain identified individuals are likely to experience performance decrements under threatening condition; thus, if the student sample did not contain enough highly domain identified individuals, we would not have seen the expected threat effects.

While this explanation could account for the attenuation of a significant threat effect, we feel this explanation may be too hasty given that the available evidence for the role of individual difference moderators has been inconclusive or even incompatible in relation to the propositions of the underlying theory. For example, the Nguyen and Ryan (2008) meta-analysis reports that highly domain identified women experienced no statistically significant threat effects whereas moderately identified women were slightly affected. Furthermore, a separate study conducted by the authors of this paper which examined the role of domain identification in the context of subtle stereotype threat manipulations (i.e., small alterations to item content similar to those used here) failed to find evidence of a significant threat by domain identification interaction (Hmurovic, Ryan, Schmitt, & Grand, 2009). In sum, we believe that more substantive theoretical and empirical treatments are needed to determine precisely why, how, when, and to what extent domain identification (and other similar moderators, cf., Nguyen & Ryan, 2008) influence the performance of individuals across a variety of stereotype threat manipulations before conceding the likelihood of this alternative explanation.

A final concern with using the college student sample is that the reading ability of these participants was likely higher than that of a pool of applicants for which tests of mathematical and mechanical ability might typically be administered. Given that the face valid versions of both tests were nearly one grade level higher in reading comprehension (as measured by the Flesch-Kincaid statistic) than their generic counterparts, the construct-irrelevant difficulty of the face valid tests could have been significantly increased had the test taker's levels of reading comprehension been lower (Messick, 1995). While this should not have been a major concern in our sample (the most verbally complex test administered was written at an eighth grade reading level), this consideration reemphasizes the need for test developers to ensure that the required reading level of their evaluative instruments is appropriate. This is particularly important given that improving face validity by adding more job relevant context at the item level can easily increase the required reading proficiency for an instrument.

Conclusion

In conclusion, the present study offers one of the first attempts to question the common belief that improvements in a test's face validity always result in positive gains for test takers. Despite the potential for inducing greater levels of stereotype threat, the face validity manipulation did not significantly affect the psychometric properties of the administered ability tests and appeared to demonstrate mostly beneficial (or non-negative) effects on test performance and perceptions. An important implication from this study is the need to examine manipulations of face validity and stereotype threat relative to perceptions of these characteristics, as the interpretations drawn from each of these approaches differed substantially. The present research supports the notion that test developers should attempt to improve the face validity of cognitive ability exams if possible, though consideration of the content domain of the instrument and the characteristics (demographics, reading comprehension, etc.) of the population for whom the test is intended remain important considerations when investing in such efforts.

References

American Educational Research Association. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Psychological Association.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Arbuckle, J. L. (2003). Amos 5.0.1 [computer software]. Chicago: SmallWaters Corp.

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test-taking. *Personnel Psychology, 43*, 695-716.

Baggozzi, R. P., & Edwards, J. R. (1998). A general approach for representing constructs in organizational research. *Organizational Research Methods*, *1*, 45-87.

Baggozzi, R. P., & Heatherton, T. F. (1994). A general approach to representing multifaceted personality constructs: Application to state self-esteem. *Structural Equation Modeling*, *1*, 35-67.

Bennett, G. K. (2006). *Bennett mechanical comprehension test, Form S*. San Antonio, TX: Harcourt Assessment Incorporated.

Bennett, G., & Cruikshank, R. (1942). Sex differences in the understanding of mechanical problems. *Journal of Applied Psychology, 26*(2), 121-127.

Bornstein, R. F. (1996). Face validity in psychological assessment: Implications for a unified model of validity. *American Psychologist, 51*(9), 983-984.

Bornstein, R. F., Rossner, S. C., Hill, E. L., & Stepanian, M. L. (1994). Face validity and fakability of objective and projective measures of dependency. *Journal of Personality Assessment, 63*, 363-386.

Brutus, S., & Ryan, A. M. (1996). Individual characteristics as determinants of the perceived job relatedness of selection procedures. Unpublished manuscript.

Cascio, W. F. (1987). *Applied psychology in personnel management* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall, Inc.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*(1), 143-159.

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology, 82*(2), 300-310.

Davies, P. G., Spencer, S. J., Quinn, D. M., & Gerhardstein, R. (2002). Consuming images: How television commercials that elicit stereotype threat can restrain women academically and professionally. *Personality and Social Psychology Bulletin, 28*, 12, 1615–1628.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5-18.

Dwight, S. A., & Alliger, G. M. (1997). Reaction to overt integrity test items. *Educational and Psychological Measurement*, 57(6), 937-948.

Elkins, T. J., & Phillips, J. S. (2000). Job context, selection decision outcome, and perceived fairness of selection tests: Biodata as an illustrative case. *Journal of Applied Psychology, 85*(3), 479-484.

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist, 43*(2), 95-103.

Gilliland, S. W. (1994). Effects of procedural and distributive justice on reactions to a selection system. *Journal of Applied Psychology, 79*(5). 691-701.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest, 8*(1), 1-51.

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*(3), 639-683.

Highhouse, S., Lievens, F., & Sinar, E. F. (2003). Measuring attraction to organizations. *Educational and Psychological Measurement, 63*(6), 986-1001.

Hmurovic, J., Ryan, A.M., Schmitt, N., & Grand, J.A. (2009). *Sensitivity or stereotype threat? Effects of gendered test content*. Poster session presented at the meeting of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Holtz, B. C., Ployhart, R. E., & Dominguez, A. (2005). Testing the rules of justice: The effects of frame-of-reference and pre-test validity information on personality test responses and test perceptions. *International Journal of Selection and Assessment, 13*(1), 75-86.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection, and amelioration of adverse impact in personnel selection procedures: Issues, evidence, and lessons learned. *International Journal of Selection and Assessment, 9*, 152-194.

Kray, L. J., Thompson, L., Galinsky, A. (2001). Battle of the sexes: Gender stereotype confirmation and reactance in negotiations. *Journal of Personality and Social Psychology, 80*, 942-958.

Lepore, L., & Brown, R. (2000). Category and stereotype activation: Is prejudice inevitable? In C. Stangor (Ed.), *Stereotypes and prejudice* (pp. 119-137). Philadelphia, PA: Psychology Press.

Levy, B. (1996). Improving memory in old age by implicit self-stereotyping. *Journal of Personality & Social Psychology, 71*, 1092-1107.

Levy, S. R., Stroessner, S. J., & Dweck, C. S. (1998). Stereotype formation and endorsement: The role of implicit theories. *Journal of Personality and Social Psychology, 74*(6), 1421-1436.

Lievens, F., De Corte, W., & Schollaert, E. (2008). A closer look at the frame-of-reference effect in personality scale scores and validity. *Journal of Applied Psychology, 93*(2), 268-279.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessments: Expectations and validation criteria. *Educational Researcher, 13,* 15-31.

McFarland, L. A., Lev-Arey, D. M., Ziegert, J. C. (2003). An examination of stereotype threat in a motivational context. *Human Performance, 16*(3), 181-205.

McGrath, J.E. (1986). Dilematics: The study of research choices and dilemmas. In J.E. McGrath, J. Martin, & R.A. Kulka (Eds.), *Judgment calls in research* (pp. 69-101). Beverly Hills, CA: Sage.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement, 7*(2), 191-205.

Muchinsky, P. M. (2004). Mechanical aptitude and spatial ability testing. In J. C. Thomas (Ed.), *Comprehensive handbook of psychological assessment: Vol. 4. Industrial and organizational assessment* (pp. 21-33). Hoboken, NJ: John Wiley & Sons, Inc.

Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement, 22*(4), 287-293.

Nguyen, H.-H., O'Neal, A., & Ryan, A. M. (2003). Relating test-taking attitudes and skills and stereotype threat effects to the racial gap in cognitive ability test performance. *Human Performance, 16*(3), 261-293.

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

O*NET. (2004). *Summary report for 49-9042.00 – Maintenance and repair workers, general*. Retrieved October 22, 2007 from http://online.onetcenter.org/link/summary/49-9042.00.

Ployhart, R. E., Schneider, B., & Schmitt, N. (2006). *Staffing organizations: Contemporary practice and theory* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Ployhart, R. E., Ziegert, J. C., & McFarland, L. A. (2003). Understanding racial differences on cognitive ability tests in selection contexts: An integration of stereotype threat and applicant reactions research. *Human Performance, 16*(3), 231-259.

Robie, C., Schmit, M. J., Ryan, A. M., & Zickar, M. J. (2000). Effects of item context specificity on the measurement equivalence of a personality inventory. *Organizational Research Methods, 3*(4), 348-365.

Sackett, P. R. (2003). Stereotype threat in applied selection settings: A commentary. *Human Performance, 16*, 295-309.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High stakes testing in employment, credentialing, and higher education. *American Psychologist, 56*, 302–318.

Sackett, P. R., Hardison, C. M., Cullen, M. J. (2005). On interpreting research on stereotype threat and test performance. *American Psychologist, 60*(3), 271-272.

Seymour, R. T. (1988). Why plaintiffs' counsel challenge tests, and how they can successfully challenge the theory of "Validity Generalization." *Journal of Vocational Behavior, 33*(3), 331-364.

Shotland, A., Alliger, G. M., & Sales, T. (1998). Face validity in the context of personnel selection: A multimedia approach. *International Journal of Selection and Assessment, 6*(2), 124-130.

Smither, J. W., Reilly, R. R., Millsap, R. E., Pearlman, K., & Stoffey, R. W. (1993). Applicant reactions to selection procedures. *Personnel Psychology, 46*(1), 49-76.

Spence, J. T., Helmreich, R., & Stapp, J. (1975). Ratings of self and peers on sex role attributes and their relation to self-esteem and conceptions of masculinity and femininity. *Journal of Personality and Social Psychology, 32*(1), 29-39.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4-28.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*(6), 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*(5), 797-811.

Steele, C.M., & Davies, P.G. (2003). Stereotype threat and employment testing: A commentary. *Human Performance, 16*(3), 311-326.

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. *Advances in Experimental Social Psychology, 34*, 379-400.

Tett, R. P, Anderson, M. G., Ho, C., Yang, T. S., Huang, L., & Hanvongse, A. (2006). Seven nested questions about faking on personality tests. In R. Griffith & M. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 43-83). Greenwich CT: Information Age Publishing.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*, 139-158.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70.

Walton, G. M., & Cohen, G. L. (2003). Stereotype lift. *Journal of Experimental Social Psychology, 39*, 456-467.

Wiechmann, D., & Ryan, A. M. (2003). Reactions to computerized testing in selection contexts. *International Journal of Selection and Assessment, 11*(2-3), 215-229.

Whitney, D. J., Diaz, J., Mineghino, M. E., & Powers, K. (1999). Perceptions of overt personality-based integrity tests. *International Journal of Selection and Assessment*, *7*(1), 35-45.

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology*, *89*(5), 696-716.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley.

Footnotes

[1] Although the next step in the MGCFA procedure is typically to check for invariance of error variances, Vandenberg and Lance (2000) note this is not a necessary step if the goal is to establish equivalence in latent means; thus the results of this step are not presented. However, equivalence in error variances was examined and no significant change in model fit was found.

[2] Given that we used a fully crossed between-subjects design, our data also permitted an examination of the interactive effects of the threat activation methods to assess whether the combination of threat cues (e.g., comparing women in the face valid/no explicit threat condition to women in the generic/no explicit threat condition, thus partialling out the effects of explicit activation, etc.) produced differential effects on women's test performance. However, the results comparing the cell means (cf., Table 2) did not reveal a significantly different pattern of findings from those obtained with the marginal means. That is, for Hypothesis 2b, the face validity by explicit threat interaction term failed to reach statistical significance indicating that female's performance on each test was not significantly different across study cells. For Hypothesis 3b, the pattern of effect size differences (d_{diff}) between males and females indicated that the performance differences were again largest in the condition where threat was not activated (generic/no explicit threat) than when either implicit threat (face valid/no explicit threat) or explicit threat (generic/explicit threat) were present. These results are available from the first author upon request.

Table 1

Means, Standard Deviations, and Correlations of Study Variables

Variable	*M*	*SD*	1	2	3	4	5	6	7	8
1. FV[a]	.50	.50	--							
2. Stereotype Threat[b]	.49	.50	.01	--						
3. Math Ability	14.97	5.81	.01	-.07	(.83)					
4. Mechanical Ability	17.49	4.10	**-.20**	-.05	**.42**	(.76)				
5. Perceived FV (Manipulation check)	3.51	.68	**-.17**	.01	**.21**	**.13**	(.87)			
6. Test-taking Motivation	4.21	.68	-.04	.02	**.20**	.07	**.11**	(.96)		
7. Self-assessed Performance	3.19	.85	.00	-.04	**.51**	**.23**	**.22**	**.29**	(.82)	
8. Test Ease	2.57	.78	.01	-.05	**.36**	**.28**	.04	.04	**.44**	(.89)
9. Pursuit Intentions	2.85	.94	-.01	-.02	**.13**	.10	.03	**.15**	**.20**	**.23**
10. Recommendation Intentions	2.92	.86	-.01	-.03	**.24**	**.20**	**.24**	**.19**	**.39**	**.35**
11. Job Attractiveness	2.47	.96	-.04	-.05	**.17**	.10	.03	**.15**	**.21**	**.29**
12. Procedural Fairness	3.37	.85	.00	.01	**.29**	**.15**	**.50**	**.20**	**.29**	**.19**
13. Perceived Predictive Validity	2.80	.84	**-.17**	.09	**.13**	**.13**	**.49**	**.13**	**.19**	.08
14. Perceived Stereotype Threat (Mechanical)	2.89	.74	.06	**-.25**	-.10	**-.25**	-.04	-.01	**-.11**	**-.16**
15. Perceived Stereotype Threat (Math)	2.70	.77	.08	**-.23**	-.09	**-.22**	-.08	-.08	**-.14**	**-.16**

Table 1 (Cont.)

9	10	11	12	13	14	15
(.87)						
.57	(.88)					
.80	**.48**	(.92)				
.15	**.44**	**.15**	(.89)			
.05	**.27**	**.11**	**.51**	(.86)		
.00	-.05	-.01	.00	.07	(.77)	
.00	-.08	.01	-.04	.01	**.83**	(.85)

Note. Numbers in bold represent significant correlations at p < .05 or better. α coefficients are presented along the diagonal where applicable. FV = Face validity.
[a]Dummy coded variable (0 = face valid condition, 1 = generic condition). [b]Dummy coded variable (0 = explicit stereotype activation, 1 = no stereotype activation)

Table 2

Means (Standard Deviations) and Cell Sizes for Males and Females on the Mathematical and Mechanical Ability Tests by Condition

	Females		Males	
	Stereotype Threat	No Stereotype Threat	Stereotype Threat	No Stereotype Threat
Face Valid	*Math* = 14.52 (5.38) *Mech* = 17.41 (4.15) *n* = 61	*Math* = 14.88 (5.82) *Mech* = 16.96 (3.55) *n* = 52	*Math* = 16.82 (6.55) *Mech* = 20.68 (2.65) *n* = 28	*Math* = 14.00 (6.09) *Mech* = 20.06 (4.71) *n* = 33
Generic	*Math* = 14.76 (5.70) *Mech* = 16.10 (3.37) *n* = 62	*Math* = 13.51 (4.77) *Mech* = 15.34 (3.29) *n* = 61	*Math* = 17.25 (6.23) *Mech* = 19.00 (4.23) *n* = 24	*Math* = 17.46 (6.58) *Mech* = 19.17 (4.21) *n* = 24

Table 3

Measurement Invariance of Ability Tests across Women in Face Valid (n = 113) and Generic (n = 123) Conditions (Hypothesis 1)

Models	χ^2	*df*	Model Comparison	$\Delta\chi^2$	Δdf	RMSEA	CFI	SRMR
M1: Free factor loadings, intercepts and error variances	86.27	68	--	--	--	.03	.97	.05
M2: Fixed factor loadings; free intercepts and error variances	93.08	76	M1 vs. M2	6.81	8	.03	.97	.06
M3: Fixed factor loadings and intercepts; free error variances	138.74	86	M2 vs. M3	45.66*	10	.05	.91	.06
M4: Fixed factor loadings and intercepts; free intercepts for MechParcel 1 and 2 and error variances	105.50	84	M2 vs. M4	12.42	8	.03	.96	.06

Note. RMSEA = root mean squared error of approximation; CFI = comparative fit index; SRMR = standardized root mean square residual.

*p < .05

Table 4

Summary of Regression Analyses for Hypotheses 2a and 2b

Dependent Variable	Predictor	β	R^2	ΔR^2
Math Test Performance	(Control Variables[a])	--	.32*	
	FV[b]	.05		
	Explicit threat[c]	-.04	.32*	.00
	FV x Explicit threat	-.12	.33*	.01
Mechanical Test Performance	(Control Variables)	--	.29*	
	FV	-.13*		
	Explicit threat	-.06	.32*	.02*
	FV x Explicit threat	-.09	.32*	.00

Note. n = 236 for both regression models. FV = Face validity.
[a] Test order, test time, general cognitive ability (ACT score). [b] Dummy coded variable (0 = face valid condition, 1 = generic condition). [c] Dummy coded variable (0 = explicit stereotype activation, 1 = no stereotype activation)
*p < .05

Table 5

Summary of Regression Analyses for Hypotheses 3a and 3b

Dependent Variable	Predictor	*β*	R^2	ΔR^2
Math Test Performance	(Control Variables[a])	--	.30*	
	Gender[b]	.16*		
	FV[c]	.11*		
	Explicit threat[d]	-.06	.33*	.03*
	Gender x FV	.15*		
	Gender x Explicit Threat	-.04		
	FV x Explicit threat	-.04	.35*	.02
	Gender x FV x Explicit Threat	.14	.35*	.00
Mechanical Test Performance	(Control Variables)	--	.23*	
	Gender	.35*		
	FV	-.12*		
	Explicit threat	-.03	.37*	.14*
	Gender x FV	.02		
	Gender x Explicit Threat	.07		
	FV x Explicit threat	-.03	.37*	.00
	Gender x FV x Explicit Threat	.08	.37*	.00

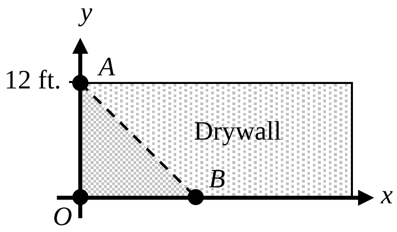
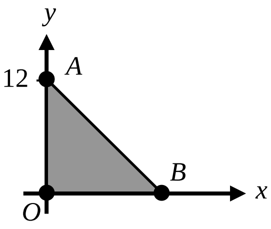
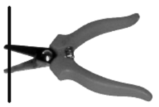
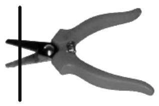
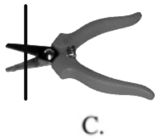
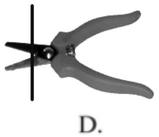
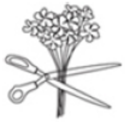
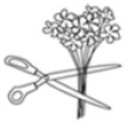
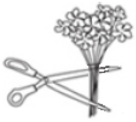
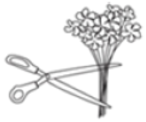
Note. *n* = 345 for both regression models. FV = Face validity.

[a]Test order, test time, general cognitive ability (ACT score). [b]Dummy coded variable (0 = female, 1 = male). [c]Dummy coded variable (0 = face valid condition, 1 = generic condition).

[d]Dummy coded variable (0 = explicit stereotype activation, 1 = no stereotype activation)

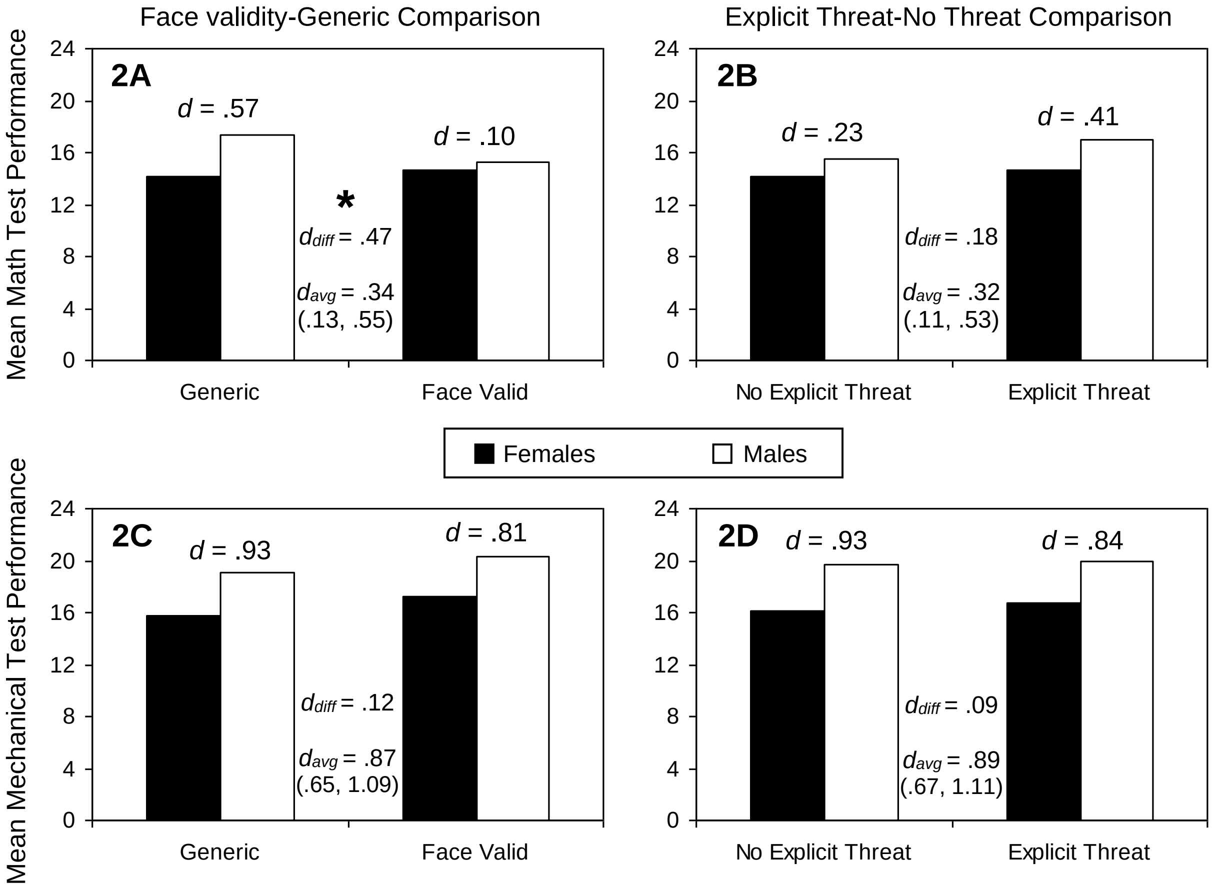
*$p < .05$

Figure 1. Example items from the face valid and generic versions of the mathematical and mechanical ability tests.

	Face Valid Items	Generic Items
Mathematical Items	 Note: Figure not drawn to scale. A carpenter is cutting a triangular notch from a large piece of drywall as shown above. If the slope of the cut (line AB) is -.75, what is the area of the piece of drywall the carpenter is removing (**Δ***ABO*)? A. 54 square feet B. 72 square feet C. 96 square feet D. 108 square feet E. 192 square feet	 Note: Figure not drawn to scale. In the figure above, if the slope of the line (AB) is -.75, what is the area of **Δ***ABO*?† A. 54 B. 72 C. 96 D. 108 E. 192
Mechanical Items	 A.  B.  C.  D. Which picture shows where the wire clippers should be placed to cut a wire with the least amount of force?	 A.  B.  C.  D. In which of the pictures would it be easiest to cut the flower stems with the scissors?

†Item adapted from Black, C., & Anestis, M. (2008). *McGraw Hill's SAT: 2008 edition*. New York: McGraw-Hill.

Figure 2. Gender effect sizes (*d*s) across test version and threat activation type for math and mechanical ability test performance.


Face validity-Generic Comparison
Explicit Threat-No Threat Comparison
Mean Math Test Performance
Mean Mechanical Test Performance
2A
d = .57
d = .10
*
d_diff = .47
d_avg = .34
(.13, .55)
Generic
Face Valid
2B
d = .23
d = .41
d_diff = .18
d_avg = .32
(.11, .53)
No Explicit Threat
Explicit Threat
■ Females
□ Males
2C
d = .93
d = .81
d_diff = .12
d_avg = .87
(.65, 1.09)
Generic
Face Valid
2D
d = .93
d = .84
d_diff = .09
d_avg = .89
(.67, 1.11)
No Explicit Threat
Explicit Threat