Journal of Personality and Social Psychology
2005, Vol. 89, No. 5, 696–716

0022-3514/05/$12.00 DOI: 10.1037/0022-3514.89.5.696

Stereotype Threat and Group Differences in Test Performance: A Question of Measurement Invariance

Jelte M. Wicherts, Conor V. Dolan, and David J. Hessen
University of Amsterdam

Studies into the effects of stereotype threat (ST) on test performance have shed new light on race and sex differences in achievement and intelligence test scores. In this article, the authors relate ST theory to the psychometric concept of measurement invariance and show that ST effects may be viewed as a source of measurement bias. As such, ST effects are detectable by means of multigroup confirmatory factor analysis. This enables research into the generalizability of ST effects to real-life or high-stakes testing. The modeling approach is described in detail and applied to 3 experiments in which the amount of ST for minorities and women was manipulated. Results indicate that ST results in measurement bias of intelligence and mathematics tests.

Keywords: measurement invariance, stereotype threat, ethnic differences, sex differences, test performance

> The greatest social benefit will come from applied psychology if we can find for each individual the treatment to which he can most easily adapt. This calls for the joint application of experimental and correlational methods. (Cronbach, 1957, p. 679)

Recent developments in experimental social psychology concerning the effects of stereotypes on test performance have contributed to the understanding of the nature of race and sex differences in achievement and intelligence test scores. Specifically, the theory of stereotype threat (Steele, 1997) states that stereotypes concerning the ability of groups (e.g., women are bad at mathematics) can have an adverse impact on test performance of members of such groups, particularly in those who identify strongly with the domain of interest (e.g., female math students). Considering the widespread use of achievement and intelligence tests in college admission and personnel selection, and the high stakes involved in their use, stereotype threat effects on test performance may have serious personal and social consequences. There is general agreement on the importance of fair, unbiased assessment in the sense that individual latent abilities should be measured validly and accurately. This means that measurements of ability should not depend on group membership based on, for instance, ethnicity or sex. Therefore, the absence of measurement bias with respect to groups (i.e., measurement invariance) is an essential aspect of valid measurement (e.g., Millsap & Everson, 1993). Both research into stereotype threat and research into measurement invariance are aimed at disentangling measurement artifacts related to group membership from individual differences in the construct that a particular test is supposed to measure (e.g., latent mathematics ability). In the current article, our aim is to explicitly relate stereotype threat to the concept of measurement invariance and to show that stereotype threat effects on test performance may be viewed as a source of measurement bias.

This conceptualization of stereotype threat effects has statistical as well as practical advantages. It gives rise to an analytical framework in which individual and group differences in latent abilities and (experimental) stereotype threat effects on test performance can be modeled simultaneously. Of more practical importance is the fact that tests for measurement invariance with respect to groups can shed light on the degree to which stereotype threat plays a role in real-life and high-stakes settings. This provides a means to study the effects of stereotype threat in settings in which it is ethically and pragmatically difficult to manipulate the debilitating effects of stereotype threat on test performance (Cullen, Hardison, & Sackett, 2004; Sackett, 2003; Steele & Davies, 2003; Steele, Spencer, & Aronson, 2002).

Below, we first discuss some methodological and statistical issues concerning experimental tests of stereotype threat effects on test performance. Next, we relate the effects of stereotype threat to measurement invariance and discuss how such effects can be detected by means of multigroup confirmatory factor analysis (MGCFA). Finally, we illustrate this approach by analyzing the results of three experiments in which the effects of stereotype threat on the test performance of stigmatized groups were investigated.

Investigating Stereotype Threat Effects

The experimental paradigm, which is used to study the effect of stereotype threat on test performance, usually involves the comparison of existing groups (e.g., Blacks and Whites) and the

Jelte M. Wicherts, Conor V. Dolan, and David J. Hessen, Department of Psychology, University of Amsterdam, Amsterdam, the Netherlands.

The preparation of this article was supported by a grant from the Netherlands Organization for Scientific Research. We are indebted to the indefatigable women of OP4425 for assistance in collecting the data of Study 1. We thank Hannah-Hanh Nguyen and Laurie O'Brien for kindly providing the descriptive statistics of their studies. We also thank Patricia Rodriguez Mosquera for her comments on a draft of this article.

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

manipulation of stereotype threat. The latter is accomplished, for instance, by labeling a test as either diagnostic or nondiagnostic for the stereotyped ability (e.g., Steele & Aronson, 1995, Study 2), or by asking for biographical information either prior to or after completion of the test (e.g., Steele & Aronson, 1995, Study 4). Stereotype threat is expected to negatively affect test performance of stigmatized groups but to have no (or a small positive; see Walton & Cohen, 2003) effect on test performance of nonstigmatized groups. Stereotype threat theory thus predicts an interaction between group and threat manipulation.

Generalizability of Stereotype Threat

Within laboratory experiments, stereotype threat has been found to depress scores on various achievement and intelligence tests, in diverse stigmatized groups (Steele et al., 2002). The extent to which stereotype threat generalizes to test settings outside the laboratory is an important issue. Only few experimental studies have looked into the debilitating effects of stereotype threat on test performance in test settings high in ecological validity, and/or settings with consequential test outcomes. Stricker and Ward (2004) conducted two field studies within an actual high-stakes test situation but were unable to replicate the strong negative effects of asking for biographical information prior to taking a test (i.e., group prime) on minority and female test performance (cf. Steele & Aronson, 1995). In addition, three recent laboratory experiments addressed the effects of stereotype threat on Blacks' test performance in a job selection context (McFarland, Lev Arey, & Ziegert, 2003; Nguyen, O'Neal, & Ryan, 2003; Ployhart, Ziegert, & McFarland, 2003). In these studies, test-taking motivation was enhanced by the promise of financial rewards for high test scores. Despite the use of manipulations with well-established effects (i.e., race prime and test diagnosticity), the debilitating effects of stereotype threat on minority test performance were generally absent. Sackett (2003) suggested that these results imply that the generality of stereotype threat effects to (motivational) job selection contexts is limited. Along similar lines, Stricker and Ward (2004) suggested that their studies indicate that high test stakes appear to be capable of overriding the negative effects of stereotype threat on test performance.

From a theoretical point of view, however, the internal validity of these real-life or contextualized experiments appears questionable. Steele and colleagues argued that stereotype threat probably always occurs within such settings because of features that have been shown to elicit stereotype threat in the laboratory (Steele & Davies, 2003; Steele et al., 2002). For instance, promising incentives or placing a test in a selection context makes a test diagnostic for the stereotyped ability, thereby triggering stereotype threat even within control conditions. Heightening stereotype threat by means of explicit test diagnosticity or group prime then fails to depress test performance of stigmatized groups much further, resulting in ineffective stereotype threat manipulations (Steele & Davies, 2003; Steele et al., 2002). In that respect, stereotype threat theory predicts that stereotype threat studies, which are high in ecological validity, are low in internal validity, and vice versa. More important, whereas inductive reasoning leads one to expect that most real-life test settings do evoke stereotype threat, empirically the question of generalizability appears hard to answer (Steele et al., 2002).

Analyzing Stereotype Threat Effects

Given the pragmatic and ethical problems of experimentation within real-life settings, correlational methodology (e.g., regression analysis) may be used to investigate the presence of stereotype threat on actual achievement tests. Osborne (2001) reasoned that stereotype threat effects may be mediated by anxiety (cf. Blascovich, Spencer, Quinn, & Steele, 2001). He found that the racial gap and, to a lesser extent, the gender gap on several achievement tests in the High School and Beyond Study were partly mediated by self-reported anxiety, which supports the notion that stereotype threat affected test performance. Cullen et al. (2004) proposed that the strong identification of high-ability persons with the domain of interest (cf. Steele, 1997) renders them more sensitive to stereotype threat (Aronson et al., 1999). Cullen et al. reasoned that if stereotype threat affects test performance of stigmatized groups on a predictor (e.g., SAT, formerly known as the Scholastic Aptitude Test), then this differential sensitivity to stereotype threat would lead to group-specific and nonlinear relations between the affected predictor and criteria that are supposedly unaffected by stereotype threat, such as job performance or grade points of classes unrelated to stereotypes. However, Cullen et al. found neither prediction bias nor any nonlinear effects, and they concluded that stereotype threat effects on the predictors (SAT and Armed Services Vocational Aptitude Battery) were small or nonexistent.

These seemingly inconsistent results may be due to the strong assumptions underlying the use of such regression approaches. For instance, Cullen et al. (2004) had to assume the absence of group differences on academic criteria (cf. the underperformance phenomenon; Steele, 1997), whereas Osborne (2001) rightly expressed some concern about the causal link involved. Moreover, these correlational studies address the effects of stereotype threat on test performance in an indirect manner. It is well established that group differences in prediction (i.e., prediction bias) do not necessarily imply that measurements are biased with respect to groups, and vice versa (Millsap, 1997a).

Measurement Models

The indirectness of these regression approaches can be avoided by adopting measurement models that explicitly relate test scores to the latent constructs that are supposed to underlie those test scores. Instead of the latent abilities, stereotype threat affects the test scores in a group-specific manner. As we see below, a comparison of stigmatized and nonstigmatized groups with respect to the test scores–construct relationship (i.e., test for measurement invariance) allows for a direct study of the presence of stereotype threat effects within a particular test situation.

An additional advantage of using measurement models is that they can be used to analyze experimental data (cf. Donaldson, 2003), thereby overcoming some difficulties associated with traditional use of analysis of variance (ANOVA) within stereotype threat experiments. The groups under investigation in such studies are expected to differ considerably with respect to the latent ability that is supposed to underlie the dependent variable(s) (i.e., test scores). This may give rise to analytical problems because of preexisting group differences in the average or variability of latent ability (e.g., gender differences in math variability; Hedges &

Nowell, 1995). In numerous stereotype threat studies, prior test scores (e.g., SAT) and analysis of covariance (ANCOVA) are used to equate groups for mean differences in ability. However, several expectations derived from stereotype threat theory do not sit well with the assumptions underlying the traditional use of ANCOVA (Wicherts, 2005; Yzerbyt, Muller, & Judd, 2004). For instance, stereotype threat may lower the regression weight of the dependent variable on the covariate in the stereotype threat condition, which violates regression weight homogeneity over all experimental cells (Wicherts, 2005). The use of statistical methods that differentiate between the construct (i.e., latent ability) and the measurement of that construct circumvents such problems. More important, measurement models provide the necessary ways to test for measurement invariance.

Measurement Invariance

Measurement invariance revolves around the issue of how groups differ in the way the measurement of a psychological construct (e.g., mathematics test score) is related to that construct (e.g., mathematical ability). Measurement invariance means that measurement bias with respect to groups is absent (Lubke, Dolan, Kelderman, & Mellenbergh, 2003a, 2003b; Meredith, 1993). Below, we explain measurement invariance conceptually in relation to stereotype threat. We first look at the formal definition of measurement invariance (Mellenbergh, 1989), which is expressed in terms of the conditional distribution of manifest test scores Y [denoted by $f(Y \mid)$]. Measurement invariance with respect to v holds if:

$$f(Y|\eta, v) = f(Y|\eta), \quad (1)$$

(for all Y, η, v), where η denotes the scores on the latent variable (i.e., latent ability) underlying the manifest random variable Y (i.e., the measured variable), and v is a grouping variable, which defines the nature of groups (e.g., ethnicity, sex). Note that v may also represent groups in experimental cells such as those that differ with respect to the pressures of stereotype threat. Equality 1 holds if, and only if, Y and v are conditionally independent given the scores on the latent construct η (Lubke et al., 2003b; Meredith, 1993).

One important implication of this definition is that the expected value of Y given η and v should equal the expected value of Y given only η. In other words, if measurement invariance holds, then the expected test score of a person with a certain latent ability (i.e., η) is independent of group membership. Thus, if two persons of a different group have exactly the same latent ability, then they must have the same (expected) score on the test. Suppose v denotes sex and Y represents the scores on a test measuring mathematics ability. If measurement invariance holds, then test scores of male and female test takers depend solely on their latent mathematics ability (i.e., η)[1] and not on their sex. Then, one can conclude that measurement bias with respect to sex is absent and that manifest test score differences in Y correctly reflect differences in latent ability between the sexes.

However, the situation changes when stereotype threat has an impact on test performance. Suppose v represents two groups (e.g., Blacks and Whites) that differ with respect to stereotypes that concern Y (e.g., intelligence tests). If stereotype threat directly affects (i.e., lowers) the observed scores (i.e., Y) in the Black group (or in a subsample of this group), then measurement invariance is violated. The reason for this is that conditioning on the latent construct (i.e., latent ability) does not remove all group differences in Y because of the debilitating effects of stereotype threat on Y, which are limited to the Black group. This becomes particularly clear if one images a Black test taker with a particular latent ability, who, because of stereotype threat, underperforms in comparison with a White test taker with the same latent ability. Clearly, the relationship between test score and latent ability now depends on group membership, and the requirements for measurement invariance no longer hold. Therefore, stereotype threat effects are by definition a source of measurement bias. Conversely, if measurement invariance holds in a particular group comparison, stereotype threat does not play a differential role in test score differences between those groups, because then test score differences rightly reflect group differences in the latent construct.

The definition of measurement invariance is quite general (Mellenbergh, 1989). It does not depend on the kind of test, selection variable, or the size of group differences in latent ability. Although measurement invariance may be investigated by many methods (Millsap & Everson, 1993; Raju, Laffitte, & Byrne, 2002) that use different types of measurement models (e.g., item response models), we restrict our attention to the confirmatory factor model. We now present this model, relate it to measurement invariance, and show how stereotype threat may result in measurement bias. After that, we investigate in three studies whether experimental stereotype threat effects indeed lead to measurement bias.

MGCFA

Here we describe the measurement model (i.e., MGCFA) in a nontechnical fashion, restricting our attention to the one factor case and assuming multivariate normality throughout. Appendix A contains a more technical and more general presentation of the model (see also Bollen, 1989; Dolan, 2000; Dolan, Roorda, & Wicherts, 2004; Lubke et al., 2003a). The confirmatory factor model is essentially a linear regression model in which scores on several indicators (i.e., subtest scores) are regressed upon scores on the latent (i.e., unobserved) construct η. Like in ordinary regression, the model includes for each indicator the following measurement parameters: a regression weight or factor loading (expressed by the symbol λ), a residual term, and an intercept. The residual term of an indicator is expressed by the symbol ε, and contains both random measurement error and specific factors tapped by that particular indicator (i.e., all uncommon sources of variance; Meredith & Horn, 2001). In most applications of confirmatory factor analysis (e.g., one-group studies), the regression intercept is uninformative and is not modeled. However, we are also interested in studying between-groups differences in means. Therefore, we add the mean structure to the analysis, which is accomplished by incorporating an intercept term for each indicator, expressed by ν (Sörbom, 1974). The extension to multiple groups enables tests of specific hypotheses concerning between-groups differences in measurement parameters (i.e., measurement bias) and between-

[1] However, measurement invariance with respect to one selection variable does not necessarily imply measurement invariance with respect to another selection variable (however, see Lubke et al., 2003b).

groups differences in the parameters that describe the distribution of the common factor within each group (i.e., group differences in mean latent ability). The simultaneous analysis of mean and mean structure[2] provide a test of measurement invariance, or *strict factorial invariance*, as it is denoted in this context (Meredith, 1993).

The model for subtest score Y_1 of a person j in group (or condition) i is as follows:

$$Y_{1ij} = \nu_{1i} + \lambda_{1\eta i}\eta_{ij} + \varepsilon_{1ij}. \quad (2)$$

Suppose we have four subtests. Of course, the latent ability score η_{ij} of person j is the same for all subtests, so we can conveniently arrange the expressions using vector notation (e.g., Bollen, 1989):

$$\begin{bmatrix} Y_{1ij} \\ Y_{2ij} \\ Y_{3ij} \\ Y_{4ij} \end{bmatrix} = \begin{bmatrix} \nu_{1i} \\ \nu_{2i} \\ \nu_{3i} \\ \nu_{4i} \end{bmatrix} + \begin{bmatrix} \lambda_{1\eta i} \\ \lambda_{2\eta i} \\ \lambda_{3\eta i} \\ \lambda_{4\eta i} \end{bmatrix} \times [\eta_{ij}] + \begin{bmatrix} \varepsilon_{1ij} \\ \varepsilon_{2ij} \\ \varepsilon_{3ij} \\ \varepsilon_{4ij} \end{bmatrix}. \quad (3)$$

This, in turn, is more parsimoniously expressed by the following matrix notation:

$$Y_{ij} = \nu_i + \Lambda_i\eta_{ij} + \varepsilon_{ij}. \quad (4)$$

Except for the difference in notation, Equations 3 and 4 are identical. For example, in Equation 4, ν_i is a four-dimensional vector containing the measurement intercepts and Λ_i is a 4×1 matrix containing the factor loadings of group i. Equation 4 presents a model for the observations. To obtain estimates of the parameters in this model, we fit the observed covariance matrices and mean vectors to the implied (by Equation 4) covariance matrices and mean vectors (cf. Appendix A). The parameters of interest are the factor loadings (Λ_i), the vector of intercepts (ν_i), the variances of the residuals—incorporated in a matrix denoted Θ_i—and the means and variances of the common factor scores in group i, denoted by α_i and Ψ_i, respectively. In fitting the model, we introduce two types of constraints: identifying constraints, which are required in all confirmatory factor analyses (e.g., scaling; Bollen, 1989), and substantive constraints, which relate specifically to the issue of measurement invariance (Meredith, 1993). As we explain next in a two-group context, the latter concern the factor loadings, intercepts, and residual variances.

Consider the top half of Figure 1. Here we see the regression lines for subtest Y_1 in two groups. The factor loading gives the slope of this line (for each increment Δ of latent ability η, the expected test score changes by $\Delta\eta$ times λ), and the intercept ν gives the point of Y_1 associated with the point $\eta = 0$. Also depicted are the normally distributed residuals in each group. Note that the residual variances appear equal in both groups. As can be seen, the regression slopes (i.e., factor loadings) are also equal in both groups. However, the intercepts differ over groups. This has serious consequences. Namely, for each possible latent factor score, the expected value on the Test Y_1 is higher for members of Group 1 than for members of Group 2.[3] Clearly, this violates measurement invariance with respect to both groups. Hence, the equality of measurement intercepts (i.e., $\nu_{11} = \nu_{12}$) is an essential requirement for measurement invariance (cf. Meredith, 1993). The reader may have already guessed a possible source for such an intercept difference between groups: the uniform (i.e., irrespective of latent ability) depression of test scores due to stereotype threat in Group 2.

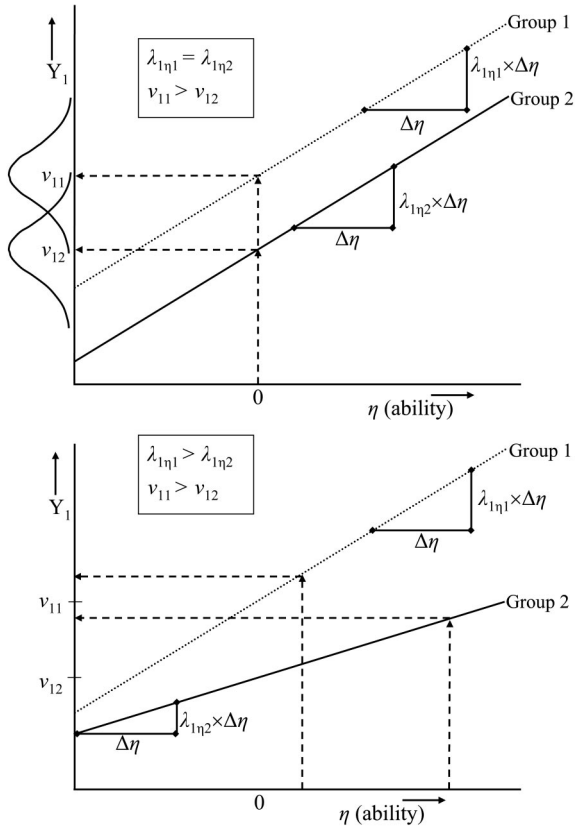

Figure 1. Top half: Regression lines of scores on Subtest Y_1 in two groups with different intercepts. Bottom half: Regression lines of scores on Subtest Y_1 in two groups with different intercepts and different factor loadings.

The bottom half of Figure 1 displays another two-group scenario. Here, the regression lines for both groups again show different intercepts. In addition, the slope of the regression line in Group 2 now differs from the slope in Group 1. Specifically, the factor loading in Group 2 is lower (i.e., $\lambda_{1\eta2} < \lambda_{1\eta1}$). This means that in Group 2, the test scores do not measure latent ability as well as in Group 1. Again, given a particular latent factor score, the expected test score depends on group membership. Even worse, the negative effect of "being" a Group 2 member now depends on the particular latent ability level. Higher ability scores result in more bias than lower ability scores. As is graphically depicted by the dashed arrows, it is even conceivable that a member of Group 2 with a fairly high ability score has an expected test score below

[2] We are also interested in and should allow for possible differences in variances between the groups. For that reason, in MGCFA, covariance matrices are analyzed instead of correlation matrices.

[3] Note the resemblance of this picture to what Steele (1997, p. 626) called the *parallel lines phenomenon* when he referred to the academic underperformance of Black college students in comparison with White college students with equal standardized test scores. The differences lie in that Steele's predictor was a standardized test score and his criterion was 1st-year GPA, whereas our predictor is the latent ability score, and the criterion is the test score.

that of someone in Group 1 who has a considerably lower ability. Clearly, for measurement invariance to hold between groups, factor loadings should be equal across groups (i.e., $\lambda_{1\eta 1} = \lambda_{1\eta 2}$). Note that a depressed factor loading could be due to stereotype threat affecting test performance in Group 2 in a nonuniform manner. Again, the lowering of the intercept may be viewed as a main effect for stereotype threat. Moreover, the lowering of a factor loading in Group 2 can be interpreted as an interaction effect between stereotype threat and latent ability on test performance. The latter may occur because domain identification is known to heighten stereotype threat effects, and domain identification may be strongly related to latent ability (Cullen et al., 2004; Steele, 1997). In such a scenario, higher ability persons suffer more under stereotype threat, resulting in a depressive effect on the factor loading.

We have presented a graphic exposition of why factor loadings and measurement intercepts need to be invariant for measurement invariance to hold. In fact, under measurement invariance, the regression lines of each group coincide. If so, then the expected value of the test scores depends solely on latent ability, regardless of group membership. An additional requirement for strict factorial invariance is that residual variances need to be invariant. This is because residual variances contain all uncommon sources of variance. Larger residual variance in one group means less reliable measurement. Moreover, added residual variance may also be due to stereotype threat variance. Meredith (1993) has provided a rigorous statistical discussion of why group-invariant factor loadings (Λ), residual variances (Θ), and intercepts (ν) are essential requirements for strict factorial invariance. Indeed, if measurement invariance holds, as defined above (Equation 1), then these equality constraints should hold to reasonable approximation (Meredith, 1993; Millsap, 1997b).

The Stereotype Threat Factor

To better understand the specific effects of stereotype threat on measurement parameters, we find it convenient to imagine the presence of an additional common factor (denoted by σ), which incorporates all the mediating effects of stereotype threat on test performance. Such an additional stereotype threat "factor" is neither measured nor modeled, but it still affects test performance in a manner that is restricted to the stigmatized group, resulting in group-specific changes of measurement parameters. Hence, constraining measurement parameters of a group under stereotype threat to group(s) without such effects (i.e., nonstigmatized group and/or control condition) would demonstrate a violation of strict factorial invariance. It is well established that stereotype threat specifically affects performance on the more difficult tasks (Blascovich et al., 2001; O'Brien & Crandall, 2003; Quinn & Spencer, 2001; Spencer, Steele, & Quinn, 1999; Steele et al., 2002). Therefore, we expected the effects to be subtest specific and mostly related to the most difficult subtests in a test battery.

Figure 2 displays the common factor model within a group (in a particular setting), where stereotype threat affects the scores on subtest Y_4 (conceivably a particularly difficult subtest). As we show in Appendix B, such a stereotype threat effect results in a lowering of the measurement intercept of the affected subtest (cf. Figure 1, top half). In addition, if stereotype threat effects vary over persons within this group, perhaps because of individual differences in domain identification or group identification, then

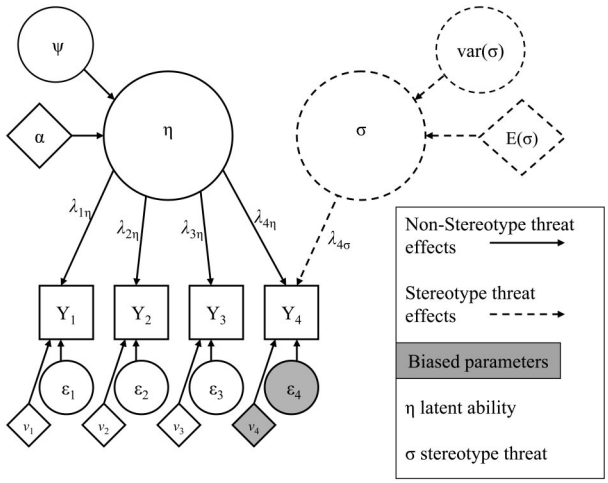

Figure 2. The common factor model in which stereotype threat affects the scores on Subtest Y_4 but not on Subtests Y_1–Y_3. var = variance; E = expected value.

the variance due to the unmeasured stereotype threat factor results in an increase of the residual variance of subtest Y_4. However, it is also conceivable that two of the four subtests are affected by stereotype threat. This situation is displayed in Figure 3. Again, this would result in negative effects on the measurement intercepts of these subtests (cf. Appendix B). In addition, if stereotype threat effects vary over persons, then this would lead to increased residual variances of both affected subtests. Furthermore, the two affected subtests now covary more strongly than would be expected from their corresponding factor loadings on the η factor. This additional covariance due to stereotype threat constitutes a violation of the dimensionality of the factor model within this group (i.e., residual covariance), resulting in model misfit. This scenario of stereotype threat affecting the performance on two subtests can be extended to cases in which more than two (or even all) subtests are affected. Of course, in such cases, stereotypethreat also violates strict factorial invariance.[4] As a final scenario, consider Figure 4, in which a nonlinear effect on subtest Y_4 is expressed as an interaction between latent ability and stereotype threat. As is shown in Appendix B, such an effect results in a lowering of the factor loading of subtest Y_4 (cf. Figure 1, bottom half). Additionally, one would expect an increase of the residual variance of the affected subtest and a downward bias of the intercept.

In conclusion, the effects of stereotype threat are detectable by tests for measurement invariance using MGCFA. Possible stereotype threat effects would show up particularly in group differences in the measurement parameters of difficult subtests. We now turn to three experiments in which the amount of stereotype threat for stigmatized groups was manipulated. We thus use tests of strict factorial invariance with respect to groups and conditions to identify the effects, if any, of stereotype threat on the test scores.

[4] However, if a (relatively) large number of subtests are affected by stereotype threat, then model misfit due to such stereotype threat effects disperses over the model. This makes it difficult to interpret measurement bias in terms of sole parameters.

Study 1: Dutch Minorities and the Differential Aptitude Test (DAT)

On average, Dutch minority students attain lower educational levels and have a higher dropout rate than Dutch majority students (Dagevos, Gijsberts, & van Praag, 2003). Several studies have indicated that Dutch high school students often view minority students as less smart (Verkuyten & Kinket, 1999) and minority groups as less educated (Kleinpenning & Hagendoorn, 1991). Recently, Verkuyten and Thijs (2004) found that academic disidentification among Dutch minority students was moderated by the perception of being discriminated in scholastic domains. The first aim of Study 1 was to study the effects of stereotype threat on intelligence test performance in a sample of minority high school students in the Netherlands. To this end, we administered a short intelligence test that contained three subtests, and we varied the amount of stereotype threat related to ethnic minorities by changing the presentation of the test and by altering the timing of an ethnicity questionnaire. The second aim of this study was to find out whether tests for measurement invariance that use MGCFA can successfully highlight the effects of stereotype threat. Furthermore, we compared the results of confirmatory factor analysis with the results of ANOVA to find out whether both analyses led to the same conclusions.

Method

Participants. A total of 295 students from nine high schools in large cities in the Netherlands participated during obligatory classes, which were aimed at counseling the students in choosing a major ("profile") in the second phase ("tweede fase") of their high school education. The students were between the ages of 13 and 16 years (M = 14.86, SD = 0.64) and attended the 3rd year of the Hoger Algemeen Voortgezet Onderwijs (HAVO), or higher general secondary, education. Given that the HAVO level is the second-highest level in the Dutch high school system, the sample was expected to be heterogeneous in terms of identification with the academic domain, which is considered an important moderator of stereotype effects (cf. Aronson et al., 1999).

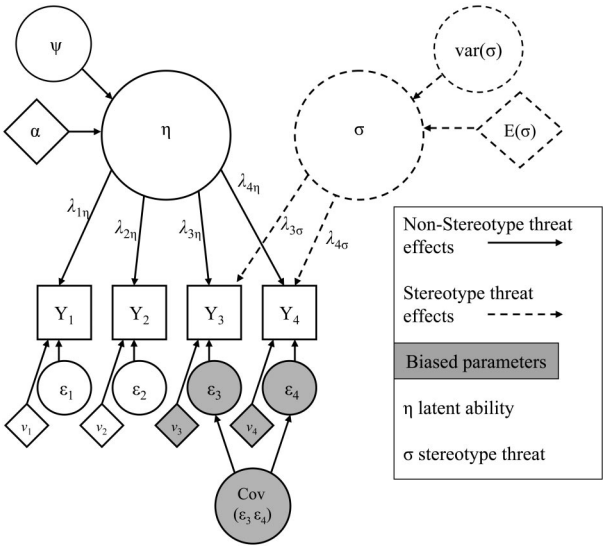

Figure 3. The common factor model in which stereotype threat affects the scores on Subtests Y_3 and Y_4. var = variance; E = expected value; Cov = covariance.

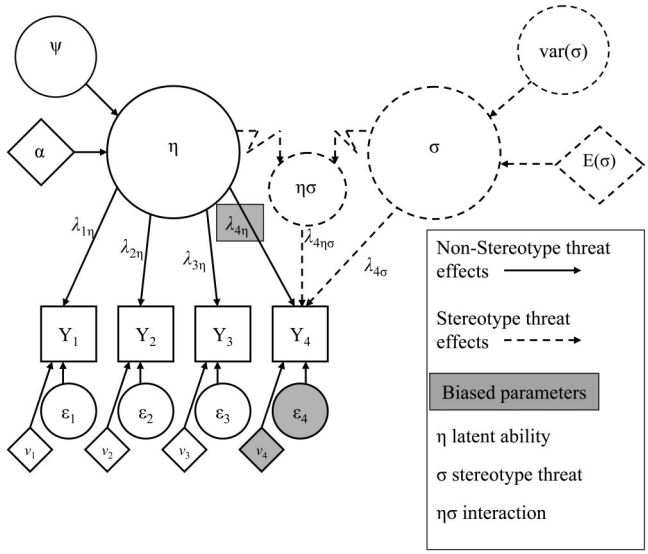

Figure 4. The common factor model in which a nonlinear effect of stereotype threat on the scores of Subtest Y_4 are expressed as an interaction between latent ability and stereotype threat. var = variance; E = expected value.

All 157 students in the majority group were born in the Netherlands, as were all their parents and grandparents. Of the 138 minority students, most were born in the Netherlands (76%), but all of them had one (10%) or two (90%) parents born outside the Netherlands. The (grand)parents of the minority students were immigrants from (former) Dutch colonies (Surinam/Antilles; n = 47, Turkey (n = 36), or Morocco (n = 55).[5] Because of the absence of large test score differences between these minority groups, and to increase the sample sizes, these minority groups were pooled.[6] When asked to indicate the cultural group they identified with, most (n = 93; 67%) of the minority students indicated their own minority group. A total of 23 minority students (17%) indicated the Dutch majority group, and 22 minority students (16%) indicated both the Dutch group and their minority group as the group with which they identified. The total sample consisted of 119 boys and 176 girls. Both ethnic groups did not differ in sex and age composition.[7]

[5] These data stem from a larger study that contained 430 students (Wicherts et al., 2003). We selected only students that could be categorized unambiguously in the majority group (student, his/her parents, and grandparents are all born in the Netherlands) or in one of these three minority groups.

[6] Although there may be differences between these minority groups in terms of general stereotypes, in terms of academic stereotypes, differences between these groups are quite small (see, e.g., Kleinpenning & Hagendoorn, 1991).

[7] To ensure the existence of stereotypes concerning the intellectual ability of minority groups, we conducted a pilot study in which we asked a group of 41 students in comparable schools and classes whether they believed that there existed prejudices concerning the intellectual ability of their cultural group (direction unspecified). On a scale ranging from 1 (*no prejudice*) to 5 (*strong prejudice*), the 20 majority students (M = 2.00, SD = 1.12) scored significantly lower, $t(39) = 4.53$, $p < .001$, than the 21 minority students (M = 3.62, SD = 1.16), indicating that the minority students reported a strong awareness of the stereotypes concerning the intellectual abilities of their group.

Procedure and design. Three shortened subtests of the DAT (Evers & Lucassen, 1992) were administered during classes, which were attended by 17–27 students. On arrival in the classroom, students found a test booklet on their desks, and a female tester of Dutch origin told them that they would be taking a counseling test. The tester said that the test booklet contained questions about their personal interests and abilities and that their answers would be used for guidance in their choice of specialty. No explicit mention of intelligence was made. The tester told the students that the test booklet consisted of several sections and that they would be told when to start and stop with a particular section. This enabled the timing of each of the following sections of the test booklet: an ethnicity questionnaire, the DAT tests, an interest inventory, an additional language test (used for exploratory purposes), and the actual profile-counseling test (administered last). After the test session, students were debriefed extensively on the purpose of the experiment. After 1 week, all students received written counseling on their specialty choice, which was based solely on the profile-counseling test (cf. Zand Scholten, 2003). Special care was taken to ensure that the answers on this test were not affected by the stereotype threat manipulation or by ethnicity (Wicherts et al., 2003).

Participants were assigned to two conditions that differed in the features that elicit stereotype threat for the minority students. Assignment to conditions was achieved by randomly distributing two versions of a test booklet, which were indistinguishable by the cover. In the stereotype threat condition, this test booklet presented each DAT subtest as an "intelligence test." The test booklet of participants in the control condition made no mention of intelligence, and the tests were simply presented as a section of the test booklet. In addition, in the stereotype threat condition, an ethnicity questionnaire was administered prior to the DAT. This questionnaire consisted of 14 questions concerning ethnic and cultural background (religion, language use) and questions about the place of birth of the students, their parents, and grandparents. In the control condition, the ethnicity questionnaire was administered after the DAT. While participants in the stereotype threat condition filled in the ethnicity questionnaire, participants in the control condition filled in an interest inventory that contained 15 items without any connection to ethnicity. This interest inventory was administered to students in the stereotype threat condition after the intelligence tests. Thus, two stereotype threat manipulations were used in concert to increase stereotype threat for ethnic minorities: an ethnicity prime and a manipulation of the diagnosticity of the intelligence test (cf. Steele & Aronson, 1995).

Intelligence test. Three subtests of the Dutch DAT (Evers & Lucassen, 1992) were used as a measure of general intelligence. The subtests were shortened by selecting items with the highest item-rest correlations in the Dutch standardization sample (N = 2,100). The Numerical Ability (NA) test (originally 40 items, 25 min) contains 14 complicated mathematic problems. Abstract Reasoning (AR; originally 45 items, 25 min) contains 18 items with a logical sequence of diagrams, which had to be completed. Verbal Reasoning (VR; originally 50 items, 20 min) contains 16 verbal analogy items. All subtests were administered with a time limit of 6 min. All items have a 5-point multiple-choice answer format. On the basis of the standardization data, NA is the most difficult subtest in terms of proportion correct of the items retained in the short version (average p = .43), followed by VR (average p = .49) and AR (average p = .59). Thus, one would expect the strongest stereotype threat effects on the NA test. The instruction pages of the subtests were slightly adapted with regard to the time limit, number of items, and the presentation of the tests as either a section (control condition) or as an intelligence test (stereotype threat condition). To correct for possible order effects, and to avoid cheating (e.g., copying answers), we used two order versions of the test booklet (bringing the total number of versions to four). The order in these two versions was NA–AR–VR and VR–NA–AR, respectively. Because none of the main or interaction effects for order reached significance (ANOVA; all ps $> .10$), these order versions were pooled for the subsequent analyses.

Analyses. Considering previous factor analyses on the complete DAT (Evers & Lucassen, 1992; Te Nijenhuis, Evers, & Mur, 2000; Wicherts et al., 2004), the use of a one-factor model for these three subtests was sensible. Although our primary interest was in testing for strict factorial invariance with respect to groups, we also conducted a 2 × 2 multivariate analysis of variance (MANOVA), with stereotype threat and ethnicity as factors and the three subtests as dependent variables. MANOVA provides a means to interpret the experimental mean effects. We predicted a significant main effect for ethnicity, with majority students outscoring the minority students (see, e.g., Te Nijenhuis et al., 2000). In addition, we expected a significant Ethnicity × Condition interaction, because stereotype threat would primarily depress scores of minority students. Given the heterogeneous sample used, we also expected heterogeneity in covariances and variances over design cells. Therefore, as is common in the (M)ANOVA framework, we also conducted tests for variance and covariance heterogeneity by means of Box's M test and the univariate Levene's test.

MGCFA can be used to shed light on the nature of differences in (co)variance and mean structure between groups. Within this 2 × 2 experimental design, the tenability of strict factorial invariance with respect to groups and conditions (i.e., four groups) is investigated by fitting a series of increasingly restrictive models. These models, as well as the restrictions imposed, are presented in Table 1. In the first step, no between-groups restrictions are imposed. The next steps involve restricting all factor loadings (Step 2) and all residual variances (Step 3) to be invariant over all four groups. Because of the random assignment to experimental conditions, one does not expect there to be differences on the factor level between conditions for both existing groups. Step 4 can be used to investigate whether factor variance of the existing groups are affected by the stereotype threat manipulation. That is, in this step, the factor variance for majority students in the stereotype threat condition is restricted to be equal to the factor variance for majority students in the control condition (and similarly for the minority students). In Step 5, the invariance of the mean structure is investigated by restricting the measurement intercepts to be equal across all groups. In the same step, factor mean differences with respect to an arbitrary baseline group are estimated. Finally, in Step 6, the means of the existing groups are restricted to be equal over condition (e.g.,

Table 1
Equality Constraints Imposed in the Steps Toward Strict Factorial Invariance

No.	Description	Λ factor loadings	Θ residual variances	Ψ factor variance	ν intercepts	α factor mean
1	Configural invariance	—				
2	Metric invariance	All groups[a]	—			
3	Equal residual variances	All groups	All groups[a]	—		
4	Factor variances invariant over condition	All groups	All groups	Existing groups[a]	—	
5	Strict factorial invariance	All groups	All groups	Existing groups	All groups[a]	—
6	Factor means invariant over condition	All groups	All groups	Existing groups	All groups	Existing groups[a]

Note. Each step is nested under the previous one.
[a] Indicates that restrictions are tested in that particular step.

Table 2
Means and Standard Deviations by Experimental Condition and Ethnic Group (Study 1)

	Condition							
	Control				Stereotype threat			
	Majority (n = 79)		Minority (n = 65)		Majority (n = 78)		Minority (n = 73)	
Subtest	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Numerical	5.35	2.54	4.88	2.47	5.49	2.31	4.67	2.52
Abstract Reasoning	10.42	2.96	6.80	3.33	9.24	3.34	7.34	2.83
Verbal Reasoning	7.27	3.01	5.37	2.82	6.65	3.47	5.56	2.70

factor mean of majority group in control condition equal to factor mean of majority group in stereotype threat condition). This ensures that the experimental manipulation has no effect on the mean of the common factor. As can be seen, if the restrictions implemented in these six steps hold, then measurement invariance holds. In that case, the differences between the existing groups are a function of the differences in the means (α) and variances (Ψ) of the common factor. However, we expected the test scores to be affected in a differential manner across groups. The tenability of each restriction is judged by differences in fit between the restricted model and the less-restricted model. For instance, Step 2 versus Step 1 involves the tenability of equality of factor loadings. Because of the nesting of models, a loglikelihood test is used to test each restriction. Besides attention for chi-squares, the comparative fit index (CFI) and the root-mean-square error of approximation (RMSEA) are used in determining the absolute and relative model fit. The CFI (Bentler, 1990) ranges from 0 to 1, and is a measure of the relative fit of a model in relation to a null model of complete independence.[8] The RMSEA (Browne & Cudeck, 1993) is a so-called close fit measure that is known to be relatively insensitive to sample size. Several rules of thumb have been proposed for these fit measures. On the basis of their simulation study, Hu and Bentler (1999) proposed that RMSEA values smaller than 0.06, and CFI values larger than 0.95, are indicative of good model fit.

In case a step is accompanied by a clear deterioration in model fit, the particular restriction is rejected. In such cases, modification indices can highlight the particular parameter(s) causing the misfit. A modification index (MI) is a measure of how much chi-square is expected to decrease if a constraint on a given parameter is relaxed, and the model is refitted (Jöreskog & Sörbom, 1993). In cases in which a restriction is accompanied by a deterioration in fit, parameters with the highest MI are freely estimated, and the sequence of models is continued. We expected that stereotype threat effects on test performance would result in measurement bias expressed by high modification indices in the minority group in the stereotype threat condition. We carried out all factor analyses using LISREL 8.54 (Jöreskog & Sörbom, 2003).[9]

Results

The values for univariate skewness and kurtosis in the four groups are in an acceptable range (i.e., from −0.89 to 0.88), suggesting no large deviations from normality. Therefore, the use of maximum likelihood for estimating the factor models is justified. Table 2 contains means and standard deviations of the three subtests for both ethnic groups in the two conditions. First, we provide the ANOVA results. Box's M test suggests some covariance heterogeneity over groups, $F(18, 284863) = 1.79$, $p < .05$. The univariate Levene's test for homogeneity of variance gives a significant value for the VR subtest, $F(3, 291) = 3.63$, $p < .05$. Because MANOVA is often claimed to be robust to (co)variance heterogeneity (e.g., Stevens, 1996), we do interpret the results of the MANOVA. The multivariate main effect for ethnicity is significant, $F(3, 289) = 20.36$, $p < .001$, as well as all univariate effects—NA: $F(1, 291) = 5.07$, $p < .05$; AR: $F(1, 291) = 57.47$, $p < .001$; VR: $F(1, 291) = 17.83$, $p < .001$—with the majority group outscoring the minority group. Neither the multivariate nor any of the univariate main effects for condition reach significance (all ps $> .30$). The multivariate interaction effect between condition and ethnicity is significant, $F(3, 289) = 2.64$, $p = .050$. The only significant univariate interaction effect is found on the AR subtest, $F(1, 291) = 5.56$, $p < .05$. However, this interaction effect is due to the majority group underperforming in the stereotype threat condition. Namely, the condition simple effect is significant for majorities, $F(1, 155) = 5.45$, $p < .05$, but nonsignificant for minorities, $F(1, 136) = 1.07$, $p > .30$. All multivariate and univariate simple effects for condition within the minority group are nonsignificant (all ps $> .30$), which is opposite to what one would expect from stereotype threat theory. Whereas, the minority group scored significantly lower than the majority group, these ANOVA results indicate that on average the minority students in the stereotype threat condition did not score lower than the minority students in the control condition.

However, it is important to stress that the sample may be expected to be heterogeneous with respect to domain identification, considered an important moderator of stereotype threat effects (e.g., Steele, 1997). For instance, Aronson et al. (1999) found that test takers that identified strongly with the domain of interest (i.e., mathematics) were more susceptible to stereotype threat, whereas test takers who moderately identified with the domain performed better under stereotype threat conditions than under control conditions. This suggests that within heterogeneous samples that contain both highly identified and moderately identified test takers, effects of stereotype threat may differ substantively

[8] Widaman and Thompson (2003) argued that because of the nesting of models, it is inappropriate to use such a null model within a multigroup context with mean structure. Therefore, we used a model without any factor structure, in which intercepts and residual variances are restricted to be group invariant (i.e., Model 0A in Widaman & Thompson's, 2003, study), as the null model in computing the CFI values.

[9] All input files used here can be downloaded from the following Web site: http://users.fmg.uva.nl/jwicherts

over persons. In such samples, positive and negative effects may cancel out, resulting in no, or only a small, effect on the mean. However, the absence of a mean effect does not necessarily mean the absence of an effect. To investigate the possibility that covariance structure was affected by the stereotype threat induction, we tested for measurement invariance with respect to the four groups. The results of the multigroup confirmatory factor analyses are reported in Table 3.

Because a one-factor model with three indicators is saturated (i.e., equal number of input statistics and parameters), the baseline model without across-group restrictions has a chi-square of zero with zero degrees of freedom. In the second step, the factor loadings are restricted to be equal over the four groups. This restriction results in a significant increase in chi-square. In addition, both the RMSEA and the CFI exceed the rule-of-thumb values for good fit. The misfit in this step is almost solely due to the factor loading of the NA subtest of the minority group in the stereotype threat condition (MI = 11). Freeing this parameter leads to a significant improvement of model fit, as can be seen in Step 2a. In the minority group, stereotype threat condition, this (unstandardized) factor loading is not significantly different from zero ($\lambda_1 = -0.04$, $SE = 0.20$, $Z = -0.19$, $p > .05$), whereas in the other groups this factor loading is significantly greater than zero ($\lambda_1 = 0.92$, $SE = 0.22$, $Z = 4.19$, $p < .01$). In Step 3, the residual variances are restricted to be invariant over the four groups. This, again, leads to a significant deterioration in model fit, as can be seen by the significant increase in chi-square, increase in RMSEA, and lowering of CFI. Not surprisingly, the misfit in this step is mainly due to the residual variance of the NA subtest of the minority group in the stereotype threat condition (MI = 7). Freeing this parameter leads to a significant improvement in model fit (Step 3a). The residual variance of NA is larger in the minority group, stereotype threat condition (6.33, $SE = 1.06$) than in the other groups (3.47, $SE = 0.61$). In the fourth step, we restrict factor variances of both ethnic groups to be invariant over condition. This leads to a relative improvement in model fit. The factor variance of the minority group is slightly smaller ($\Psi = 3.32$, $SE = 1.08$) than the factor variance of the majority group ($\Psi = 4.12$, $SE = 1.23$). In the fifth step, mean structure is modeled by restricting the intercepts to be invariant over the groups and by freeing the factor means of three groups (cf. Table 1). In light of the different factor loading of the NA subtest in the minority group, stereotype threat condition, it does not make sense to restrict this particular intercept. Hence, in Step 5, this parameter is freely estimated for this particular group. Step 5 results in a significant increase in chi-square, an increase in RMSEA, and a clear drop in CFI. The restriction on intercepts is clearly rejected. The highest MI is related to the intercept of the AR test of the majority group in the control condition. Freeing this parameter results in an improvement in model fit (Step 5a). However, as judged by RMSEA ($>$ 0.06) and CFI (< 0.95), the model fit of Step 5a is still not very good. The highest MI (MI = 4) in this step is related to the intercept of the AR subtest of the minority group in the control condition. Freeing this parameter results in an improvement in model fit in terms of RMSEA and CFI (Step 5b). An interesting finding is that the intercept of the AR subtest is higher in the majority group, control condition ($\nu_2 = 8.67$, $SE = 0.47$) than in the two ethnic groups in the stereotype threat condition ($\nu_2 = 7.54$, $SE = 0.31$). This is not surprising considering the mean effect of the stereotype threat manipulation on this subtest in the majority group. In the minority group, control condition, this intercept is even lower ($\nu_2 = 6.72$, $SE = 0.37$). This suggests the presence of bias with respect to ethnicity in the control condition. In the sixth and final step, we investigated whether the factor means of both groups differed over experimental condition. This step is accompanied by a relative improvement in model fit. The factor mean of the majority is significantly higher than that of the minority group ($\alpha = 1.62$, $SE = 0.39$, $Z = 4.20$, $p < .001$).

Discussion

Although MANOVA results indicated an absence of mean effects of stereotype threat on test performance of the minority group, the stereotype threat manipulation clearly resulted in measurement bias with respect to the minority group. The measurement bias due to stereotype threat was related to the most difficult NA subtest. An interesting finding is that, because of stereotype threat, the factor loading of this subtest did not deviate signifi-

Table 3
Fit Measures of Steps Toward Strict Factorial Invariance (Study 1)

Step	Restrictions	*df*	χ^2	*p*	Δdf	$\Delta\chi^2$	*p*	RMSEA	CFI
1		0	0.00	1.000				0.000	1.000
2	**Λ**	6	14.73*	.023	6	14.73*	.023	0.145	0.942
2a	$\Lambda^{\mathbf{a}}$	5	4.74	.449	1[f]	9.99**	.002	0.000	1.000
3	Λ^{a}, **Θ**	14	23.68	.050	9	18.94*	.026	0.097	0.936
3a	Λ^{a}, $\Theta^{\mathbf{b}}$	13	16.45	.226	1[f]	7.23**	.007	0.058	0.977
4	Λ^{a}, Θ^{b}, **Ψcon**	15	16.91	.324	2	0.46	.795	0.040	0.987
5	Λ^{a}, Θ^{b}, $\boldsymbol{\nu}^{c}$, Ψcon	20	31.78*	.046	5	14.87*	.011	0.089	0.922
5a	Λ^{a}, Θ^{b}, $\nu^{c,\mathbf{d}}$, Ψcon	19	27.43	.095	1[f]	4.35*	.037	0.079	0.944
5b	Λ^{a}, Θ^{b}, $\nu^{c,d,\mathbf{e}}$, Ψcon	18	23.70	.165	1[f]	3.73	.053	0.065	0.962
6	Λ^{a}, Θ^{b}, $\nu^{c,d,e}$, Ψcon, ***α*con**	20	24.44	.224	2	0.74	.691	0.056	0.971

Note. Restrictions in bold are tested by loglikelihood test delta chi-square. RMSEA = root-mean-square error of approximation; CFI = comparative fit index; Con = restriction over conditions for existing groups.
[a] Factor loading Numerical Ability, minority group, stereotype threat. [b] Residual variance Numerical Ability, minority group, stereotype threat. [c] Intercept Numerical Ability, minority group, stereotype threat. [d] Intercept Abstract Reasoning, majority group, control. [e] Intercept Abstract Reasoning, minority group, control. [f] Parameter freely estimated.
* $p < .05$. ** $p < .01$.

cantly from zero. This change in factor loading suggests a nonuniform effect of stereotype threat. This is consistent with the third scenario discussed above (cf. Appendix B) and with the idea that stereotype threat effects are positively associated with latent ability (cf. Cullen et al., 2004). Such a scenario could occur if latent ability and domain identification are positively associated. This differential effect may have led low-ability (i.e., moderately identified) minority students to perform slightly better under stereotype threat (cf. Aronson et al., 1999), perhaps because of moderate arousal levels, whereas the more able (i.e., highly identified) minority students performed worse under stereotype threat. Such a differential effect is displayed graphically in Figure 5. This pattern could explain the absence (i.e., canceling out) of mean effects, the increased residual variance, and the smaller factor loading in the minority group. Another explanation for this effect may lie in individual differences in working memory capacity (WMC). Beilock and Carr (2005) recently found that students high in WMC underperformed on a difficult arithmetic task under pressure, whereas students low in WMC showed a slight increase in performance when put under high pressure.

The biasing effect of stereotype threat would have been completely overlooked, had we restricted ourselves to the MANOVA, and had we regarded the covariance heterogeneity as a statistical annoyance, instead of as an important source of information. The bias due to stereotype threat on test performance of the minority group is quite serious. The intelligence factor explains approximately 0.1% of the variance in the NA subtest, as opposed to 30% in the other groups. To put it differently, because of stereotype threat, the NA test has become completely worthless as a measure of intelligence in the minority group. Note, however, that such an effect changes our interpretation of the factor within the minority group under stereotype threat. It is also conceivable that the stereotype threat effects were present on the other two subtests. However, because of the rather small factor model, such an effect is hardly distinguishable from a nonuniform effect on the NA test. Nevertheless, the latter subtest is the most difficult subtest, and it is apparent that stereotype threat resulted in severe measurement bias with respect to the minority group.

In the control condition, there also appears to be measurement bias with respect to ethnicity, indicating that even in that condition test scores of minority and majority students are incomparable. It could be argued that because the test setting resembled strongly the common practice of testing in Dutch high school, the test setting could have elicited stereotype threat even in the control condition (cf. Steele & Davies, 2003). However, because the bias in this condition was related to the easiest of the three subtests, it seems unlikely that stereotype threat has caused this bias. Further research could shed light on the issue of whether stereotype threat is also present in the control condition or if perhaps the bias is caused by something else (cf. Te Nijenhuis et al., 2000). On the basis of this study, we would advise great caution in the use of these DAT scales for Dutch minority students.

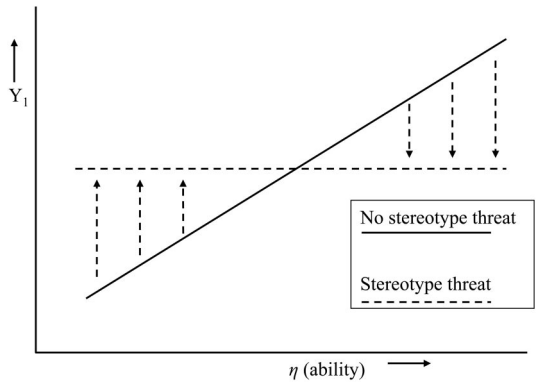

Figure 5. Non-uniform effect on factor loading of Subtest Y_1 in case of an interaction between latent ability and stereotype threat.

Surprisingly, the manipulation also had a depressing effect on the AR subtest in the majority group. Perhaps this is due to a priming effect of the ethnicity questionnaire (cf. Wheeler & Petty, 2001). Further research could shed light on why the scores on this relatively easy subtest were depressed in the majority group. Nevertheless, the depressing effect of stereotype threat on this subtest became apparent in the ANOVA, and clearly resulted in measurement bias in the factor analyses.

The presence of covariance effects in the absence of mean effects in this first study, led us to reanalyze the results of another stereotype threat study, in which a clear mean effect on test performance was also absent. In an experiment by Nguyen et al. (2003), the effects of stereotype threat on Black students' test performance were studied within a job-selection context. A timed short version of a cognitive ability test that contained three subtests was used to assess cognitive ability. A total of 86 Blacks and 86 Whites were randomly assigned to a stereotype threat or control condition. Similar to Study 1 above, stereotype threat was manipulated by both an ethnicity prime and by test diagnosticity (Nguyen et al., 2003). Using ANOVA, Nguyen et al. found that Whites outscored the Blacks on all subtests (i.e., significant multivariate and univariate main effects for ethnicity). However, MANOVA indicated no significant interaction between stereotype threat manipulation and race, as would be expected from stereotype threat theory. Therefore, Nguyen et al. concluded that stereotype threat effects on test performance were absent. We submitted these data to MGCFA, and our reanalysis suggested that (besides an increased residual variance for Whites in the stereotype threat condition) strict factorial invariance with respect to conditions and race was mainly tenable. Although the power may have been low, this result suggests that the race differences in test performance in either condition appear not to be caused by stereotype threat. Therefore, the argument that the stereotype threat manipulation in Nguyen et al.'s study was unsuccessful because of the fact that stereotype threat was already present in the control condition (Steele & Davies, 2003) appears implausible.

From an experimental perspective, the results of the first study are unusual in the sense that experimental mean effects on test performance of the stigmatized group were absent. Hence, it is desirable to investigate the merits of our modeling approach in the presence of clear experimental mean effects.

Study 2: O'Brien and Crandall's (2003) Reanalysis

O'Brien and Crandall (2003) studied the effects of stereotype threat on performance of female students on three mathematics

Table 4
Means and Standard Deviations of Male Students and Female Students by Experimental Condition (Study 2)

	Condition							
	Control				Stereotype threat			
	Male students (n = 50)		Female students (n = 30)		Male students (n = 51)		Female students (n = 28)	
Subtest	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Easy	7.50	4.34	6.37	3.91	7.80	3.93	8.18	3.98
Difficult	9.13	2.36	7.99	2.88	9.19	2.51	6.81	2.55
Persistence	18.72	5.79	15.30	6.13	19.53	4.67	16.43	6.30

Note. Descriptive statistics provided by O'Brien and Crandall (2003).

tests, which differed in difficulty level: a difficult test, an easy test, and a relatively easy math persistence test. Here, we reanalyze these data with our modeling approach to investigate whether a test of strict factorial invariance can highlight the stereotype threat effects on test performance. We briefly describe the original study. For more details, the reader is referred to O'Brien and Crandall (2003).

Method

Participants. A total of 164 students enrolled in a psychology class participated in this study in exchange for course credit. Because of missing data for 5 participants on the math persistence test, the current analysis is based on a sample of 58 female students and 101 male students.

Design and procedure. Participants were randomly assigned to two conditions that differed in the amount of stereotype threat for women. In the control condition, the gender stereotype was made irrelevant for the test setting by a text stating that the test at hand had "NOT been shown to produce gender differences" (O'Brien & Crandall, 2003, p. 785). In the stereotype threat condition, the text indicated that the test had been shown to produce gender differences. After reading this text, participants completed a questionnaire regarding their feelings concerning test taking. After that, the three math tests were administered in a counterbalanced order.

Materials. The easy math test had a time limit of 10 min and consisted of 20 relatively easy multiplication problems. The difficult test was administered with a time limit of 11 min and consisted of 15 difficult items from the quantitative SAT. Items were in a five-option multiple-choice format. The math persistence test contained 24 addition and subtraction problems, which were to be solved mentally (i.e., without the aid of paper and pencil) within 8 min (O'Brien & Crandall, 2003).

Analyses. Reasoning that the effects of heightened arousal on task performance depend on task difficulty, O'Brien and Crandall (2003) expected that stereotype threat would heighten scores of female students on the easy math test while depressing their scores on the difficult test. The math persistence test was originally used as a control for effort. However, because of quite high correlations between all three subtests, and in light of the clear mathematical nature of the three tests, the use of a one-factor model in describing these data is justified. In the male groups in both conditions and in the female group, stereotype threat condition, all inter-subtest correlations are significantly greater than zero ($p < .05$; range = 0.33–0.55). However, the correlation between the easy and the difficult test of the female group in the control condition is not significant. Furthermore, the correlation between the easy test and the math persistence test in this group is negative. This appears not to be caused by any distinguishable bivariate outliers (L. T. O'Brien, personal communication, June 7, 2004). Moreover, in this group, the math persistence test has a platykurtotic distribution (kurtosis = −1.3). In combination with the small sample size (n = 30), this makes the data of this group less suitable for maximum likelihood estimation. Therefore, we limited the factor analyses to three groups: the female group in the stereotype threat condition, and the male groups in both conditions. For our modeling approach, this poses no problem. We expected measurement bias because of stereotype threat in the female group. We again use the steps given in Table 1 to assess the tenability of restrictions over these three groups.

Results

Except for the math persistence test scores of the male group in the stereotype threat condition,[10] the kurtosis and skewness values are in the moderate range, making the data suitable for maximum likelihood estimation. The means and standard deviations of the four groups are reported in Table 4. Using repeated-measures ANOVA on the standardized scores of the easy and the difficult tests, O'Brien and Crandall (2003) found a significant main effect for gender, with male students outscoring the female students. More important, this test showed a significant three-way interaction between gender, condition, and test difficulty, which indicated that stereotype threat lowered scores of women on the difficult test while heightening the scores on the easy test. In a separate two-way ANOVA on the math persistence scores, O'Brien and Crandall found a significant main effect for sex (male students outscoring female students), although the interaction between sex and condition was not significant. Thus, these ANOVA results indicate no effects of condition for male students. For female students, ANOVAs indicate a clear mean effect of stereotype threat on the easy and difficult tests but no effect on the math persistence test.

The results of the factor analyses on the three groups are reported in Table 5. Again, the first step involves a saturated model with perfect model fit. The second step (equal factor loadings), the third step (equal residual variances), and the fourth step (equal factor variance in male groups) all result in nonsignificant increases in chi-square. Moreover, the CFI and RMSEA clearly indicate that these three restrictions are tenable. This is not the case for the restriction on measurement intercepts, which is tested in the

[10] The high kurtosis value (2.6) in this group was due to a very low scoring male student. Excluding this outlier does not change the results of the factor analyses.

Table 5
Fit Measures of Steps Toward Strict Factorial Invariance (Study 2)

Step	Restrictions	*df*	χ^2	*p*	Δdf	$\Delta\chi^2$	*p*	RMSEA	CFI
1		0	0.00	1.000				0.000	1.000
2	**Λ**	4	2.74	.602	4	2.74	.602	0.000	1.000
3	Λ, **Θ**	10	5.87	.826	6	3.13	.792	0.000	1.000
4	Λ, Θ, **Ψcon**	11	6.40	.846	1	0.53	.467	0.000	1.000
5	Λ, Θ, ***ν***, Ψcon	15	22.78	.089	4	16.38**	.003	0.113	0.896
5a	Λ, Θ, ν^{a}, Ψcon	14	12.42	.572	1[c]	10.36**	.001	0.000	1.000
5b	Λ, Θ, $\nu^{a,b}$, Ψcon	13	6.66	.919	1[c]	5.76*	.016	0.000	1.000
6	Λ, Θ, $\nu^{a,b}$, Ψcon, ***α*con**	14	7.02	.934	1	0.36	.549	0.000	1.000

Note. Restrictions in bold are tested by loglikelihood test delta chi-square. RMSEA = root-mean-square error of approximation; CFI = comparative fit index; Con = restriction over conditions for existing groups.
[a] Intercept Difficult subtest, women, stereotype threat. [b] Intercept Easy subtest, women, stereotype threat. [c] Parameter freely estimated.
* $p < .05$. ** $p < .01$.

fifth step. This restriction clearly results in a worsening in model fit, as is clear in the significant increase in chi-square and the clear worsening in CFI and RMSEA values. The largest modification indices are related to the intercepts of the difficult test (MI = 7) and the easy test (MI = 6) of the female group in the stereotype threat condition. Freeing both parameters (Steps 5a and 5b) results in clear improvements in model fit. The intercept of the difficult test is lower in the female group under stereotype threat ($\nu_2 = 7.72$, $SE = 0.57$) than in both male groups ($\nu_2 = 9.05$, $SE = 0.30$). The intercept of the easy test is higher in the female group ($\nu_3 = 9.65$, $SE = 0.94$) than in both male groups ($\nu_3 = 7.48$, $SE = 0.50$). In the sixth step, the factor mean of the male students in both conditions is restricted to be equal. This does not result in a worsening in model fit. In this last step, the factor mean of the female group is significantly lower than that of the male group ($\alpha = 2.70$, $SE = 1.28$, $Z = 2.11$, $p < .05$). However, because of the two freely estimated intercepts, this factor mean difference is actually a significance test of the difference between male students and female students on the math persistence test.

Discussion

The reanalysis of O'Brien & Crandall's (2003) data demonstrated one drawback of the current modeling approach. Because of the platykurtotic distribution of test scores, and the negative correlation between tests in the female group control condition, this group had to be excluded from the test for measurement invariance. Nevertheless, the factor analysis approach remained feasible. Even without the possibility to compare the female group in the stereotype threat condition with a female group without such threat effects (i.e., in the control condition), we were able to establish that test scores of male students and female students were incomparable. It became apparent that intercepts were not invariant across groups, and that strict factorial invariance was violated because of stereotype threat. Suppose that these data would have been nonexperimental data, stemming from a real-life, or even a high-stakes, test setting. Even then, a test for strict factorial invariance would have pointed toward the measurement bias with respect to sex. The reanalysis of these data illustrates our point that because of their nature, stereotype threat effects are detectable in principle by means of tests for measurement invariance.

Of course, O'Brien and Crandall (2003) especially selected their math tests to show this pattern of effects. However, their study can contribute to future studies into stereotype threat effects within real-life test settings. A careful selection of easy and more difficult tests, together with the current modeling approach, enables one to investigate the existence of stereotype threat effects on test performance. In sum, the results of the current reanalysis are clearly in line with the results of ANOVA by O'Brien and Crandall. Moreover, the current results support the notion that whenever stereotype threat affects test performance on a collection of tests, it does so in a way incompatible with the requirements for measurement invariance within the common factor model.

One drawback of the first two studies is the small number of subtests. In Study 3, we used a test battery that consisted of four subtests that measured arithmetic/mathematic ability. In addition, we wanted to investigate strict factorial invariance in three conditions that differed with respect to stereotype threat related to female test takers: a control condition with no explicit reference to sex differences, a nullified condition in which gender stereotype was made irrelevant to the test, and a stereotype threat condition with explicit mention of sex differences. The latter condition is interesting because it has well-known negative effects on female test performance, whereas male test performance is often enhanced (i.e., a stereotype lift effect; Walton & Cohen, 2003). We expected that both this negative and this positive effect would result in measurement bias. The comparison with regard to strict factorial of three conditions that differ in stereotype threat enables one to find a test setting in which stereotype threat is absent and test scores of male students and female students are comparable.

Study 3: Sex Differences in Arithmetic Test Performance

The first aim of this third study was to replicate the effects of stereotype threat on women's test scores on a collection of arithmetic/mathematic ability tests in a sample of psychology undergraduates in the Netherlands. The second aim was to investigate whether tests for measurement invariance using MGCFA can successfully differentiate between conditions in which stereotype threat is manipulated. To this end, we administered an arithmetic

test battery to male students and female students, varied the amount of stereotype threat for female students over conditions, and tested for strict factorial invariance with respect to groups.

Method

Participants. A total of 283 undergraduate psychology students of the University of Amsterdam participated as part of course requirements.[11] On average, the 142 female students were slightly younger (age: $M = 20.40$ years, $SD = 3.76$) than the 141 male students ($M = 21.64$ years, $SD = 4.97$). The sample was highly educated but not especially selected for good arithmetic/mathematic skills. The sample was expected to be heterogeneous with respect to identification with the arithmetic/mathematical domain.

Design and procedure. An arithmetic test battery was administered by computer during two large mixed-sex group sessions. Participants were randomly assigned by the computer to one of three conditions, in which the introductory texts were used to manipulate the amount of stereotype threat. All three texts started by mentioning that the test of arithmetic ability contained four timed subtests. The three versions differed with respect to the next section in the instruction text. In the control condition, meant to resemble the usual testing circumstances, no mention was made of sex differences. In the nullified condition, however, the instruction read (translated from Dutch): "Although on many arithmetic tests sex differences have been found, previous research has shown that on this arithmetic test females achieve as well as males. Mean scores of males and females on the four subtests are equal." This nullified condition was created to make the gender stereotype irrelevant for the test that participants were taking, thereby hopefully reducing the effects of stereotype threat on female students (cf. Brown & Pinel, 2003; O'Brien & Crandall, 2003; Smith & White, 2002; Spencer et al., 1999). In the stereotype threat condition, the text was changed to (translated from Dutch): "Previous research has shown that females and males score differently on this arithmetic test. On the average females score lower than males on all four subtests." This instruction text was meant to increase stereotype threat for female test takers in the stereotype threat condition. (cf. Keller, 2002; O'Brien & Crandall, 2003; Spencer et al., 1999). After this manipulation, the participants completed the four subtests. Each subtest consisted of a page with a specific instruction, an example item, and a test page containing the test items. The computer automatically stopped the subtests when the allocated test period had passed. Total test time was 21 min. After the test session, all participants were debriefed extensively on the purpose of the experiment.

Materials. We used a selection of subtests that measure arithmetic/mathematical proficiency. The four subtests differ in form and difficulty level but are nevertheless expected to measure one single trait, which we henceforth denote by arithmetic ability. In order of presentation, these subtests are as follows: Arithmetic, Number Series, Worded Problems, and Sums.

The Arithmetic test is a timed test of 3 min containing 40 items that stem from an arithmetic ability test by Elshout (1976). The latter test is part of the standard test program of psychology undergraduates at the University of Amsterdam. The original test has high internal consistency and validity (Vorst & Zand Scholten, 2000). The items have an open-ended answer format (e.g., $43 \times 6 =$).

The Number Series test is a test developed to be parallel to the Number Series Test by Elshout (1976). The latter test is also part of the standard test program of the University of Amsterdam's Psychology Department, and it has high internal consistency and validity (Elshout, 1976; Vorst & Zand Scholten, 2000). The test used in the current study contains 20 items in a five-option multiple-choice format and has a time limit of 6 min. An example item follows: "0 1 3 7 15 (options: 25, 29, 31, 32, 23)."

The Worded Problems test has a time limit of 4 min and contains 23 worded arithmetic problems. This test is based on the Arithmetic subtest of the Weschler Adult Intelligence Scale (WAIS)—Dutch edition (Stinissen, Willems, Coetsier, & Hulsman, 1970) and contains some additional and comparable items from the CMS test by Elshout (1976). All items have an open-ended answer format and were adapted to increase difficulty. An example item follows: "Someone has a loan at a 5% interest rate per year. After three years he has paid 225 Euros interest. What is his debt in Euros?"

The Sums test is the NA test of the Primary Mental Abilities (Thurstone, 1958, 1962). It contains 60 items and was administered with a (adapted) time limit of 5 min. The respondents are required to indicate whether a sum is correct or incorrect (e.g., $13 + 39 + 99 + 32 = 183$). To correct for guessing on this subtest, the total score is computed by subtracting half the number of incorrect responses from the number of correct responses.

Although speediness increases the difficulty of all subtests, the items themselves are fairly easy to solve. The Number Series subtest is the most difficult in terms of abstractness and item difficulty. We therefore expected that stereotype threat would particularly affect scores on this subtest.

Analyses. Again, we also provide the results of a two-way MANOVA with sex and condition (three levels) as factors and the four tests as dependent variables. On the basis of research in previous cohorts of psychology undergraduates (e.g., Vorst & Zand Scholten, 2000), we anticipated that male students would outscore the female students on all subtests. We expected that the instruction texts would particularly influence female test performance. Specifically, we expected that female students in the nullified condition would outscore the female students in the control and stereotype threat conditions. In addition, we predicted female students in the stereotype threat condition to score lowest of all groups. We expected no negative effects for male students, although stereotype lift effects (Walton & Cohen, 2003) could conceivably provide a pattern of mean differences for the male students opposite to those of female students.

As the four subtests were expected to load on a general arithmetic ability factor, we fitted a single common factor model in the confirmatory factor analyses. We again followed the stepwise approach given in Table 1, this time involving six groups. We expected to find measurement bias for female students in the stereotype threat condition. This should result in the rejection of strict factorial invariance, particularly because of the induced bias in the relatively difficult Number Series subtest. Whether strict factorial invariance with respect to sex is tenable in the control and nullified conditions depends on the degree of stereotype threat. However, we expected the degree of measurement bias to be greatest in the stereotype threat condition.

Results

With two exceptions (i.e., Arithmetic subtest for male students in control and stereotype threat conditions), univariate skewness and kurtosis values are moderate (-1, 1), suggesting univariate normality of most subtests in most of the cells. Therefore, use of maximum likelihood in estimating the factor models seems appropriate. Means and standard deviations of the subtests for male students and female students in the three conditions are given in Table 6. The Box test shows that homogeneity of covariance matrices across conditions is rejected, $F(50, 139810) = 1.75$, $p < .01$. Levene's tests for equal variances across conditions show significant values for Arithmetic, $F(5, 277) = 4.68$, $p < .001$, and Number Series, $F(5, 277) = 4.62$, $p < .001$, but nonsignificant values for the other two subtests. Assuming robustness to this violation of (co)variance homogeneity, we continue with the MANOVA. The multivariate sex main effect is associated with a

[11] Because of computer failure, 3 additional participants, 1 male student and 2 female students, were excluded from the analyses.

Table 6
Means and Standard Deviations of Subtests per Sex and Condition (Study 3)

	Condition											
	Control				Nullified				Stereotype threat			
	Men (n = 46)		Women (n = 48)		Men (n = 50)		Women (n = 47)		Men (n = 45)		Women (n = 47)	
Subtest	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Arithmetic	13.28	7.46	10.23	4.62	14.18	7.78	11.70	3.53	12.20	5.53	9.96	6.16
Number Series	8.52	3.74	7.60	2.86	8.56	4.36	7.11	2.66	9.22	3.33	5.62	2.35
Worded Problems	8.39	3.43	6.40	2.80	7.60	3.09	6.72	2.32	7.44	2.88	5.74	2.72
Sums	12.90	5.92	11.55	5.14	13.14	5.86	11.21	4.66	12.97	5.11	11.81	5.18

significant F value, $F(4, 274) = 7.35$, $p < .001$. The univariate analyses of variance show significant sex main effects on all subtests—Arithmetic: $F(1, 277) = 12.89$, $p < .001$; Number Series: $F(1, 277) = 25.79$, $p < .001$; Worded Problems: $F(1, 277) = 19.58$, $p < .001$; Sums: $F(1, 277) = 5.43$, $p < .05$—with male students outscoring the female students on all subtests. Furthermore, compared with the nullified and control conditions, there is a clear trend for female students in the stereotype threat condition to score lower. For the male students, the picture is less clear, with highest scores in conditions depending on the subtest used. The multivariate main effect of condition does not reach significance, $F(8, 548) = 1.71$, $p > .05$. Most important, the multivariate interaction of condition and sex is significant: $F(8, 548) = 2.37$, $p < .05$. None of the univariate condition main effects reach significance (all ps $> .10$). As expected, the only significant univariate interaction effect between sex and condition is found on the Number Series subtest: $F(2, 277) = 4.32$, $p < .05$. Within the female group, the simple effect for condition is significant, $F(2, 139) = 7.29$, $p < .01$. Paired comparisons show that female students in the stereotype threat condition scored significantly lower than female students in the control condition ($p < .01$) and significantly lower than female students in the nullified condition ($p < .05$), but that female scores did not differ significantly between nullified and control conditions ($p > .50$). Although male scores on the Number Series subtest are highest in the stereotype threat condition, the condition simple effect for male students did not reach significance, $F(2, 138) = 0.48$, $p > .50$, and the paired comparisons for male students also did not reach significance (all ps $> .50$). In other words, the stereotype lift effect for male students did not reach significance with the traditional ANOVA approach. To summarize, these ANOVA results indicate a clear suppression of scores on the Number Series subtest for female students in the stereotype threat condition.

Results of factor analyses in the six groups are reported in Table 7. In the first step, we assessed the fit of the one-factor model, which is acceptable. The second step does not result in a significant increase in chi-square. Therefore, factor loadings appear invariant over the six groups. The restriction on residual variances in the third step results in a clear deterioration in model fit. The largest modification indices are found in the male group, nullified condition, and are related to the residual variance of the Number Series subtest (MI = 23) and of the Arithmetic subtest (MI = 18). Furthermore, the residual variance of the Arithmetic test in the female students in the stereotype threat condition is also partly responsible for misfit (MI = 13). Freeing these three parameters in a stepwise fashion (Steps 3a, 3b, 3c) results in clear improvements

Table 7
Fit Measures of Steps Toward Strict Factorial Invariance (Study 3)

Step	Restrictions	*df*	χ^2	*p*	Δdf	$\Delta\chi^2$	*p*	RMSEA	CFI
1		12	9.61	.650				0.000	1.000
2	**Λ**	27	18.39	.891	15	8.78	.889	0.000	1.000
3	Λ, **Θ**	47	64.17*	.049	20	45.78**	.001	0.099	0.967
3a	Λ, Θ^{a}	46	47.18	.424	1[f]	16.99**	.000	0.031	0.998
3b	Λ, $\Theta^{a,b}$	45	36.74	.805	1[f]	10.44**	.001	0.000	1.000
3c	Λ, $\Theta^{a,b,c}$	44	26.00	.986	1[f]	10.74**	.001	0.000	1.000
4	Λ, $\Theta^{a,b,c}$, **Ψcon**	48	35.39	.912	4	9.39	.052	0.000	1.000
5	Λ, $\Theta^{a,b,c}$, **ν**, Ψcon	63	76.73	.115	15	41.34**	.000	0.072	0.973
5a	Λ, $\Theta^{a,b,c}$, ν^{d}, Ψcon	62	63.99	.407	1[f]	12.74**	.000	0.040	0.996
5b	Λ, $\Theta^{a,b,c}$, $\nu^{d,e}$, Ψcon	61	55.19	.685	1[f]	8.80**	.003	0.000	1.000
6	Λ, $\Theta^{a,b,c}$, $\nu^{d,e}$, Ψcon, **αcon**	65	59.12	.682	4	3.93	.416	0.000	1.000

Note. Restrictions in bold are tested by loglikelihood test delta chi-square. RMSEA = root-mean-square error of approximation; CFI = comparative fit index; restriction over conditions for existing groups.
[a] Residual variance Number Series, Men, Nullified. [b] Residual variance Arithmetic, Men, Nullified. [c] Residual variance Arithmetic women, stereotype threat. [d] Intercept Number Series, Women, stereotype threat. [e] Intercept Number Series, Men, stereotype threat. [f] Parameter freely estimated.
* $p < .05$. ** $p < .01$.

in model fit. These freely estimated residual variances are larger in the corresponding groups than in the other groups. In the fourth step, the factor variances of the male group and of the female group are restricted to be equal over conditions. This results in a slight, but nonsignificant, increase in chi-square. Considering the perfect values of RMSEA and CFI in Step 4, we conclude that factor variances of the sex groups are invariant over conditions. The factor variance of the female group is smaller ($\psi = 15.08$, $SE = 2.47$) than the factor variance of the male group ($\psi = 38.09$, $SE = 5.56$).

Considering the mean effects that we found by means of the MANOVA, one would expect intercept differences across groups. In the fifth step, the intercepts are restricted to be invariant across groups. This clearly results in a deterioration in model fit, with a highly significant increase in chi-square, worsening in RMSEA, and drop in CFI. Inspection of the modification indices shows that this restriction is untenable because of the intercept of the Number Series subtest in the stereotype threat condition in both sex groups (female students: MI = 12; male students: MI = 8). Indeed, freeing both parameters results in clear improvement in model fit (i.e., Steps 5a and 5b). As expected, the intercept of this difficult subtest is lower in the female group in the stereotype threat condition ($\nu_2 = 5.92$, $SE = 0.45$) than in the groups in the other conditions ($\nu_2 = 7.19$, $SE = 0.31$). In the male group, under stereotype threat this intercept is higher ($\nu_2 = 8.40$, $SE = 0.45$), thus nicely reflecting the stereotype lift effect on this relatively difficult subtest. In the sixth step, factor means of each sex group are restricted to be equal over conditions. This restriction appears tenable. The factor mean of the male groups is significantly higher than the factor mean of the female groups ($\alpha = 2.61$, $SE = 0.67$, $Z = 3.92$, $p < .001$). In terms of the pooled within-group standard deviation units of the latent factor, this difference in latent ability has an effect size of 0.52.

The current stepwise approach has the risk of path dependence, in the sense that the results of later restrictions (i.e., steps in the lower part of Table 1) may depend on the particular parameters, which were freed in previous steps because of high modification indices. In addition, within a particular test setting, one would normally test for strict factorial invariance with respect to the existing groups. Therefore, both as an illustration, and as a check, we also report tests for strict factorial invariance with respect to sex within each of the three conditions. This enables us to investigate whether these tests can differentiate between situations (i.e., conditions) in which stereotype threat is, or is not, present. Note that in this situation, it does not make sense to restrict factor variances and factor means, thus Steps 4 and 6 are skipped. The results of the tests per condition are reported in Table 8. As can be seen, in the control condition, restricting factor loadings, residual variances, and intercepts does not result in a worsening in model fit. In this condition, strict factorial invariance with respect to sex is clearly tenable. Test scores of male students and female students in this condition are therefore comparable, and sex differences in test performance can be explained by differences in factor mean ($\alpha = 3.16$, $SE = 1.28$, $Z = 2.47$, $p < .01$). This sex difference in factor mean has an effect size of 0.55, which is comparable with the effect size estimate in the six-group analysis.

In the nullified condition, restricting the residual variances leads to a clear deterioration in fit, as is evident by the significant chi-square difference between Steps 3 and 2, increased RMSEA, and lowered CFI. With the added restriction on intercepts, model fit does not appear to worsen any further, indicating that the mean structure is sex invariant. The largest modification indices are related to the residual variances of the Arithmetic and the Number Series subtests.

In the condition in which the gender stereotype was activated, we see that the baseline model (Step 1) shows sufficient fit,

Table 8
Fit Measures of Stepwise Test of Strict Factorial Invariance Over Sex Per Condition (Study 3)

Step	Restrictions	*df*	χ^2	*p*	Δdf	$\Delta\chi^2$	*p*	RMSEA	CFI
				Control condition					
1		4	2.33	.675				0.000	1.000
2	**Λ**	7	4.72	.694	3	2.39	.495	0.000	1.000
3	Λ, **Θ**	11	6.39	.846	4	1.67	.796	0.000	1.000
5	Λ, Θ, **ν**	14	10.03	.760	3	3.64	.303	0.000	1.000
				Nullified condition					
1		4	2.56	.634				0.000	1.000
2	**Λ**	7	5.04	.655	3	2.48	.479	0.000	1.000
3	Λ, **Θ**	11	18.69	.067	4	13.65**	.009	0.104	0.946
5	Λ, Θ, **ν**	14	19.42	.150	3	0.73	.866	0.071	0.962
				Stereotype threat condition					
1		4	4.72	.317				0.063	0.996
2	**Λ**	7	7.23	.406	3	2.51	.473	0.000	0.999
3	Λ, **Θ**	11	17.89	.084	4	10.66*	.031	0.113	0.958
5	Λ, Θ, **ν**	14	40.31**	.000	3	22.42**	.000	0.197	0.839

Note. Restrictions in bold are tested by loglikelihood test delta chi-square. Restrictions = equality constraints over sex group; RMSEA = root-mean-square error of approximation; CFI = comparative fit index.
* $p < .05$. ** $p < .01$.

although RMSEA is somewhat large (i.e., RMSEA > .06). Here, again, the restriction on factor loadings is not accompanied by any substantial worsening in model fit. In the third step, in which residual variances are restricted to be sex invariant, the fit does deteriorate. However, the clearest deterioration in model fit is found when mean structure is modeled (Step 5). All fit measures show that strict factorial invariance is untenable in this condition. As expected, the largest modification indices are found with the intercept of the Number Series subtest and the residual variance of the Arithmetic subtest.

Discussion

The MANOVA results indicate that stereotype threat affected the arithmetic test scores of the male and female groups in a differential manner. As expected, the clearest effect of stereotype threat was found on the difficult Number Series subtest. Female students clearly underperformed on this subtest when they were reminded of the gender stereotype that female students perform less well than male students on arithmetic ability tests. This corroborates the typical result that stereotype threat negatively affects math performance of female test takers on difficult tests (e.g., Spencer et al., 1999).

The factor analyses showed that strict factorial invariance over sex clearly failed in the stereotype threat condition. Specifically, stereotype threat resulted in bias with respect to sex in the Number Series subtest. In the nullified condition, we saw that residual variances were larger in the male group, indicating the presence of slight measurement bias with respect to male students. Perhaps this is because the instruction text had a sort of stereotype threat effect on these male students. Therefore, the instruction text (falsely) stressing the absence of sex differences appears not to create ideal test circumstances for male students. In the control condition, strict factorial invariance with respect to sex was tenable. Thus, in that condition, test scores of male and female students are comparable, and sex differences in test scores can be interpreted in terms of differences in the latent construct.

In contrast with several studies conducted in the United States (Ben Zeev, Fein, & Inzlicht, 2005; Smith & White, 2002; Spencer et al., 1999), we did not find a significant mean difference on female math performance between control and nullified conditions. This may be due to a difference in test setting. In the majority of American studies, participants were tested alone as opposed to in large mixed-sex groups. Such differences in setting are known to affect the strength of stereotype threat (Inzlicht & Ben Zeev, 2003; Sekaquaptewa & Thompson, 2003). Alternatively, gender stereotypes may be less strong in the Netherlands.

When test takers were reminded of gender stereotypes concerning math ability, this resulted in stereotype threat negatively affecting female performance and in stereotype lift positively affecting male performance. An interesting finding is that this stereotype lift effect did not reach significance in the MANOVA analysis but was clearly detected with MGCFA. In sum, the results of the MGCFA analyses clearly indicate that tests for strict factorial invariance are capable of determining whether stereotype threat plays a role in a particular test situation.

General Discussion

There is a large and still-growing body of research that supports the notion that stereotype threat can negatively affect test performance in stigmatized groups (Steele et al., 2002). The magnitude of these negative effects is often investigated in laboratory experiments in which stereotype threat can be manipulated. However, such research within real-life settings is difficult for ethical and logistical reasons (Sackett, 2003; Steele & Davies, 2003; Steele et al., 2002). Nevertheless, viewing and modeling stereotype threat effects as a source of measurement bias, the seriousness of stereotype threat for the comparability of groups can be investigated by testing for measurement invariance with respect to groups (regardless of the type of group, test setting, or test under investigation), provided, of course, that a reasonable factor structure is tenable.

Stereotype Threat as a Biasing Variable

Measurement invariance with respect to groups is an essential aspect for interpreting group differences in scores of any kind of psychological measurement. Tests for measurement invariance enable one to differentiate between group differences in the latent constructs that a certain test is supposed to measure (i.e., real-ability differences) and measurement artifacts related to group membership. We view stereotype threat as a source of measurement bias. Surely, no one would suggest that stereotype threat affects real (i.e., latent) abilities, at least not in the short term. Instead, stereotype threat affects the measurements of ability, and this is precisely what tests of measurement invariance are designed to investigate. Formally, if measurement invariance holds, and one conditions on latent ability, then there should be, by definition, no group differences in (manifest) test scores. This is clearly not the case if stereotype threat lowers scores of members of a group that is subject to negative ability stereotypes. Therefore, measurement invariance is expected to be violated if stereotype threat differentially affects test scores of groups. Note that the same applies to stereotype lift effects (Walton & Cohen, 2003) and priming effects on test scores (e.g., Wheeler & Petty, 2001). For instance, in Study 3 we saw that the stereotype lift effect of male students on the difficult subtest resulted in a heightening in the measurement intercept of this subtest. Moreover, the enhanced performance of female students on the easy test due to stereotype threat in Study 2 was also clearly detected.

Recent studies into the mediating variables of stereotype threat effects have shown that stereotype threat negatively affects WMC (Schmader & Johns, 2003) or increases disruptive mental load (Croizet et al., 2004). This research suggests that the mediatory principle underlying stereotype threat effects has a strong relation to the construct of intelligence. If indeed stereotype threat affects test performance through the construct, then this could result in stereotype threat effects that are completely collinear with the subtests' factor loadings. In that case, the relative strength of stereotype threat effects on each subtest correlates perfectly with the relation of each subtest with the construct. If this occurs, then stereotype threat effects could conceivably be accompanied by measurement invariance with respect to groups. However, constructs such as intelligence and mathematic ability are stable characteristics, and stereotype threat effects are presumably short-lived effects, depending on factors such as test difficulty (e.g.,

O'Brien & Crandall, 2003; Spencer et al., 1999). Furthermore, stereotype threat effects are often highly task specific. For instance, Seibt and Förster (2004) found that stereotype threat leads to a more cautious and less risky test-taking style (i.e., prevention focus), the effects of which depend on whether a particular task is speeded or not, or whether a task demands creative or analytical thinking (cf. Quinn & Spencer, 2001). In light of such task specificity, we view stereotype threat effects as test artifacts, resulting in measurement bias. Steele (1997) appears to subscribe to this view when he states that "stereotype threat effects may be a possible source of bias in standardized tests" (p. 622). It is an empirical question whether stereotype threat effects could ever be accompanied by measurement invariance. However, the results of the studies reported here lend support to the conceptualization of stereotype threat effects as a source of measurement bias.

It should be noted that within our empirical examples, sample sizes were rather small. The power to find subtle group differences in model parameters may therefore be low. Nevertheless, the fact that bias was clearly detected in our studies indicates that MGCFA is a powerful tool in detecting measurement bias (cf. Cheung & Rensvold, 2002; Meade & Lautenschlager, 2004), even if these effects are only present at the covariance level (Study 1). In light of the fact that measurement invariance is basically a null hypothesis (Borsboom, in press), the failure to reject measurement invariance may always be due to a lack of power. Fortunately, power studies within MGCFA can be conducted readily (Saris & Satorra, 1993).

Using MGCFA in Experiments

Our results show that MGCFA provides a fruitful means to investigate stereotype threat effects. It is unfortunate that many investigators do not go beyond mean differences as tested by ANOVA in analyzing experimental data. Variance and covariance differences are a potential source of information. For instance, the absence of an increase in residual variance of the affected subtests in Study 2 suggests that the stereotype threat effect did not vary over women (see Appendix B, Scenario 1). The effect of stereotype threat on the factor loading in the minority group in Study 1 suggests that the stereotype threat effects interacted with latent ability (see Appendix B, Scenario 3). Moreover, MGCFA allows for more specific tests of experimental effects, thereby increasing power. For example, the stereotype lift effect for male students in Study 3 did not reach significance in the MANOVA framework, yet with MGCFA, the corresponding intercept differed significantly from those in the other groups. If possible, the use of a measurement model such as MGCFA should be preferred to ANOVA. Moreover, the use of measurement models can add to our understanding of stereotype threat effects.

Many recent stereotype threat studies have been aimed at identifying the mediating factor underlying its effects on test performance (see, e.g., Smith, 2004, for an overview). The current modeling framework may greatly contribute to this exercise, because mediators such as anxiety (e.g., Ben Zeev et al., 2005), WMC (Schmader & Johns, 2003), and regulatory focus (Seibt & Förster, 2004) can be measured. Such measured mediators as well as many conceivable moderators (e.g., domain identification; Smith & White, 2001) may be incorporated in the model in a way that may eventually capture the stereotype threat factor as displayed in Figures 2–4. Lubke et al. (2003a) discussed the incorporation of covariates in the MGCFA framework. When studying mediators, this method boils down to extending the factor model by adding factors, which are believed to be responsible for the depressing effect of stereotype threat. For instance, one may measure arousal (e.g., Ben Zeev et al., 2005), add to the factor model an arousal factor (besides the ability factor), and see whether this arousal factor shows an increase in factor mean (or variance) under stereotype threat. Then, in a model that takes into account latent ability, one can test whether the stereotype threat effect on test performance is mediated by arousal. Moreover, one can compare various alternative models statistically, such as whether arousal also affects the ability factor, whether arousal fully mediates the effect, whether arousal interacts with ability, and so forth. In comparison with traditional approaches of studying mediation (e.g., Baron & Kenny, 1986), the advantage of using MGCFA lies in the fact that MGCFA allows for a differentiation between effects on measurements of ability and effects on ability itself. This distinction is of substantive interest and may have consequences for statistical power, which is often an issue in mediation analysis (cf. MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002). The flexibility of the common factor model and structural equation modeling in general to incorporate many factors, mediators, and moderators in a linear or nonlinear fashion, opens many doors that can contribute to our understanding of stereotype threat.

Understanding Measurement Bias

Of course, measurement bias may have many causes besides stereotype threat. It is important to stress that the broad definition of measurement invariance does not suppose anything about the possible causes of measurement bias. Unfortunately, measurement bias has been, and still is, mostly interpreted incorrectly in terms of item content. For instance, a test item could contain a concept (e.g., a football term such as "40-yard line") that is less known to one group (e.g., women), resulting in increased difficulty of that item for that particular group. However, measurement bias is not a fixed characteristic of a certain test or test item but a characteristic of how test scores relate to the construct that a test is supposed to measure. Although item content may be used to interpret the causes of measurement bias, the latter may be due to characteristics of test settings. Therefore, stereotype threat theory provides a better understanding of why measurement bias occurs. Unfortunately, the use of bias detection methods is rarely accompanied by theoretical expectations regarding why and how measurement bias occurs (however, see Oort, 1992). Needless to say, understanding the sources of measurement bias can increase the chances of measurement bias being detected, either when bias is studied by MGCFA or when bias is studied by item response models.

Stereotype Threat and Item Response Modeling

As we saw in our three studies, within MGCFA, the effects of stereotype threat are particularly evident in the performance on the more difficult subtests. This differential aspect of stereotype threat is also relevant to the study of measurement invariance within the framework of item response theory, in which item difficulty is modeled explicitly. The item level can be very informative in investigating stereotype threat effects, particularly when these are

viewed as sources of measurement bias. Within item response theory, several methods have been developed to investigate measurement bias, which in this respect is usually denoted by differential item functioning (DIF; see Millsap & Everson, 1993). If only difficult items are subject to the interference of stereotype threat, then this implies that easy items should be hardly affected (e.g., Spencer et al., 1999). This enables one to use easy items of tests for conditioning in testing for measurement bias with respect to stigmatized groups. In addition, only the complex or difficult items in a test would show bias in the presence of stereotype threat. Therefore, DIF analyses can also be used to investigate the effects of stereotype threat on test scores in real-life settings. In this respect, recent results of a study into DIF with respect to sex on the SAT–Math are of interest. Bielinski and Davison (1998, 2001) found that particularly difficult items are biased with respect to sex, which is consistent with the idea that stereotype threat has depressed scores of female students on this test.

Generalizability

The generality of stereotype threat effects on test performance in real-life settings is an important issue. The number of studies investigating strict factorial invariance with respect to ethnic groups is rather small (however, see Dolan, 2000; Dolan et al., 2004; Dolan & Hamaker, 2001). Clearly, there is a need for more research on this topic. If a certain test score gap is accompanied by measurement invariance (and power is not an issue), then stereotype threat is not likely to play a differential role in those particular group differences. If, however, strict factorial invariance with respect to groups is violated, then stereotype threat is one of the probable causes of measurement bias. Then, measures of mediators or moderators of stereotype threat could be used to model the sources of measurement bias (Lubke et al., 2003a).

As argued by Steele et al. (2002), it depends on the test situation, domain identification of a person, the content of the stereotype, and the kind of test whether stereotype threat has an effect on test performance. We argue that its effects are detectable by means of tests for measurement invariance, regardless of test situation. Clearly, tests for measurement invariance can be useful to investigate the seriousness of stereotype threat on test performance, particularly in high-stakes test situations. We hope that by using the current modeling approach within an experimental context, we can bridge the gap between differential psychology (with its interest in individual differences) and experimental psychology (with its interest in experimental effects) to gain a better understanding of when individual abilities are correctly reflected in test scores and when they are not (cf. Cronbach, 1957).

References

Aronson, J., Lustina, M. J., Good, C., Keough, K., Steele, C. M., & Brown, J. (1999). When White men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology, 35,* 29–46.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182.

Beilock, S. L., & Carr, T. H. (2005). When high-powered people fail: Working memory and "choking under pressure" in math. *Psychological Science, 16,* 101–105.

Bentler, P. M. (1990). Comparative fix indexes in structural models. *Psychological Bulletin, 107,* 238–246.

Ben Zeev, T., Fein, S., & Inzlicht, M. (2005). Arousal and stereotype threat. *Journal of Experimental Social Psychology, 41,* 174–181.

Bielinski, J., & Davison, M. L. (1998). Gender differences by item difficulty interactions in multiple-choice mathematics items. *American Educational Research Journal, 35,* 455–476.

Bielinski, J., & Davison, M. L. (2001). A sex difference by item difficulty interaction in multiple-choice mathematics items administered to national probability samples. *Journal of Educational Measurement, 38,* 51–77.

Blascovich, J., Spencer, S. J., Quinn, D., & Steele, C. M. (2001). African Americans and high blood pressure: The role of stereotype threat. *Psychological Science, 12,* 225–229.

Bollen, K. A. (1989). *Structural equations with latent variables.* Oxford, England: Wiley.

Borsboom, D. (in press). When does measurement invariance matter? *Medical Care.*

Brown, R. P., & Pinel, E. C. (2003). Stigma on my mind: Individual differences in the experience of stereotype threat. *Journal of Experimental Social Psychology, 39,* 626–633.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255.

Croizet, J. C., Després, G., Gauzins, M. E., Huguet, P., Leyens, J. P., & Méot, A. (2004). Stereotype threat undermines intellectual performance by triggering a disruptive mental load. *Personality and Social Psychology Bulletin, 30,* 721–731.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671–684.

Cullen, M. J., Hardison, C. M., & Sackett, P. R. (2004). Using SAT-grade and ability-job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology, 89,* 220–230.

Dagevos, J., Gijsberts, M., & van Praag, C. (2003). *Rapportage minderheden 2003* [*Report minorities 2003*]. Hague, the Netherlands: Sociaal en Cultureel Planbureau.

Dolan, C. V. (2000). Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivariate Behavioral Research, 35,* 21–50.

Dolan, C. V., & Hamaker, E. L. (2001). Investigating Black–White differences in psychometric IQ: Multi-group confirmatory factor analyses of the WISC-R and K-ABC and a critique of the method of correlated vectors. In F. Columbus (Ed.), *Advances in psychology research* (Vol. 6, pp. 31–59). Huntington, NY: NOVA Science Publishers.

Dolan, C. V., Roorda, W., & Wicherts, J. M. (2004). Two failures of Spearman's hypothesis: The GAT-B in Holland and the JAT in South Africa. *Intelligence, 32,* 155–173.

Donaldson, G. W. (2003). General linear contrasts on latent variable means: Structural equation hypothesis tests for multivariate clinical trails. *Statistics in Medicine, 22,* 2893–2917.

Elshout, J. J. (1976). *Karakteristieke moeilijkheden in het denken* [*Characteristic difficulties in thinking*]. Unpublished doctoral dissertation, University of Amsterdam, Amsterdam, the Netherlands.

Evers, A., & Lucassen, W. (1992). *Handleiding DAT '83* [DAT '83 manual]. Lisse, the Netherlands: Swets & Zeitlinger.

Hedges, L. V., & Nowell, A. (1995, July 7). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269,* 41–45.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Inzlicht, M., & Ben Zeev, T. (2003). Do high-achieving female students underperform in private? The implications of threatening environments on intellectual processing. *Journal of Educational Psychology, 95,* 796–805.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409–426.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: User's reference guide.* Chicago: Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (2003). *LISREL 8.5.* Lincolnwood, IL: Scientific Software International.

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

Kleinpenning, G., & Hagendoorn, L. (1991). Contextual aspects of ethnic stereotypes and interethnic evaluations. *European Journal of Social Psychology, 21,* 331–348.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003a). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence, 31,* 543–566.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003b). Weak measurement invariance with respect to unmeasured variables: An implication of strict factorial invariance. *British Journal of Mathematical and Statistical Psychology, 56,* 231–248.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7,* 83–104.

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

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling, 11,* 60–72.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 127–143.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58,* 525–543.

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change: Decade of behavior* (pp. 203–240). Washington, DC: American Psychological Association.

Millsap, R. E. (1997a). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2,* 248–260.

Millsap, R. E. (1997b). The investigation of Spearman's hypothesis and the failure to understand factor analysis. *Cahiers de Psychologie Cognitive, 16,* 750–757.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297–334.

Nguyen, H. H. D., 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,* 261–293.

O'Brien, L. T., & Crandall, C. S. (2003). Stereotype threat and arousal: Effects on women's math performance. *Personality and Social Psychology Bulletin, 29,* 782–789.

Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika, 6,* 150–166.

Osborne, J. W. (2001). Testing stereotype threat: Does anxiety explain race and sex differences in achievement? *Contemporary Educational Psychology, 26,* 291–310.

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,* 231–259.

Quinn, D. M., & Spencer, S. J. (2001). The interference of stereotype threat with women's generation of mathematical problem-solving strategies. *Journal of Social Issues, 57,* 55–71.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87,* 517–529.

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

Saris, W. E., & Satorra, A. (1993). Power evaluations in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 181–204). Newbury Park, CA: Sage.

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology, 85,* 440–452.

Seibt, B., & Förster, J. (2004). Stereotype threat and performance: How self-stereotypes influence processing by inducing regulatory foci. *Journal of Personality and Social Psychology, 87,* 38–56.

Sekaquaptewa, D., & Thompson, M. (2003). Solo status, stereotype threat, and performance expectancies: Their effects on women's performance. *Journal of Experimental Social Psychology, 39,* 68–74.

Smith, J. L. (2004). Understanding the process of stereotype threat: A review of mediational variables and new performance goal directions. *Educational Psychology Review, 16,* 177–206.

Smith, J. L., & White, P. H. (2001). Development of the domain identification measure: A tool for investigating stereotype threat effects. *Educational and Psychological Measurement, 61,* 1040–1057.

Smith, J. L., & White, P. H. (2002). An examination of implicitly activated, explicitly activated, and nullified stereotypes on mathematical performance: It's not just a woman's issue. *Sex Roles, 47,* 179–191.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27,* 229–239.

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

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613–629.

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

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

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 34, pp. 379–440). San Diego, CA: Academic Press.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences.* Mahwah, NJ: Erlbaum.

Stinissen, J., Willems, P. J., Coetsier, P., & Hulsman, W. L. L. (1970). Wechsler Adult Intelligence Scale (WAIS)—Dutch Edition Test Manual. Lisse, The Netherlands: Swets & Zeitlinger.

Stricker, L. W., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and sex, and standardized test performance. *Journal of Applied Social Psychology, 34,* 665–693.

Te Nijenhuis, J., Evers, A., & Mur, J. P. (2000). Validity of the Differential Aptitude Test for the assessment of immigrant children. *Educational Psychology, 20,* 99–115.

Thurstone, T. G. (1958). *Manual for the SRA Primary Mental Abilities Test 11–17.* Chicago: Science Research Associates.

Thurstone, T. G. (1962). *Primary mental abilities for Grades 9–12.* Chicago: Science Research Associates.

Verkuyten, M., & Kinket, B. (1999). The relative importance of ethnicity: Ethnic categorization among older children. *International Journal of Psychology, 34,* 107–118.

Verkuyten, M., & Thijs, J. (2004). Psychological disidentification with the academic domain among ethnic minority adolescents in the Netherlands. *British Journal of Educational Psychology, 74,* 109–125.

Vorst, H. C. M., & Zand Scholten, A. (2000). *Psychometrische analyse van metingen op het cognitieve, structurele en affectieve domein afgenomen in testweek 31* [*Psychometric analyses of measures in the cognitive, structural, and affective domains, administered during test week 31*]. (Internal report). Amsterdam, the Netherlands: University of Amsterdam, Psychological Methods Department.

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

Wheeler, S. C., & Petty, R. E. (2001). The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychological Bulletin, 127,* 797–826.

Wicherts, J. M. (2005). Stereotype threat research and the assumptions underlying analysis of covariance. *American Psychologist, 60,* 267–269.

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., van Baal, G. C. M., Boomsma, D. I., et al. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence, 32,* 509–537.

Wicherts, J. M., Van Asten, E. J., Balcombe, S. J., Boom, S. M., Van den Heuvel, B. B., & Sylva, H. (2003). *Stereotype threat and intelligence test scores of minority and majority high school students* (Internal Report). Amsterdam, the Netherlands: University of Amsterdam, Psychological Methods Department.

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods, 8,* 16–37.

Yzerbyt, V. Y., Muller, D., & Judd, C. M. (2004). Adjusting researchers' approach to adjustment: On the use of covariates when testing interactions. *Journal of Experimental Social Psychology, 40,* 424–431.

Zand Scholten, A. (2003). *Profielkeuzeadvies via internet.* [*Profile counseling via the Internet*]. Unpublished masters thesis, University of Amsterdam, Amsterdam, the Netherlands.

Appendix A

General Formulation Multigroup Confirmatory Factor Analysis Model

Let Y_{ij} denote the observed p-dimensional random column vector of subject j in group (or experimental condition) i. We specify the following linear factor model for Y_{ij}:

$$Y_{ij} = \nu_i + \Lambda_i \eta_{ij} + \varepsilon_{ij}, \tag{A1}$$

where η_{ij} is a q-dimensional random vector of correlated common factor scores ($q < p$), and ε_{ij} is a p-dimensional vector of residuals that contain both random error and unique measurement effects (Meredith, 1993). The $(p \times q)$ matrix Λ_i contains factor loadings, and the $(p \times 1)$ matrix ν_i contains measurement intercepts. It is generally assumed that ε_{ij} is p-variate normally distributed with zero means and a diagonal covariance matrix Θ_i, that is, residual terms are mutually uncorrelated. Furthermore, the vector η_{ij} is assumed to be q-variate normally distributed with mean α_i and a $(q \times q)$ positive definite covariance matrix Ψ_i. In addition, η_{ij} and ε_{ij} are assumed to be uncorrelated. Given these assumptions, the observed variables are normally distributed $Y_{ij} \sim N_p(\mu_i, \Sigma_i)$, where

$$\mu_i = \nu_i + \Lambda_i \alpha_i. \tag{A2}$$

$$\sum\nolimits_i = \Lambda_i \Psi_i \Lambda_i' + \Theta_i, \tag{A3}$$

where the superscript t denotes transposition. Equations A2 and A3 represent the implied mean vector and implied covariance matrix, respectively. In case of several correlated common factors, a sufficient number of elements in Λ_i should be fixed to zero to avoid rotational indeterminacy (Bollen, 1989; Jöreskog, 1971). In the same matrix Λ_i, q elements should be fixed to equal one to identify the variances of the common factors. Similarly, for reasons of identification, latent group differences in means instead of latent means themselves are modeled (Sörbom, 1974).

Appendix B

Measurement Bias Due to Stereotype Threat

Here we present three scenarios in which measurement bias due to stereotype threat (ST) is present. We use the one factor model presented in Equations 2–4 and the assumptions given above. We assume the presence of an unmeasured ST factor that incorporates all the mediating variables of ST. The scores on this ST factor are represented by σ. We assume that ST effects are uncorrelated with latent ability, that is, $\mathrm{Cov}(\eta, \sigma) = 0$. For clarity, we leave out person and group indices and restrict our attention to the group that is affected by ST (i.e., stigmatized group). Our aim is to highlight the effects of ST on the measurement parameters of the manifest variables. For an extensive discussion of the implications of strict factorial invariance, see Lubke et al. (2003b).

Scenario 1: ST Effects on Subtest L (See Figure 2)

Let Y_1 denote the scores on a biased Subtest L, and let Y_k denote the scores on a Subtest K that is not affected by ST. In that case, the linear model for Y_k is given by the following:

$$Y_k = \nu_k + \lambda_{k\eta}\eta + \varepsilon_k, \tag{B1}$$

where $\lambda_{k\eta}$ represents the factor loading of Y_k on the latent ability factor η. The linear model for Y_1 (i.e., scores on the affected subtest) is given by the following:

$$Y_1 = \nu_1 + \lambda_{1\eta}\eta + \lambda_{1\sigma}\sigma + \varepsilon_1, \tag{B2}$$

where $\lambda_{1\sigma}$ denotes the factor loading of Y_1 on the ST factor. Note that $\lambda_{1\sigma}$ has a negative value by definition, indicating the debilitating effect of ST on test performance on Subtest L. From this model, one can derive (see, e.g., Bollen, 1989) the following expressions for the implied variance (Var), and the expected value (E) of Y_k and Y_1, as well as the implied covariance (Cov) between Y_1 and Y_k:

$$\mathrm{Var}(Y_k) = \lambda_{k\eta}^2 \mathrm{Var}(\eta) + \mathrm{Var}(\varepsilon_k), \tag{B3}$$

$$\mathrm{Var}(Y_1) = \lambda_{1\eta}^2 \mathrm{Var}(\eta) + \lambda_{1\sigma}^2 \mathrm{Var}(\sigma) + \mathrm{Var}(\varepsilon_1), \tag{B4}$$

(*Appendixes continue*)

$$\mathrm{Cov}(Y_k, Y_1) = \lambda_{k\eta}\lambda_{1\eta}\mathrm{Var}(\eta), \quad \text{(B5)}$$

$$\mathrm{E}(Y_k) = \nu_k + \lambda_{k\eta}\mathrm{E}(\eta), \quad \text{(B6)}$$

$$\mathrm{E}(Y_1) = \nu_1 + \lambda_{1\eta}\mathrm{E}(\eta) + \lambda_{1\sigma}\mathrm{E}(\sigma), \quad \text{(B7)}$$

where $\mathrm{E}(\sigma)$ is greater than zero. Because the effects of ST (i.e., σ) are unknown and not modeled, the effects of the ST factor on Y_1 are incorporated in the measurement parameters of this subtest on the latent factor (η). This leads to measurement bias in the corresponding parameters. The residual variance of the affected subtest is larger in the stigmatized group because of the added variance of ST: $\mathrm{Var}(\varepsilon_1)^* = \lambda^2_{1\sigma}\mathrm{Var}(\sigma) + \mathrm{Var}(\varepsilon_1)$. In addition, the intercept (ν_1) in the stigmatized group would be lower because of the ST effects: $\nu_1^* = \nu_1 + \lambda_{1\sigma}\mathrm{E}(\sigma)$, reflecting increased difficulty and lowered scores of the affected subtest. Note that, because the covariance between the scores on the affected subtest and the scores on any unaffected subtest (such as Y_k) is unrelated to σ, the factor loading of the biased Subtest L (i.e., $\lambda_{1\eta}$) remains unchanged. In homogeneous samples, ST effects may not vary over persons, that is, $\mathrm{Var}(\sigma) = 0$. This would result in the absence of added variance, whereas intercept bias is still present. Furthermore, it is conceivable that the mean of the ST effect is zero, that is, $\mathrm{E}(\sigma) = 0$, resulting in the absence of intercept bias. Finally, if the mean the ST factor is negative, that is, $\mathrm{E}(\sigma) < 0$, then σ may be viewed as a stereotype lift effect (Walton & Cohen, 2003).

Scenario 2: ST Effects on Subtests L and M (See Figure 3)

Suppose that Subtests L and M are affected by ST. Let Y_1 and Y_m denote the scores on these two affected subtests. Suppose again that scores Y_k on Subtest K are unaffected by ST. The linear model for Y_k is given by (B1), whereas those for Y_1 and Y_m are as follows:

$$Y_1 = \nu_1 + \lambda_{1\eta}\eta + \lambda_{1\sigma}\sigma + \varepsilon_1, \quad \text{(B8)}$$

$$Y_m = \nu_m + \lambda_{m\eta}\eta + \lambda_{m\sigma}\sigma + \varepsilon_m. \quad \text{(B9)}$$

The implied variance and the expected value of Y_k are given by (B3) and (B6), respectively. Similarly, $\mathrm{Var}(Y_1)$ and $\mathrm{E}(Y_1)$ are given by (B4) and (B7), respectively. In addition, we derive the following expressions for Y_1 and Y_m:

$$\mathrm{Var}(Y_m) = \lambda^2_{m\eta}\mathrm{Var}(\eta) + \lambda^2_{m\sigma}\mathrm{Var}(\sigma) + \mathrm{Var}(\varepsilon_m), \quad \text{(B10)}$$

$$\mathrm{Cov}(Y_k, Y_m) = \lambda_{k\eta}\lambda_{m\eta}\mathrm{Var}(\eta), \quad \text{(B11)}$$

$$\mathrm{Cov}(Y_1, Y_m) = \lambda_{1\eta}\lambda_{m\eta}\mathrm{Var}(\eta) + \lambda_{1\sigma}\lambda_{m\sigma}\mathrm{Var}(\sigma), \quad \text{(B12)}$$

$$\mathrm{E}(Y_m) = \nu_m + \lambda_{m\eta}\mathrm{E}(\eta) + \lambda_{m\sigma}\mathrm{E}(\sigma). \quad \text{(B13)}$$

$\mathrm{Cov}(Y_k, Y_1)$ is given by (B5). The effects on residual variances and intercepts for both the affected subtests are parallel to the effects in the first scenario. Thus, the residual variances of L and M are increased, and the intercepts of L and M are lowered because of ST. In addition, the covariance between Y_1 and Y_m is now increased by the effect due to the ST factor: $\lambda_{1\sigma}\lambda_{m\sigma}\mathrm{Var}(\sigma)$. This added covariance shows up as a subdiagonal element in the residual covariance matrix. Specifically, this results in an additional covariance between the residuals of Subtests L and M: $\mathrm{Cov}(\varepsilon_1, \varepsilon_m) = \lambda_{1\sigma}\lambda_{m\sigma}\mathrm{Var}(\sigma)$. However, if the effects of ST do not vary over persons, that is, $\mathrm{Var}(\sigma) = 0$, then the bias due to ST is only apparent in between-group differences of the intercepts of the affected Subtests L and M, and the residual variances and residual covariance are unbiased.

Scenario 3: Nonuniform ST Effects on Subtest L (See Figure 4)

Nonuniform effects of ST can occur if ST effects depend on the level of latent ability. This may occur, for instance, if domain identification and latent ability are positively correlated with higher ability reflecting stronger identification with the domain and hence stronger ST effects. Suppose Subtest L is nonuniformly affected by ST, and Subtest K is again unaffected by ST. Let Y_k and Y_1 represent the scores on Subtests K and L. The usual linear model for Subtest K is given by (B1). Nonuniform ST effects on Y_1 can be modeled by adding an interaction factor $\eta\sigma$, resulting in this nonlinear expression for the affected subtest:

$$Y_1 = \nu_1 + \lambda_{1\eta}\eta + \lambda_{1\sigma}\sigma + \lambda_{1\eta\sigma}\eta\sigma + \varepsilon_1, \quad \text{(B14)}$$

where $\lambda_{1\eta\sigma}$ represents the negative factor loading of Y_1 on the interaction factor. This model gives rise to the following expressions for Y_1:

$$\mathrm{Var}(Y_1) = \lambda^2_{1\eta}\mathrm{Var}(\eta) + \lambda^2_{1\sigma}\mathrm{Var}(\sigma) + \lambda^2_{1\eta\sigma}\mathrm{Var}(\eta\sigma) + 2\lambda_{1\eta}\lambda_{1\eta\sigma}\mathrm{Cov}(\eta, \eta\sigma) + 2\lambda_{1\sigma}\lambda_{1\eta\sigma}\mathrm{Cov}(\sigma, \eta\sigma) + \mathrm{Var}(\varepsilon_1), \quad \text{(B15)}$$

$$\mathrm{Cov}(Y_k, Y_1) = \lambda_{k\eta}\lambda_{1\eta}\mathrm{Var}(\eta) + \lambda_{k\eta}\lambda_{1\eta\sigma}\mathrm{Cov}(\eta, \eta\sigma), \quad \text{(B16)}$$

$$\mathrm{E}(Y_1) = \nu_1 + \lambda_{1\eta}\mathrm{E}(\eta) + \lambda_{1\sigma}\mathrm{E}(\sigma) + \lambda_{1\eta\sigma}\mathrm{E}(\eta\sigma). \quad \text{(B17)}$$

As can be seen, this scenario leads to an increased residual variance:

$$\mathrm{Var}(\varepsilon_1)^* = \mathrm{Var}(\varepsilon_1) + \lambda^2_{1\sigma}\mathrm{Var}(\sigma) + \lambda^2_{1\eta\sigma}\mathrm{Var}(\eta\sigma) + 2\lambda_{1\eta}\lambda_{1\eta\sigma}\mathrm{Cov}(\eta, \eta\sigma) + 2\lambda_{1\sigma}\lambda_{1\eta\sigma}\mathrm{Cov}(\sigma, \eta\sigma), \quad \text{(B18)}$$

where $2\lambda_{1\eta}\lambda_{1\eta\sigma}\mathrm{Cov}(\eta, \eta\sigma)$ is negative, whereas the other terms increase the variance. Furthermore, the ST effect depresses the intercept of the affected subtest: $\nu_1^* = \nu_1 + \lambda_{1\sigma}\mathrm{E}(\sigma) + \lambda_{1\eta\sigma}\mathrm{E}(\eta\sigma)$. What most clearly characterizes the interaction effect, however, is the fact that the value of the factor loading of Subtest L is lowered because of the nonuniform effect. This effect is due to the fact that the covariance of Y_1 with all other unaffected subtests, such as Y_k, is lowered by the negative term $\lambda_{k\eta}\lambda_{1\eta\sigma}\mathrm{Cov}(\eta, \eta\sigma)$, provided that the mean of η is different from zero. If the mean of the biasing factor $\mathrm{E}(\sigma)$ is zero, then this can account for the absence of mean effects, that is, $\lambda_{1\sigma}\mathrm{E}(\sigma) = \lambda_{1\eta\sigma}\mathrm{E}(\eta\sigma) = 0$, and for the fact that the direction of the effect changes for low- and high-ability persons (cf. Figure 5). Finally, whereas the factors η and σ can have a normal distribution, the nonlinear effects lead to nonnormal distribution of Y_1. Therefore, besides the fact that kurtosis and skewness values can point toward such nonlinear effects, such nonnormality leads the normal-theory maximum likelihood estimator to show an upward bias in terms of model fit.

Received July 20, 2004
Revision received May 27, 2005
Accepted May 27, 2005 ■