Intelligence 38 (2010) 169–178

Contents lists available at ScienceDirect

Intelligence

Test anxiety and the validity of cognitive tests: A confirmatory factor analysis perspective and some empirical findings

Jelte M. Wicherts *, Annemarie Zand Scholten

University of Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 6 June 2008
Received in revised form 21 September 2009
Accepted 21 September 2009
Available online 25 October 2009

Keywords:
Confirmatory factor analysis
Measurement bias
Test anxiety
Validity
Latent variable modeling
Stereotype threat

ABSTRACT

The validity of cognitive ability tests is often interpreted solely as a function of the cognitive abilities that these tests are supposed to measure, but other factors may be at play. The effects of test anxiety on the criterion related validity (CRV) of tests was the topic of a recent study by Reeve, Heggestad, and Lievens (2009) (Reeve, C. L., Heggestad, E. D., & Lievens, F. (2009). Modeling the impact of test anxiety and test familiarity on the criterion-related validity of cognitive ability tests. Intelligence, 37, 34—41.). They proposed a model on the basis of classical test theory, and concluded on the basis of data simulations that test anxiety typically decreases the CRV. In this paper, we view the effects of test anxiety on cognitive ability test scores and its implications for validity coefficients from the perspective of confirmatory factor analysis. We argue that CRV will be increased above the effect of targeted constructs if test anxiety affects both predictor and criterion performance. This prediction is tested empirically by considering convergent validity of subtests in five experimental studies of the effect of stereotype threat on test performance. Results show that the effects of test anxiety on cognitive test performance may actually enhance the validity of tests.

1. Introduction

The predictive value of cognitive ability tests for criteria such as academic success and different aspects of job performance is normally attributed to the effects of the latent abilities that cognitive tests are supposed to measure (e.g., *g*; Gottfredson, 2002). Yet, it is well known that test anxiety may decrease cognitive ability test scores and college grades (Hembree, 1988; Zeider, 1998). Psychometric models help to elucidate the effects of such non-target variables on test performance and explain the degree to which test scores are correlated with variables of interest. Recently, Reeve, Heggestad, and Lievens (2009) studied the impact of test anxiety (and familiarity with tests)[1] on the criterion-related validity (CRV) of cognitive ability tests. Using results from Classical Test Theory (CTT) they proposed a psychometric model in which the true score components of test scores were due both to the target ability construct and to test anxiety. On the basis of correlations between test scores and test anxiety as reported in the literature, data were simulated on the basis of this model to determine the degree to which the CRV of tests was affected by test anxiety. The main result was that CRVs are mostly biased downwards because of test anxiety, and that this downward bias is particularly strong when test anxiety is related to both the test scores and the criterion. In this paper, we present the results from an alternative point of view on the effects of test anxiety on test performance and validity coefficients. Although we agree with their reading of the extant literature, we disagree with Reeve et al. on several crucial aspects of their CTT model. Our goal is to show how a more sophisticated psychometric model on the basis of Confirmatory Factor Analysis (CFA) can shed new light on the effect of test anxiety on CRV. It can do so because it allows us to specify and test in more detail theories on the manner in which the latent ability and test anxiety contribute to the

* Corresponding author. Department of Psychology, University of Amsterdam, Roetersstraat 15, 1018 WB, Amsterdam, The Netherlands.
E-mail address: j.m.wicherts@uva.nl (J.M. Wicherts).

[1] The current paper focuses on test anxiety effects, while Reeve et al. also considered test familiarity. A longer version of the present paper, which also addressed test familiarity is available upon request.

0160-2896/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2009.09.008

observed test scores and the criterion. We use CFA to model three substantive theories on the relations between test anxiety, cognitive ability, and observed scores. These CFA models result in specific implications concerning the correlations between the latent and observed variables and the CRV. We discuss the plausibility of the implications based on general research findings. The most plausible CFA models of test anxiety have implications for the CRV that stand in stark contrast with those based on the CTT model. Specifically, we demonstrate that the effects of test anxiety may lead to *increases* of the CRV and alternative validity coefficients (e.g., convergent validity) over and above the effects of the targeted construct(s). This result might seem counter-intuitive at first, but it becomes obvious when viewed together with the other model specifications determined by substantive theory. In order to empirically substantiate this prediction of increased CRV due to test anxiety, we present the results of five experiments into the effects of test anxiety due to stereotype threat on test performance. We first explain why we view CFA as superior to CTT.

2. Confirmatory factor analysis versus classical test theory

We prefer to use CFA rather than CTT for several reasons. First, CFA includes an explicit model of the relation between test scores and latent variables, which is absent in CTT (Bollen, 1989). Second, CFA allows for statistical tests of alternative models. Different models with vastly different practical and theoretical implications, such as different theoretical models of test anxiety, can be readily specified and compared using CFA (Reeve & Bonaccio, 2008; Wicherts, Dolan, & Hessen, 2005). Third, any CTT model with linear relations can be expressed in terms of CFA (Bollen, 1989). Fourth, CFA models can be extended to include non-linear effects that may be relevant to the issues of test anxiety and test familiarity. For instance, if the effects of test anxiety on test scores are a function of the level of *g*, then this entails an interaction between *g* and test anxiety. Such nonlinear CFA models (e.g., Marcoulides & Schumacker, 1998) can be used to study the effects of test artifacts on, say, the *g*-loadedness of tests (cf. Te Nijenhuis, van Vianen & van der Flier, 2007; Wicherts et al., 2005). Fifth, the distinction between manifest test scores and the underlying latent variable(s) that a test is supposed to measure (Bartholomew, 2004) is less clear in the CTT perspective. CTT centers on the outdated notion of the true score (Borsboom & Mellenbergh, 2002), which is defined as a person's expected score on a test obtained by averaging over scores obtained by repeated (independent) administrations of the test (Lord & Novick, 1968). Given this operational definition, the true score is theoretically hollow (Borsboom, 2005), which makes it difficult to define measurement bias unambiguously within the CTT framework.

3. Reeve et al.'s psychometric model

Although Reeve et al. (2009) acknowledged that CTT was not formulated to deal with latent variables, they adopted the CTT framework to study the biasing effects of test anxiety on the CRV of tests. Their model was based on the idea that the test scores *X* are a combination of different effects: (1) the true score component due to *g* (T_g), (2) a component of the true score due to test anxiety (T_A), (3) the true score component due to test familiarity (T_F), and (4) random measurement error. Their psychometric model is expressed in their Eq. (3):

$$X = (T_g + T_A + T_F) + \text{Error}.$$

In this model, true scores of *g*, anxiety, and familiarity contribute to the overall true scores. The expected correlation between X and T_A as implied by this model can be calculated readily (Bollen, 1989). Having computed this model-implied correlation between *X* and T_A in all conditions in Reeve et al.'s simulation study, we found that the mean correlation between test scores and the effects of test anxiety equaled 0.21.This positive correlation is at odds with both theoretical reasoning on the effects of test anxiety on test performance (e.g., Zeider, 1998), and the results of a good deal of empirical studies, which indicate that self-reported test anxiety and test scores are negatively correlated (Ackerman & Heggestad, 1997; Hembree, 1988). Furthermore, as we discuss in more detail in the Appendix, the effect of test anxiety on test scores *X* in the CTT model is of a different sign than the correlation between test anxiety and criterion scores (which is negative). This suppresses the CRV, especially when both the test scores and the criterion are related to test anxiety. This problem does not occur if one uses a negative weight for test anxiety, i.e., a negative factor loading. In the next section, we present an alternative model of the potentially biasing effects of test anxiety on the basis of CFA.

4. A linear CFA model of the effects of test anxiety

Measurement bias due to non-target constructs can be defined formally in a latent variables framework. Consider the effect of test anxiety on the scores of a test that is meant to measure *g*. Formally, scores on test *X* are said to be biased with respect to test anxiety if $F(X|g, A) \neq F(X|g)$, where $F(.|.)$ denotes the conditional distribution function, *X* denotes the manifest test scores, *g* denotes a given *g* score, and *A* denotes a given score on the latent variable test anxiety. This definition of measurement bias was first proposed by Mellenbergh (1989) and has explicit and testable consequences in the CFA model (Meredith, 1993). If test anxiety does not bias test scores *X*, the test scores are measurement invariant with respect to test anxiety: $F(X|g, A) = F(X| g)$. An important implication of these equations is that the presence of bias suggests that test anxiety has a *direct* influence on test scores *X*, above and beyond the effects of test anxiety on *X* that run via *g*. On the other hand, the absence of a biasing effect of test anxiety on *X* suggests that any relation between test anxiety and test scores is fully explained by *g*. This does not imply that test anxiety and scores on *X* are unrelated; they are independent *conditional* on *g*. In the absence of measurement bias, test anxiety does not directly affect performance on test *X*. Note that this definition of measurement bias differs from classic definitions of selection bias and predictive bias, which normally do not refer to latent variables (Borsboom, Romeijn, & Wicherts, 2008).

The correlation between test anxiety scores and intelligence tests scores is substantial: $r = -.23$ (Hembree, 1988) or $r = -.33$ (Ackerman & Heggestad, 1997). These correlations are open to alternative interpretations (Reeve & Bonaccio, 2008). According to the interference model

(Wine, 1971), test anxiety artificially lowers the performance on cognitive ability tests. In this case, test anxiety introduces measurement bias into the test scores. That is to say, two test takers who have the same level of *g*, but who differ in test anxiety, will differ in their expected test score, such that the person with the higher level of test anxiety has a lower expected test score than the person with the lower level of test anxiety. On the other hand, the deficit model (see Zeider, 1998) states that test anxiety does not cause lower test scores, but rather that the correlation between (measures of) test anxiety and test scores exists because people with lower ability levels tend to be higher in test anxiety. In this case, there is no measurement bias because of test anxiety. Two test takers with the same level of *g* will have the same expected score on the test, regardless of their (relative) level of test anxiety. These alternative models can be readily distinguished within the CFA framework (Reeve & Bonaccio, 2008; Wicherts et al., 2005), given multivariate test scores and an identified (Bollen, 1989) CFA model.

Before we consider the issue of test anxiety according to several different CFA models, we first introduce some assumptions and definitions. We adopt the usual assumptions in the CFA approach (linearity, normality, uncorrelated residuals etc.). We set variances of the latent variables to equal one, i.e., a common scaling constraint in CFA. The correlation between test score *X* and the criterion score *Y* as implied by the model represents the CRV.

4.1. Scenario 1. Deficit model

The top model in Fig. 1 displays the first scenario that explicates the deficit model. Recently, Reeve and Bonaccio (2008) used a CFA framework to study the relation between test anxiety and test performance and found support for this deficit model. In this scenario, test anxiety does not affect the test and criterion scores directly, because the correlation between test anxiety and *X* and/or *Y* is fully explained by *g*. In this case, the CRV is not affected in any way by test anxiety. The reason that *X* and *Y* are correlated with test anxiety is that both *X* and *Y* load on *g*, and *g* is correlated with test anxiety. In this scenario, the interpretation of the CRV can be safely based on the idea that both the test scores and the criterion performance are caused by individual differences in *g*. An interesting implication of this scenario is that the (relative) strength of the correlations of *X* and *Y* with test anxiety provides an indication of the *g* loading of *X* and *Y*. Suppose the test *X* shows a stronger correlation with test anxiety than the criterion *Y*. This is does not mean that the test is more “biased” by test anxiety than the criterion, rather it means that the test is a better indicator of *g* than the criterion is. Because the test *X* is correlated more strongly with *g*, it is also correlated more strongly with test anxiety.

In his elaborate meta-analysis, Hembree (1988) reports a correlation between test scores and test anxiety of −.23 and between Grade Point Average (GPA) and test anxiety of −.29. If Scenario 1 were true, then this result would imply that the *g* loading of GPA is higher than the *g* loading of the test. This strikes us as rather awkward, because GPA is only moderately correlated with *g*, while test scores are more strongly correlated with it. Jensen indicates that the *g* loading of tests is around 0.80 (Jensen, 1998). If we take this estimate as valid and suppose Scenario 1 was true, then the *g* loading of GPA would be impossibly high. That is, if the standardized *g* loading of 0.80 results in a correlation between test scores and test anxiety of −.23, then this implies that the correlation between *g* and test anxiety equals −.23/.8 = −.2875. This is (slightly) weaker than the correlation found by Hembree between test anxiety and GPA, which would suggest that the *g* loading of GPA is higher than 1, which is impossible. But the exact *g* loading of the test is immaterial. If we take these particular results from Hembree's meta-analysis at face value[2], Scenario 1 would have it that the *g* loading of GPA is always higher than the *g* loading of the test. Hence, combined,

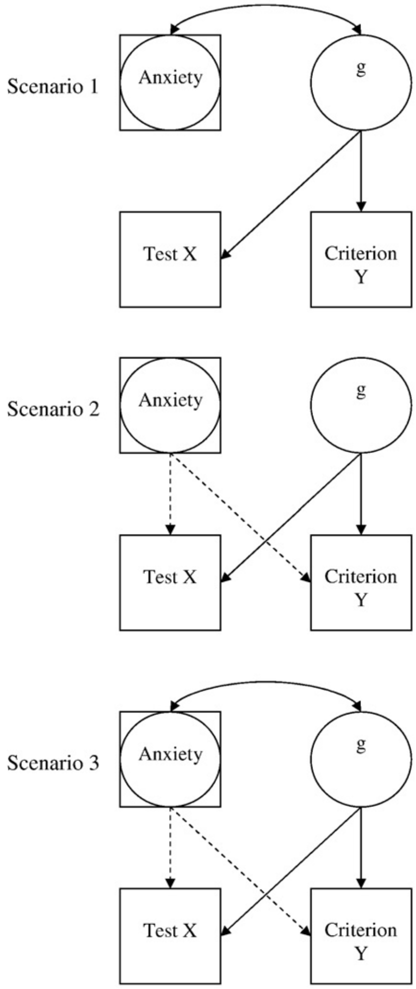

Fig. 1. Three CFA models for the effects of test anxiety.

[2] Note that this implication is based on a rather literal interpretation of some particular findings in Hembree's meta-analysis. Our interpretation is not completely consistent with Hembree's own writings. We chose these particular values from Hembree's meta-analysis because they were used by Reeve et al. and because they illustrate nicely the potential implications of the CFA model. Other values provided by Hembree may give rise to other implications.

these particular results from Hembree's, Jensen's estimate of typical *g*-loadings of cognitive tests, and the average CRV in case of GPA would appear to be at odds with Scenario 1. Another empirically testable implication of Scenario 1 is that other criteria that are (supposedly) caused by *g* (e.g., Gottfredson, 2002), such as a particular aspect of job performance, should also show a (weak) correlation with test anxiety. Thus, a correlation between some aspect of job performance and test anxiety equal to zero argues against Scenario 1.[3]

4.2. Scenario 2. Interference model

The middle diagram of Fig. 1 displays the interference model. According to this model, the correlation between test anxiety and *X* and *Y* is caused by a direct effect of test anxiety on both the test and the criterion. In this case, both test scores on *X* and *Y* are said to be biased by test anxiety. Given that these effects of test anxiety are both negative, the implication is that the CRV will be higher than is to be expected from the effects of *g* alone. This alters the interpretation in terms of latent variables of the CRV fundamentally: one of the reasons that the test predicts the criterion is that both also measure (i.e., are influenced by) test anxiety. Test anxious test takers will score lower on the test as to be expected from their level of *g*, but they will also evidence a lower average criterion performance because of their test anxiety. Their level of *g*, however, is not lower than that of others who do not suffer from test anxiety. That is, the latent factors test anxiety and *g* are uncorrelated in Scenario 2. An implication of this scenario is that the relative correlations of test scores *X* and criterion scores *Y* with test anxiety provide an indication of the relative degree to which test scores *X* and the criterion scores *Y* are affected by test anxiety. So according to Scenario 2, Hembree's findings would imply that test anxiety has a stronger effect on GPA than on test scores. Because GPAs are often derived from tasks (e.g., assignments) that are less affected by test anxiety than actual test scores, Hembree's results appear not to sit well with Scenario 2 either. Another implication of this model would be that other criteria, which are caused by *g*, but are not affected by test anxiety, would not show a correlation with test anxiety. Suppose that *g* affects a second criterion, say some aspect of job performance, in the same way as it affects GPA (i.e., both criterions are equally *g* loaded). All other things being equal, CRV of GPA would then be higher than the CRV of the other criterion, simply because the former is partly caused by test anxiety.

4.3. Scenario 3. Full model

The third scenario is a combination of the interference and deficit models. In this general scenario, which obviously is less parsimonious than the other two models, the correlation between test anxiety and *X* and *Y* is partly explained by *g*, but there are also direct effects of test anxiety on *X* and *Y*. Now the CRV is biased in the sense that the CRV is higher than expected on the basis of the *g* loadings of *X* and *Y* alone. If *X* were used in selection, test anxious test takers would be discriminated against, just like in Scenario 2, because they score lower on test *X* than expected given their level of *g*. On the other hand, according to a purely instrumental view on the CRV, the test functions well as a predictor of *Y*, because test *X* captures test anxiety just as *Y* does. Although this interpretation may not be problematic from a practical point of view, it goes against the theoretical interpretation of the CRV in terms of latent cognitive variables. By our definition of measurement invariance, the direct effect of test anxiety on test scores represents measurement bias because test anxiety and *g* are considered to be distinct entities. If, however, a lack of bias is defined in a classical framework, e.g., in terms of a lack of differential prediction, then Scenario 3 may not represent bias. We, however, subscribe to the more theoretically based definition of measurement bias (Borsboom et al., 2008), in which this scenario means that the test displays measurement bias due to test anxiety.

5. Effects of test anxiety on the validity coefficient

Reeve et al. concluded on the basis of their CTT model and simulations that "it seems unlikely that the observed relationships between cognitive ability test scores and a criterion represent overestimates of the true relationship between g and criterion performance" (p. 40). In order to study the implications of these three scenarios on the CRV in our CFA model, we fixed the *g* loading of the test at 0.8 and the *g* loading of the criterion at 0.625. In the absence of direct effects of test anxiety on both the test and the criterion, this implies (see, e.g., Bollen, 1989) a CRV of $(0.80*1*0.625=)$ 0.50. Without the effects of test anxiety, the residual variances of *X* and *Y* equal $1–.8^2=.36$ and $1–.625^2=.609$, so that the variances of *X* and *Y* equal 1. It is easy to compute the effects of any additional direct effects of test anxiety on *X* and *Y* on the CRV. Fig. 2 displays the CRV as a function of changes in the correlation between test anxiety and *g* (top panel) and changes in the factor loading of *X* on test anxiety (bottom panel). In Scenario 1 (top panel), the CRV is unaffected by test anxiety, even if the correlation between *g* and test anxiety is substantial. In Scenarios 2 and 3, the variance due to test anxiety either captures some of the residual variance (and thus reduces it) or it increases the total variance of *X*, such that the variance of test scores *X* assumes values larger than 1. If the variance of *X* increases, the standardized *g* loading decreases, but if the residual variance of *X* is partly captured by test anxiety, the standardized *g* loading of *X* remains constant at .80. In Scenario 3 (top panel in Fig. 2), the CRV normally increases as the correlation between *g* and test anxiety becomes more negative. This does not occur when (1) there is an effect of test anxiety on *X* alone, (2) test anxiety does not capture some of the residual variance in *X* (so that the variance of *X* gets larger than 1) and (3) the correlation between test anxiety and *g* lies between 0 and –.21. In Scenario 2, the CRV usually increases, but can also remain constant (when there is no effect of test anxiety on *Y*, and test anxiety captures residual variance in *X*). In Scenario 2, the CRV only lies below .50 when (1) there is no direct

[3] As rightly remarked by one of our reviewers, the relation between *g* and test anxiety in the deficit model may also be implemented as a causal path from *g* to anxiety. As long as test anxiety is not fully explained by the effects of g (and so has some variance of its own), the use of this causal path does not alter our main results.

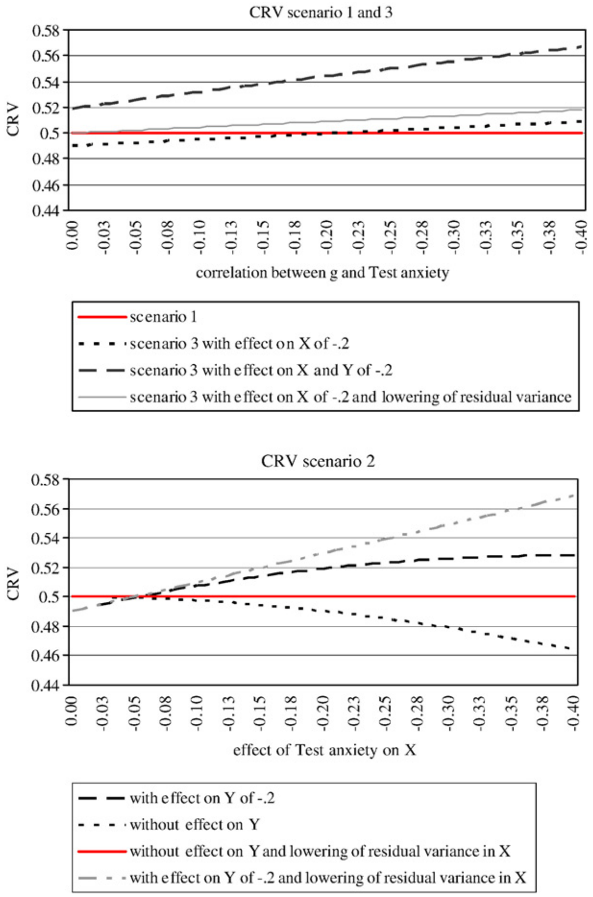

Fig. 2. Implication for the value of the Criterion Related Validity under scenarios in Fig. 1.

effect of test anxiety on *Y* and (2) test anxiety increases the total variance of *X*.[4]

So, the CRV is unaffected by test anxiety in Scenario 1. In Scenarios 2 and 3, the CRV will *increase* as the effects of test anxiety on the test and criterion increase, provided the directions of the effect of test anxiety on *X* and *Y* are the same, and the covariance between the factors has the same sign. This is completely opposite to the results based on the CTT model, which led to the conclusion that the effects of test anxiety normally reduce the CRV, especially when the criterion is related to test anxiety (see Appendix for further discussion).

6. The effects of stereotype threat on validity coefficients

According to the CTT model, the CRV goes down considerably when both the criterion and the test score are related to test anxiety. On the other hand, according to our CFA model, the CRV will become larger than the nominal value (i.e., the effect due to *g* only) under these circumstances. This prediction may be hard to study in real-life settings involving real criteria (e.g., for ethical reasons). However, experimental studies in which test anxiety is induced may shed light on this issue. The paradigm of stereotype threat is ideal to study what happens when test anxiety affects test performance over and above the effects of the targeted ability. Stereotype threat is a form of test anxiety (Jensen, 1998) induced by stressing stereotypes related to the performance of one's group (e.g., "females are bad at math")[5]. Several studies have shown that stereotype threat induces test anxiety that lowers test performance of ethnic minorities on cognitive tests, as well as the performance of females on math tests (Steele, Spencer, & Aronson, 2002). Our interest here is not in supporting or contesting the validity of the effect of stereotype threat in real-life test settings (to which these results may not generalize; Evers, Te Nijenhuis, and van der Flier, 2005; Sackett, 2003), but in investigating the effect of test anxiety on CRV in an experimental setting and we will treat the stereotype threat research results as such. Several of these studies involved more than one achievement or intelligence subtest. For instance, in one of their studies, Wicherts et al. (2005) administered four mathematics tests to female undergraduate students under conditions that differed in stereotype threat effects. Because stereotype threat effects can be expected to affect all the subtests to at least some degree, this situation is analogous to the situation in which both the test scores (e.g., the first subtest) and the criterion (e.g., the second subtest) are affected by test anxiety (i.e., Scenario 2 in Fig. 1). Hence, we consider the effects of test anxiety on the validity coefficient within the stereotyped group by comparing the correlations between experimental conditions in which test anxiety was either present (because of induced stereotype threat) or absent (i.e., control condition where stereotype threat was supposedly not at play). According to the CFA model, the correlation between subtests is expected to be higher when the effects of test anxiety are present, resulting in a higher correlation under stereotype threat conditions. In this section, we consider the results of five experimental studies on the effects of stereotype threat on test performance in which ability was measured by more than one subtest in order to test the prediction that adding test anxiety to test scores may enhance the validity coefficient.

Table 1 details the results of five stereotype threat studies in which we could compare the average inter-subtest correlations of stereotyped groups between stereotype threat and control conditions. We included these studies because we had access to the data, and because in these studies there was no indication of non-linear effects (which we checked by testing for invariance of factor loadings across experimental groups, cf. Wicherts et al., 2005). Study 1 from Wicherts et al. (2005) was not considered here because in this particular study there was a clear indication of non-linear effects, which neither our CFA model nor the CTT model can handle easily (both models are concerned with linear, non-interacting, effects). We ran a standard meta-analysis (e.g., without

[4] In addition, the CRV is smaller than .50 when the effect of test anxiety on Y is −.20 and the effect of test anxiety on X that is very small (i.e., between 0 and −.05), but we consider this an unlikely combination.

[5] Although Reeve and Bonaccio (2008) raised some doubts as to whether stereotype threat resulted in test anxiety, stereotype threat is clearly strongly related to test anxiety (e.g., Jensen, 1998; Osborne, 2001; Nguyen et al., 2003). According to Zeidner (1998), stressing lower performance of one's social group is one way to induce test anxiety.

Table 1
Comparison of inter-subtest correlations between control and stereotype threat conditions in five experiments.

Study	Sample	Test	Man.check successful?	d	N_1	N_2	Corr. (Z_r) without ST	Corr. (Z_r) with ST
Klein et al. (2007)	African immigrants	Culture Fair Test	No	.405	28	30	.62	.75
Nguyen et al. (2003)	African American undergraduates	Cognitive Ability Test	Yes	.061	43	43	.17	.64 [a]
Wicherts et al. (2005)	Female psychology students	Mathematics test battery	–	.222	88	44	.58	.57
Zand Scholten et al. (2004)	Female psychology students	GRE	Yes	.071	29	27	.09	.38
Zand Scholten et al. (2005)	Female high school students	"Kangoeroe" Math Test	Yes	.030	229	119	.22	.34
Overall meta-analytic difference between conditions					417	263	Cohen's $q = .162$ [b]	

Note: d = Effect size of mean difference between conditions with and without stereotype threat; ST = Stereotype Threat.

[a] Significant difference between conditions $p < .05$.

[b] Positive value indicates higher value in stereotype threat conditions.

corrections for restriction of range) of the correlations in Table 1, because the power to detect small differences between correlations is relatively small (Cohen, 1988). For instance, to detect a difference of .10 one would require over 1000 cases to reach significance ($p < .05$). Meta-analysis is well suited to establish such small effects. For each study, we chose to consider the correlation between the two subtests that showed the strongest decrease in mean performance due to stereotype threat, but the use of all available correlations between subtests provided similar results.

In the first experiment, Klein, Pohl, and Ndagijimana (2007) studied the debilitating effects of stereotype threat on the performance of African immigrants on Cattell's Culture Fair Test (CFT). They used a randomized between-subjects design involving four conditions, two of which were meant to increase test anxiety due to stereotype threat, and the other two were meant to diminish these effects on test performance. They found that the manipulation had a significant negative mean effect on CFT performance. The CFT consists of four subtests, and we found that the effects of stereotype threat were strongest for subtests 3 and 4 of the CFT. We computed the Fischer-transformed correlation between these two subtests within the two control conditions (combined) and in the stereotype threat conditions (combined). The correlation was lower when the effects of stereotype threat were absent than when they were present (i.e., $Z_r = .62$ versus $Z_r = .75$).

Nguyen, O'Neal, and Ryan (2003) conducted an experiment to study the lowering effects of stereotype threat on the Cognitive Ability Test (CAT) performance of a group of African American undergraduates. Although the authors failed to find significant mean effects on the CAT, a manipulation check showed that the manipulation of test anxiety was successful. We reanalyzed their data and found that for two of the three subtests, the scores went down (albeit not statistically significant). The Fischer-transformed correlation between these subtests of the CAT was markedly higher in the stereotype threat condition than in the control condition, i.e., $Z_r = .64$ versus $Z_r = .17$, respectively. This difference is significant: $Z = 2.11$, $p < .05$.

The third experiment is Study 3 from Wicherts et al. (2005), in which we randomly assigned female psychology students to conditions that differed in stereotype threat, resulting in decrements in performance on two of the four subtests that comprised the battery. In this study, the correlation between these two subtests showed no clear difference between conditions with and without stereotype threat. This suggests that the test anxiety effects showed little inter-individual variation.

The fourth study is an unpublished laboratory experiment that involved female psychology students from the University of Amsterdam (Zand Scholten et al., 2004), who completed a difficult math test composed of three subtests that were derived from GRE test prep material. Participants were randomly assigned to two conditions that differed in terms of the relevance of stereotype threat. Although the manipulation check in this study showed that the manipulation of stereotypic thoughts was successful, mean effects of stereotype threat on test performance were not statistically significant. Nonetheless, this study was well executed and can be used to study the effects of test anxiety on the convergent validity. On two of the subtests the female students showed (small) mean decrements due to stereotype threat. Again, we found that the correlation was higher under stereotype threat ($Z_r = .38$) than when stereotype threat was absent ($Z_r = .09$), suggesting that test anxiety due to stereotype threat added covariance to the test scores.

The fifth study is another unpublished study on the effects of stereotype threat on the math test performance of female students in the highest level of the Dutch high school system (Zand Scholten et al., 2005). The math test consisted of three subtests and was based on the most difficult items from the so-called Kangoeroe math test, which was designed specifically for students in this educational level. Stereotype threat was manipulated by randomly assigning students to conditions in which the sex differences on the math test were either stressed or nullified prior to taking the math test. Although the manipulation check of stereotype threat proved successful, there were no significant mean effects on test performance. Nonetheless, this large-scale study was well executed, and two of the subtests showed (small) decreases due to stereotype threat. The correlation between these two subtests was higher in the stereotype threat condition ($Z_r = .34$) than in the control condition ($Z_r = .22$).

In four of five studies (combined $N = 664$), the convergent validity of subtests was highest within the conditions where test anxiety was induced by the threat of stereotypes concerning the ability of the test-takers' social group. We performed a random-effects meta-analysis on these data with Cohen's q as the dependent variable. Cohen's q is the difference between Fischer-transformed correlations (Cohen, 1988), which we computed for each study by

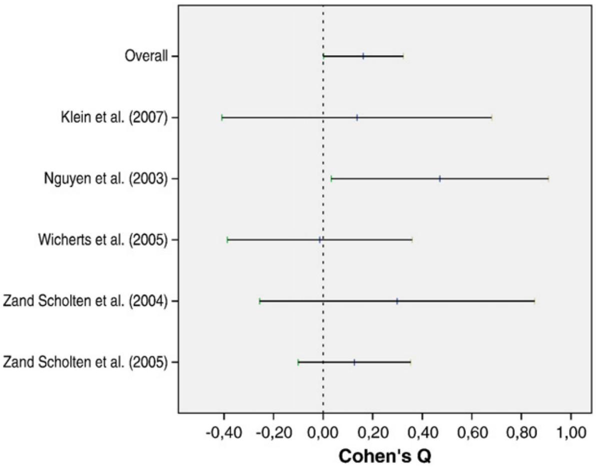

Fig. 3. Forest plot of meta-analysis of differences between correlations with or without stereotype threat.

subtracting the Z_r in the control condition from Z_r in the stereotype threat condition. Fig. 3 displays the Forest Plot of Cohen's q in each of the five studies, as well as the overall estimate. We found that the average meta-analytic difference between correlations was Cohen's $q = 0.162$, 95% confidence interval: [.001–.323], $Z = 1.97$, $p < .05$. A fixed effect model gave identical results because there was very little heterogeneity in effect sizes: $Q = 3.11$, $DF = 4$, $p = .54$. We also tested our main hypothesis with the full correlation matrices (i.e., on the basis of correlations between all subtests) and found similar results. Hence, as predicted from the CFA model, the average correlation between subtests increases under conditions in which test anxiety was experimentally induced on the test battery. This result is in line with our CFA model, but contradicts the results on the basis of the CTT model, in which the CRV went down when test anxiety was related both to the predictor test and the criterion.

According to our CFA model and the results in Fig. 2, test anxiety due to stereotype threat will have little effect on the CRV if the criterion is not affected by test anxiety. On the basis of the results in Fig. 2, we would expect a small downward bias that is caused mainly by an increase in residual variance unrelated to the target construct. This prediction is in line with the CTT model. We had at our disposal data from two stereotype threat experiments that bear on this issue. The first study was already described above and in Wicherts et al. (2005) as Study 3. The criterion for this study was GPA of the first-term of the first-year psychology students. The second study is by Brown and Day (2006) who conducted an experiment on the effects of stereotype on Raven's Advanced Progressive Matrices (APM) of African American undergraduates. They found significant mean differences on the APM between two conditions with stereotype threat and a condition with no stereotype threat. These participants also submitted their ACT scores, which may serve as a criterion. Of course, we cannot be certain that stereotype threat effects are absent on the criteria in both studies, but we consider this a reasonable assumption. As can be seen from the results in Table 2, the effects of test anxiety due to stereotype threat on the CRV were quite small. Likewise, in Wicherts et al.'s Study 3 the correlation between the math test and GPA was hardly affected by the added effect of stereotype threat. This is an interesting result, which illustrates that the criterion validity can be quite impervious to sizeable negative effects of test anxiety due to stereotype threat. It is also noteworthy that Wicherts and Millsap (2009) failed to find an indication of differential prediction between groups in both of these studies, despite the common claim (by, e.g., Sackett, Borneman, & Connelly, 2008) that test artifacts such as test anxiety necessarily result in underprediction of criterion performance for the (lower-scoring) stereotyped group. This claim has been shown to be false (Millsap, 2008).

7. Discussion

The criterion related validity (CRV) of cognitive ability tests is often thought to be due to the effects of g alone (Gottfredson, 2002), but this is not necessarily the case. In this paper, we considered the linear effects of test anxiety on cognitive ability tests scores and its implications for validity coefficients from the perspective of Confirmatory Factor Analysis (CFA). The present results show that test anxiety may improve the criterion related validity over and above the effects of g, when the criterion is also related to test anxiety. When two subtests are affected by a non-target construct in roughly the same manner, the correlation between these subtests (i.e., convergent validity) will be higher than is to be expected from the targeted construct only. These predictions were supported by the results from five experiments into the effects of stereotype threat on test performance, four of which showed that the validity coefficients were higher when test anxiety was induced by stereotype threat. Other biasing variables may have similar effects, which suggests that validity coefficients per se do not tell us much about the causal role of the latent constructs that a given test is supposed to measure. Suppose the correlation between test and criterion not only depends on how well they both measure cognitive ability, but also how well they both 'measure' other variables such as test anxiety. Then one could imagine a situation where test and criterion correlate very highly, but only because they both measure other

Table 2
Comparison of CRV between control and stereotype threat conditions in two experiments.

Study	Sample	Predictor	Criterion	N_1	N_2	Corr. (Z_r) without ST	Corr. (Z_r) with ST
Brown and Day (2006)	Black undergraduates	Raven's APM	ACT	17	36	.692	.667
Wicherts et al. (2005)	Female psychology students	Mathematics test battery	1st trimester GPA	88	44	.307	.263

variables than the target variable well. This is especially problematic when we cannot identify such biasing variables or determine their impact. In such circumstances it becomes impossible to ascertain to what extent the validity coefficient is due to the targeted ability. If the validity coefficient is high the test will probably be useful in predicting scores on the criterion, but not necessarily because they both measure the latent construct of interest. To fully understand the effects of different latent variables on test performance, psychometric modeling is clearly needed.

According to what we denoted the instrumental view on the CRV, the effects of test anxiety in Scenarios 2 and 3 may not be seen as bias if test anxiety acts to enhance the CRV above the predictive effects due to *g*. The instrumental view reflects a practical stance on test validity according to which a test is a valid predictor if it predicts well. For instance, a lack of test anxiety may be important for academic success in college and so the CRV may be high because a test that also measures test anxiety. However, these effects of test anxiety on the predictor test are problematic from both a theoretical and psychometric point of view. People with higher levels of test anxiety will systematically underperform on the test, regardless of their true latent cognitive ability. The test is meant to measure a particular latent cognitive ability, but if it also measures test anxiety, it is no longer unidimensional. In terms of a contemporary view on test validity, the test is valid if it measures what it is supposed to measure (Borsboom, Mellenbergh, & Heerden, 2004; Borsboom, Cramer, Kievit, Zand Scholten, & Franic, 2009). According to this view, (direct) test anxiety effects on test performance threaten the validity of cognitive tests, regardless of how they affect the CRV.

We based our work on the definition of measurement invariance by Mellenbergh (1989), who defined invariance in terms of latent variables. Alternative definitions of bias, such as the differential prediction definition of test bias (e.g., Sackett et al., 2008) or certain definitions of selection bias (Petersen & Novick, 1976) do not refer to latent variables but rather define bias in terms of observable test and criterion scores. The alternative definitions of invariance and bias are not consistent with each other (Borsboom et al., 2008; Millsap, 2008). Given the choice, we opt for Mellenbergh's definition because it fits contemporary conceptions of cognitive abilities (e.g., Bartholomew, 2004; Jensen, 1998) best.

Reeve et al. developed a CTT model of the effects of test anxiety and concluded that “in most real life situations […] contaminating factors such as anxiety […] will not lead to upward biases of CRV [….]. [B]ias hypotheses that claim these factors produce falsely high CRV appear to be inconsistent with our model. (p. 39)” Our results based on CFA as well as the empirical results from stereotype threat studies are at odds with this conclusion. Reeve et al. based their work on CTT. CTT fails to distinguish between observed scores and latent variables, which may result in incorrect conclusions regarding the link between true scores and constructs (Borsboom & Mellenbergh, 2002). In considering a model which features latent variables, it is better to adopt a framework that accommodates such variables explicitly (e.g., the CFA model), rather than a framework that is not developed to do so, i.e., CTT.

We tested a prediction of the CFA model by focusing on the convergent validity between cognitive tests in five experiments that involved stereotype threat as a form of test anxiety. Although the independent studies may not have been ideal, the meta-analytic results were clearly consistent with our theoretical prediction. In addition, these studies were laboratory experiments in low-stakes settings and so the results cannot be taken to mean that test anxiety or stereotype threat also enhances the actual criterion related validity on high-stakes tests or in real-life settings. The generalizability of these results to high-stakes settings remains to be studied. The study of effects of test anxiety and stereotype threat in actual test settings is hindered by ethical and practical considerations. However, as argued by Wicherts et al. (2005) and illustrated nicely by Reeve and Bonaccio (2008), CFA allows for empirical tests of particular effects such as test anxiety. For instance, Wicherts et al. (2005) found that stereotype threat effects can be detected with CFA in *any* setting because they normally result in measurement bias. The effects of non-targeted variables on test performance can not always be found with CFA because of practical considerations. For instance, the CFA model of Scenario 3 as displayed in Fig. 1 is not identified. However, by using a multivariate test battery, it is possible to put different substantive models to the test (see, e.g., Reeve & Bonaccio, 2008). Also, the use of CFA to study the effects of test anxiety requires a sufficiently large sample and a test battery that ideally fits a theoretical factor structure. Under these conditions, CFA may contribute to our theoretical understanding of the effects of test anxiety on cognitive test scores, which, in turn, may improve the use of these tests in practical settings.

In this paper, we focused on linear effects of test anxiety on test performance. We considered models that are idealizations, and it is quite likely that test anxiety effects are non-linear, or that they interact with the latent cognitive trait. Such effects require more elaborate psychometric models. For instance, suppose that tests scores are related to anxiety according to the non-linear Yerkes-Dodson's law (see, e.g., Jensen, 1998), which states that the difficulty of the test moderates the performance-lowering effect of anxiety (or arousal). If we define “test difficulty” as the expected proportion of correctly answered items in a test, it is quite clear that difficulty of a test is dependent on the ability level of the test-taker. Therefore, the Yerkes-Dodson law may be modeled so that test anxiety interacts with ability (see Wicherts et al., 2005 for some discussion of such effects). Such nonlinear effects of test anxiety may have effects on the validity of tests that are markedly different from those reported in Fig. 2. Clearly, more work on such effects is needed.

Acknowledgements

The preparation of this article was supported by VENI grant no. 451-07-016 from the Netherlands Organization for Scientific Research (NWO). We thank Raoul Grasman and Conor Dolan for their comments and Drs. Brown, Klein, and Nguyen for kindly sharing their data. We also wish to acknowledge Dr. Reeve and two anonymous reviewers for their constructive comments.

Appendix. A critique of Reeve et al.'s model

In Reeve et al.'s model, the true score of test scores X is a combination of: (1) the true score component due to g (T_g), (2) a component of the true score due to test anxiety (T_A), (3) the true score component due to test familiarity (T_F), and (4) random error:

$$X = (T_g + T_A + T_F) + \text{Error}$$

In this model, the effects of the true scores of g, anxiety, and familiarity on the overall true scores are parallel. In the original CTT framework, parallel tests were used to define important characteristics like test reliability (Lord & Novick, 1968). However, for most tests, the notion of parallelism is too restrictive and empirically untenable. In the current model, it would imply that roughly one third of the structural variance in test scores is due to each of the three sources. Acknowledging that this is unlikely, Reeve et al. did the following in their simulation study. They multiplied the effects of test anxiety and test familiarity by 0.28 or 0.50 (depending on the condition) and subsequently added the effects to arrive at the test scores X. Thus, in the conditions in which they multiplied the true scores of these variables by .50, their model may be written as follows:

$$X = 1{*}T_g + 0.50{*}T_A + 0.50{*}T_F + E$$

The fundamental problem with this approach is that the effect of the true score of test anxiety on the overall true score remains positive. This might be correct in the CTT framework, but it is empirically and theoretically awkward. This model implies that although the true scores of test anxiety and the true scores of g are negatively correlated, the effect of the true score of test anxiety on the overall "true" test scores X is positive. In other words, the implication of the model by Reeve et al. is that higher anxiety results in higher test scores. In his review of the current paper, Reeve indicated that it is to be expected that the true score components are positively correlated with the composite. Within CFA, however, this need not be the case, because factor loadings can assume negative values.

The true score is a confusing concept for several reasons (Borsboom & Mellenbergh, 2002; Borsboom, 2005) and we think we would do a good service to the field of psychometrics to get rid of it. In our view, true score component due to test anxiety in Reeve et al.'s model confuses the latent factor test anxiety and the direct effect of test anxiety on the test score. To elucidate this, let us denote the true score components as products of the factor loading of test X on the latent variables of interest:

$$\begin{aligned} T_g &= \lambda_{Xg}{*}G, \\ T_A &= \lambda_{XA}{*}A, \\ T_f &= \lambda_{XF}{*}F, \end{aligned}$$

where λ_{Xg}, λ_{XA}, and λ_{XF} denote the factor loading of test X on the following factors G (g), A (test anxiety), and F (test familiarity). Thus, in terms of CFA, Reeve et al.'s model may be expressed as follows:

$$X = \lambda_{Xg}{*}G + 0.50{*}\lambda_{XA}{*}A + 0.50{*}\lambda_{XF}{*}F + \text{Error}.$$

The factor loadings in Reeve et al. are all equal to one, but because of the multiplication of .50 for the effects of test anxiety and test familiarity, both these factor loadings may be given the values .50. The implications of their model for the CRV (i.e., the correlation between X and Y) are given in their Eq. (9):

$$r_{xy} = \frac{\sigma_{T_gT_Y} + \sigma_{T_AT_Y} + \sigma_{T_FT_Y}}{\sqrt{(\sigma^2_{T_g} + \sigma^2_{T_A} + \sigma^2_{T_F} + 2\sigma_{T_gT_F} + 2\sigma_{T_gT_F} + 2\sigma_{T_AT_F}) + \sigma^2_{E_X}} \bullet \sqrt{\sigma^2_{T_Y} + \sigma^2_{E_Y}}},$$

where $\sigma^2_{E_X}$ denotes the residual variance of X, and the sum: $\sigma^2_{T_Y} + \sigma^2_{E_Y}$ denotes the variance of criterion scores Y. Reeve et al. describe the remaining terms in this equation as follows: "$\sigma^2_{T_g}$ is the true score variance due to g, $\sigma^2_{T_A}$ is the true test score variance due to test anxiety, $\sigma^2_{T_F}$ is the true test score variance due to test familiarity, $\sigma_{T_gT_A}$ is the covariance of g and test anxiety, $\sigma_{T_gT_F}$ is the covariance of g and test familiarity, and $\sigma_{T_AT_F}$ is the covariance between test anxiety and test familiarity" (p. 7). Note that the variance components are defined in terms of the variance of the true score components, while the covariances are defined as the covariance between the *constructs* g, test anxiety, and test familiarity. However, in terms of CFA, the covariance between the true score components associated with particular constructs (i.e., COV($\lambda_{Xg}g$, $\lambda_{XA}A$)) is not the same as the covariance between the constructs (i.e., COV(g,A)). Because λ_{Xg} and λ_{XA} are constants, the covariance between the true score components should be as follows: COV($\lambda_{Xg}g$, $\lambda_{XA}A$) $=\lambda_{Xg}\lambda_{XA}$COV($g$,$A$). The covariance between constructs can be negative, whereas the covariance between the true score components can be positive, as in this case (i.e., because λ_{XA} and COV(g,A) are negative, while λ_{Xg} is positive). This has important implications for the effects of test anxiety on the CRV. In terms of a CFA model, their Eq. (9) should read as:

$$r_{XY} = \frac{\lambda_{Xg}\lambda_{YC}\psi_{gC} + \lambda_{XA}\lambda_{YC}\psi_{AC} + \lambda_{XF}\lambda_{YC}\psi_{FC}}{\sqrt{\lambda^2_{Xg}\psi_{gg} + \lambda^2_{XA}\psi_{AA} + \lambda^2_{XF}\psi_{FF} + 2\lambda_{Xg}\lambda_{XA}\psi_{gA} + 2\lambda_{Xg}\lambda_{XF}\psi_{gF} + 2\lambda_{XA}\lambda_{XF}\psi_{AF} + \theta_X} \bullet \sqrt{\lambda_{YC}\psi_{CC} + \theta_Y}}$$

(where λ_{YC} denotes the loading of the criterion scores Y on the factor associated with the criterion C, ψ_{AC} denotes the covariance between the criterion factor and test anxiety, and ψ_{FC} denotes the covariance between the criterion factor and test familiarity). This is entirely different from Eq. (9) by Reeve et al. because of the following. Both λ_{XA} and ψ_{AC} will normally be negative, indicating the negative effect of test anxiety on test scores X and the negative correlation between the criterion factor and test anxiety. For that reason the term $\lambda_{XA}\lambda_{YC}\psi_{AC}$ (which is related to the covariance between test anxiety and the criterion factor; ψ_{AC}) is positive instead of negative (as in Reeve et al.); this term does not lower the covariance between X and Y. It *increases* the covariance related to the CRV. Thus, the reason that the CRV in most of Reeve et al.'s scenarios was lowered because of test anxiety (and test familiarity) lies in the fact that the effects of test anxiety on test scores X are of a different sign than the correlation between the criterion factor and test anxiety. It should be noted that Reeve et al.'s model does not distinguish between the correlation of test anxiety with the criterion factor and the causal effect of test anxiety on criterion scores. We consider this an important distinction.

We were able to replicate the findings in Table 1 of Reeve et al. (2009) by means of direct computation using our CFA implementation of their model (additional results are available upon request). The model-implied correlation between X and test familiarity was .346 on average across the conditions of their simulations study. This is twice the empirical correlation between test familiarity and test scores (i.e., $r = .17$). Note that it might be argued that the model by Reeve et al. is in fact a CFA model. The only reason that their model may still be considered a CTT model is that it uses the true score notation and it does not explicitly use factor loadings. Negative factor loadings are required for modeling the negative impact of test anxiety on test performance. Because CTT does not entail factor loadings, classic CTT formulas cannot be used to model these negative effects. In CFA, negative effects can be accommodated readily.

References

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin*, *121*, 219–245.

Bartholomew, D. J. (2004). *Measuring intelligence. Facts and fallacies.* Cambridge, UK: Cambridge University Press.

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

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge: Cambridge University Press.

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Zand Scholten, A., & Franic, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity* (pp. 135–170). Charlotte, VA, US: Information Age Publishing.

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence*, *30*, 505–514.

Borsboom, D., Mellenbergh, G. J., & Heerden, Van (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.

Borsboom, D., Romeijn, J. W., & Wicherts, J. M. (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods*, *13*, 75–98.

Brown, R. P., & Day, E. A. (2006). The difference isn't Black and White: Stereotype threat and the race gap on Raven's Advanced Progressive Matrices. *Journal of Applied Psychology*, *91*, 979–985.

Cohen (1988). *Statistical power analysis for the behavioral sciences*, (2nd Edition), Hillsdale, NJ: Lawrence Erlbaum Associates.

Evers, A., Te Nijenhuis, J., & Van der Flier, H. (2005). Ethnic bias and fairness in personnel selection: Evidence and consequences. In A. Evers, N. Anderson, & O. F. Voskuijl (Eds.), *The Blackwell handbook of personnel selection* (pp. 306–328). Oxford, United Kingdom: Blackwell.

Gottfredson, L. S. (2002). Where and why g matters: Not a mystery. *Human Performance*, *15*, 25–46.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research*, *58*, 47–77.

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger.

Klein, O., Pohl, S., & Ndagijimana, C. (2007). The influence of intergroup comparisons on Africans' intelligence test performance in a job selection context. *The Journal of Psychology*, *141*, 453–467.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Marcoulides, G. A., & Schumacker, R. E. (1998). *Interaction and nonlinear effects in structural equation modeling.* Mahwah, NJ, US: Lawrence Erlbaum Associates.

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.

Millsap, R. E. (2008). Invariance in measurement and prediction revisited. *Psychometrika*, *72*, 461–473.

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.

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

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, *13*, 3–29.

Reeve, C. L., & Bonaccio, S. (2008). Does test anxiety induce measurement bias in cognitive ability tests? *Intelligence*, *36*, 526–538.

Reeve, C. L., Heggestad, E. D., & Lievens, F. (2009). Modeling the impact of test anxiety and test familiarity on the criterion-related validity of cognitive ability tests. *Intelligence*, *37*, 34–41.

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

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist*, *63*, 215–227.

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, US: Academic Press, Inc.

Te Nijenhuis, J., van Vianen, A. E., & van der Flier, H. (2007). Score gains on g-loaded tests: No g. *Intelligence*, *35*, 283–300.

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

Wicherts, J. M., & Millsap, R. E. (2009). The absence of underprediction does not imply the absence of measurement bias. *American Psychologist*, *64*, 281–283.

Wine, J. (1971). Test anxiety and the direction of attention. *Psychological Bulletin*, *76*, 92–104.

Zand Scholten, A., Wicherts, J. M., Korver, N., Leonardo, L., Tijsmans, N., & Trienekens, I. (2004). *Stereotype bedgreiging bij vrouwen op wiskunde taken. [Stereotype threat among females on math tasks].* Internal Report, OP4710, University of Amsterdam.

Zand Scholten, A., Wicherts, J. M., Elsenburg, F., Stoffels, M., Delsing, G., Dotsch, T., & Mulders, M. (2005). *Het effect van stereotype dreiging aangaande de scores van meisjes op een wiskundetest. [Stereotype threat effects on the scores of girls on a math test].* Internal Report, OP4810, University of Amsterdam.

Zeider, M. (1998). *Test anxiety: The state of the art.* New York: Plenum.