1-2009

Modeling the impact of test anxiety and test familiarity on the criterion-related validity of cognitive ability tests

Charlie L. REEVE
University of North Carolina at Charlotte

Eric D. HEGGESTAD
University of North Carolina at Charlotte

Filip LIEVENS
Singapore Management University, filiplievens@smu.edu.sg
DOI: https://doi.org/10.1016/j.intell.2008.05.003

Published in Intelligence, 2009 January,
Volume 37, Issue 1, Pages 34-41
http://doi.org/10.1016/j.intell.2008.05.003

Modeling the impact of test anxiety and test familiarity on the criterion-related validity of cognitive ability tests ☆

Charlie L. Reeve [a,*], Eric D. Heggestad [a], Filip Lievens [b]

[a] University of North Carolina Charlotte, United States
[b] Ghent University, United States

ARTICLE INFO

Article history:
Received 26 January 2007
Received in revised form 20 May 2008
Accepted 21 May 2008

Keywords:
Cognitive ability tests
Test anxiety
Test familiarity
Criterion-related validity
Classical test theory

ABSTRACT

The assessment of cognitive abilities, whether it is for purposes of basic research or applied decision making, is potentially susceptible to both facilitating and debilitating influences. However, relatively little research has examined the degree to which these factors might moderate the criterion-related validity of cognitive ability tests. To address this gap, we use Classical Test Theory formulas to articulate how test anxiety and test familiarity can influence observed scores, observed score variance, and most importantly, the criterion-related validity of observed scores. The resulting equations reveal that understanding the influence of test anxiety and test familiarity on criterion-related validity coefficients requires the consideration of a number of additional parameters. To elucidate the implications of the model, we present a Monte Carlo simulation. Results show that anxiety and familiarity can have a significant negative effect on the observed criterion-related validity, but also show that this effect is highly variable. In particular, the effect depends heavily upon the relation between these factors and the criterion variable. Additionally, we note that the equations we develop highlight important gaps in the literature; there are few clear empirical estimates of several of the parameters in our formulas. We call for future research to better examine these additional relations.

The use of ability tests is common in both educational and employment settings due to their robust capability to predict important outcomes (Jensen, 1998; Kuncel, Hezlett, & Ones, 2001; Schmidt & Hunter, 1998). However, concomitant relations between test scores and non-ability factors (e.g., affective traits, socio-economic indicators) continue to fuel concerns that ability test scores and/or associated predictive validity coefficients may be biased. In particular, substantial attention has been given to examining the influence of test anxiety and test familiarity[1] (aka, test-specific knowledge, test sophistication). Although both test anxiety and test familiarity are concepts with multiple dimensions, it can generally be stated that test anxiety is negatively related to ability test performance (Hembree, 1988; Zeidner, 1995), and test familiarity is positively, though weakly, related to ability test performance (Anastasi, 1981; Kulik, Bangert-Drowns, & Kulik, 1984). To the extent that test anxiety has a negative influence on ability test performance, and to the extent that test familiarity has a positive influence on performance, the use of such tests in applied settings may result in the biased assessment, placement, or selection of test-takers. For example, in the context of college admissions, high ability applicants who suffer from test anxiety may be inappropriately rejected while lower ability applicants with high levels of test familiarity may be inappropriately accepted.

Despite these concerns, little research has explicitly examined how these non-target factors might impact the criterion-related validity (CRV) of cognitive ability tests. Rather most of the extant research has investigated the relationships between test performance and test anxiety or test familiarity, or the impact of various test preparation activities on ability test performance. Although

☆ We wish to acknowledge the helpful comments and suggestions provided by the PHRRG members and Silvia Bonaccio.

* Corresponding author. Department of Psychology, University of North Carolina Charlotte, 9201 University City Boulevard, Charlotte, NC 28223-0001, United States.
E-mail address: clreeve@uncc.edu (C.L. Reeve).

[1] For brevity of expression, we use the term 'test familiarity' in a general sense to encompass all construct-irrelevant test-specific "knowledge" that is not associated with the actual ability being measured. Although we recognize there are differences between concepts such as test-taking skills, test-wiseness, test familiarity, and test-specific variance, all of these concepts do share a common theme. Namely, all of these concepts reflect, in various forms, a performance-facilitating factor that is theoretically independent of the ability measured by the test.

these studies are valuable for understanding the nexus of correlates of ability test performance, they do not by themselves demonstrate whether and to what degree these factors (or changes in these factors) artificially inflate or deflate the observed CRV coefficients. Without direct empirical evidence, both scholars and practitioners are without a clear understanding of how such factors might result in flawed estimates of CRV. This lack of understanding is especially concerning for those who use ability tests in selection contexts based on the assumption that the CRV of test scores has been accurately established. If test anxiety and test familiarity significantly alter the CRV of ability test scores, practitioners would need to be cognizant of and consider ways to reduce differences in these factors.

Moreover, few contemporary studies of the effects of test anxiety and test familiarity have formalized the conceptual models in mathematical terms (see Jensen, 1998; te Nijenhuis, van Vianen, & van der Flier, 2007, for exceptions). We believe that the failure to formalize conceptual arguments in a clearly articulated psychometric model explaining how test anxiety and test familiarity affects observed scores, observed score variance, and the CRV coefficient is a notable omission in the current literature. Typically, only arguments regarding the conceptual processes that might lead to changes in performance are provided, and correlations with these non-ability factors and test performance are reported. Unfortunately, the examination of correlations with observed scores alone does not provide insight into how the inclusion of contaminating factors might influence the CRV of the test scores. While empirical research is ultimately required, it is possible to gain a better understanding of how and to what degree test anxiety and test familiarity might impact the CRV of ability assessments through the clear articulation of a psychometric model. As such, we present a psychometric model based on Classical Test Theory that directly incorporates these two non-ability factors (namely, test anxiety and test familiarity). Following the articulation of that model, we present the results of a Monte Carlo simulation to demonstrate the potential implications of these two factors on CRV coefficients.

1. A psychometric consideration

In the following sections we use Classical Test Theory to elucidate how test anxiety and test familiarity influence observed scores on cognitive ability tests, the variance in those scores across a sample of test takers, and most importantly, the CRV of those scores.

1.1. Classical Test Theory and Observed Variance in Ability Test Scores

The central premise of classical test theory is that an individual's observed score on a test can be expressed as:

$$X = T + E \tag{1}$$

where X is the observed score (i.e., the score the individual obtains on the assessment), T is the true score, and E is error. True score can be defined as the mean score for an individual if he or she were to complete the assessment across a large number of identical testing situations (i.e., under the same testing conditions, in the same physical and psychological state). Thus, assuming the test is construct valid, the true score can be thought of as reflecting the individual's position on the latent dimension(s) the test measures (note, true score contains all systematic sources of variance and is thus not the same thing as the concept of a latent variable or a construct; however, we make the assumption that a large portion of the systematic variance is due to the target construct). Error is defined as the deviation of any observed score from the true score. Error, furthermore, is defined to be a random, normally distributed variable with a mean of zero; that is, across a large number of theoretical testing occasions, an individual's distribution of error scores would have a mean of zero. Extrapolating to a population of test-takers, it would be expected that true and error scores would be uncorrelated and that the mean of the error scores in the population would also be zero. If Eq. (1) holds for each individual in the sample, then the observed variance in a sample can be said to be a function of the variance in true scores and the variance in random errors:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \tag{2}$$

When assessing general cognitive ability, it is often assumed that true score is synonymous with standing on the ability (i.e., g). Such an assumption is not technically correct from the perspective of classical test theory, however. That is, as noted by Crocker and Algina (1986), "Any systematic errors or biasing aspects of a particular test for an individual contribute to that person's psychological true score on that test" (p. 110). Thus, from a technical standpoint, "true score" is defined as a composite of all systematic sources of variance. In the current context, the assumption is that there is one random source of variance (i.e., random measurement error) and three systematic sources of variance: the construct of interest (in this case, g), a construct-irrelevant debilitating factor (i.e., test anxiety), and a construct-irrelevant facilitating factor (i.e., test familiarity). Therefore, we can re-specify Eq. (1) as follows:

$$X = (T_g + T_A + T_F) + E \tag{3}$$

where X is the observed score, T_g is the component of true score due to g, T_A is the component of true score due to test anxiety, T_F is the component of true score due to test familiarity and E is random error variance. Accordingly, Eq. (2) can be re-expressed as follows:

$$\begin{aligned}\sigma_X^2 =&(\sigma_{T_g}^2 + \sigma_{T_A}^2 + \sigma_{T_F}^2 + 2\sigma_{T_gT_A} + 2\sigma_{T_gT_F} + 2\sigma_{T_gE_X} \\ &+2\sigma_{T_AT_F} + 2\sigma_{T_AE_X} + 2\sigma_{T_FE_X}) + \sigma_{E_X}^2\end{aligned} \tag{4}$$

where $\sigma_{T_g}^2$ is the true score variance due to g, $\sigma_{T_A}^2$ is the true score variance due to test anxiety, $\sigma_{T_F}^2$ is the true 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 of test anxiety and test familiarity. Of course, any covariance term with an error component (E_X) equals zero by definition, thus Eq. (4)[2] can be simplified as:

$$\sigma_X^2 = \left(\sigma_{T_g}^2 + \sigma_{T_A}^2 + \sigma_{T_F}^2 + 2\sigma_{T_gT_A} + 2\sigma_{T_gT_F} + 2\sigma_{T_AT_F}\right) + \sigma_{E_X}^2 \tag{5}$$

In comparing Eq. (5) to Eq. (2), we see that the variance in observed scores is not just a function of individual differences

[2] Please note that the derivation of Eq. (4) makes the simplifying assumption that there are not any higher-order interactive effects among the true scores. Thus, we include zero-order relations (i.e., covariances) among these factors, but we do not model two-way or three-way interactions.

in *g*, anxiety and test familiarity, but also of the degree to which those factors covary.

1.2. The influence on criterion-related validity

To understand how test anxiety and test familiarity might influence the CRV coefficients of observed ability test scores, we start with the standard expression of the correlation between a predictor, *X*, and a criterion, *Y*:

$$r_{XY} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \cdot \sqrt{\sigma_Y^2}}} \tag{6}$$

Next, we can replace various terms in Eq. (6) with the equations derived above. First, in the numerator, we can substitute *X* with the expression shown in Eq. (3) and *Y* with the Classical Test Theory expression shown in Eq. (1) (except here we are dealing with *Y* rather than *X*). Second, in the denominator, we can substitute σ_X^2 with the expression shown in Eq. (5) and σ_Y^2 can be replaced with the Classical Test Theory expression shown in Eq. (2). Thus, Eq. (6) can be rewritten as:

$$r_{XY} = \frac{\sigma_{(T_g+T_A+T_F+E_X)(T_Y+E_Y)}}{\sqrt{\left(\sigma_{T_g}^2 + \sigma_{T_A}^2 + \sigma_{T_F}^2 + 2\sigma_{T_gT_A} + 2\sigma_{T_gT_F} + 2\sigma_{T_AT_F}\right) + \sigma_{E_X}^2 \cdot \sqrt{\sigma_{T_Y}^2 + \sigma_{E_Y}^2}}} \tag{7}$$

Expanding the numerator using basic algebra, we get the following:

$$r_{XY} = \frac{\sigma_{T_gT_Y} + \sigma_{T_gE_Y} + \sigma_{T_AT_Y} + \sigma_{T_AE_Y} + \sigma_{T_FT_Y} + \sigma_{T_FE_Y} + \sigma_{E_XT_Y} + \sigma_{E_XE_Y}}{\sqrt{\left(\sigma_{T_g}^2 + \sigma_{T_A}^2 + \sigma_{T_F}^2 + 2\sigma_{T_gT_A} + 2\sigma_{T_gT_F} + 2\sigma_{T_AT_F}\right) + \sigma_{E_X}^2 \cdot \sqrt{\sigma_{T_Y}^2 + \sigma_{E_Y}^2}}} \tag{8}$$

Once again, the covariance terms in the numerator that contain an error term are by definition equal to zero, thus we can simplify the equation as:

$$r_{XY} = \frac{\sigma_{T_gT_Y} + \sigma_{T_AT_Y} + \sigma_{T_FT_Y}}{\sqrt{\left(\sigma_{T_g}^2 + \sigma_{T_A}^2 + \sigma_{T_F}^2 + 2\sigma_{T_gT_A} + 2\sigma_{T_gT_F} + 2\sigma_{T_AT_F}\right) + \sigma_{E_X}^2 \cdot \sqrt{\sigma_{T_Y}^2 + \sigma_{E_Y}^2}}} \tag{9}$$

Comparing Eq. (9) to Eq. (6), we see that there are seven additional terms that impact the observed correlation with two appearing in the numerator and five in the denominator. Accordingly, understanding how the non-ability factors of test anxiety and test familiarity will impact the observed correlation between observed ability test scores and a criterion variable is more complex than typically recognized.

1.3. Implications of Eq. (9)

As was noted above, a key concern for scientists and practitioners alike is the possibility that test anxiety and test familiarity may result in biased estimates of ability or validity. To the extent test anxiety artificially reduces performance on ability tests and test familiarity artificially increases performance on ability tests, it is often believed that the use of such tests in applied contexts may result in the biased evaluation of test takers. We believe the equations presented here can help inform the discussion on important theoretical questions concerning these factors.

For example, Eq. (9) can be used to understand whether and to what degree test anxiety might negatively impact on the observed CRV of cognitive ability tests. Similarly, Eq. (9) can be used to better understand the extent to which the observed CRV of cognitive ability tests varies as a function of test anxiety and test familiarity. Importantly, Eq. (9) shows that it is not just the variance due to these factors that will influence the observed CRV; rather there are several additional elements to be considered. In this sense, we believe Eq. (9) also has theoretical implications in that it highlights the need to ask additional questions that are rarely asked in the testing literature. For example, to what extent does the relationship between test anxiety (or test familiarity) and the criterion impact the observed CRV of cognitive ability tests?

Additionally, the derivation of a psychometric model (i.e., Eq. (9)) allowed us to identify all the pieces of information that are needed for a comprehensive understanding of how and when non-ability factors might negatively impact the use of ability test scores. Even a brief perusal of the literature will reveal that there are key gaps in our knowledge which prevent this comprehensive understanding. That is, there does not appear to be much research on several of the additional parameters shown in Eq. (9) (a point we will make salient below). Thus, we believe the use of such equations can have a meaningful contribution on ability testing research.

2. A simulation study of the implications of Eq. (9) on CRV

The primary purpose of this paper is the articulation of a psychometric model of the impact of anxiety and test familiarity on CRV using Classical Test Theory. Eq. (9) is the result of that effort. Above we suggested several ways that Eq. (9) might have implications for important theoretical and applied questions. To help elucidate some of these implications, we present a brief Monte Carlo simulation to show how the seven focal parameters in Eq. (9) influence CRV coefficients.

2.1. Constants

There are 12 parameters in Eq. (9). Five of these were held constant for the current study. In particular, we set the correlation between T_g and the criterion variable (*Y*) to be r_{gY}=.50. This value was chosen as a rough average of estimated ρ between *g* and useful criteria ranging from school grades, to job training, to job performance in moderately complex jobs (see Jensen, 1998; Schmidt & Hunter, 1998). In addition, the variance for *g* and the observed criterion variable were based on standardized scores (i.e., *M*=0, S.D.=1.0) and held constant across different conditions. Finally, to focus on the effects of the substantive variables, we set the two error variance terms equal to zero (i.e., the effects of the substantive parameters on the CRV can be directly interpreted without corrections for unreliability).

2.2. Variables[3]

Given the potentially extreme number of conditions generated by crossing seven independent variables, we restricted our simulation by having only two levels of each variable. One value for all parameters was set to zero, as this reflects an extremely conservative condition (i.e., no real relation exists). Consistent with recommendations of Paxton, Curran, Bollen, Kirby, and Chen (2001) to use realistic values in Monte Carlo simulations, the comparison set of values was based on the best plausible estimates found in the current literature. Hence, we believe they provide reasonable estimates for the purpose of demonstration. However, we again emphasize that researchers should consider what the various parameters are likely to be in their specific situation. Conceivably a wide variety of values could have been modeled. The purpose of this simulation is simply to demonstrate the potential impact of these factors on obtained CRVs as indicated by the formulas derived above, given a set of starting values. It should not be interpreted as providing real estimates of rho for any given situation. Additionally, please note that although the formulas clearly are based on variances and relations of true scores, many of the estimates used in the simulation are drawn from prior research on observed scores. Though psychometric meta-analysis could provide such estimates in theory, such analyses do not exist for most of the relations in the formulas.

2.2.1. g-test anxiety relationship (r_{ga})

Ackerman and Heggestad (1997) reported a meta-analytic correlation of $r = -.33$ between general intelligence tests and test anxiety while Hembree (1988) reported a meta-analytic correlation of $r = -.23$. Furthermore, Jensen (1998) reported the typical *g*-saturation of the IQ score is at least .80 (p. 91). Thus, to gain an estimate of the correlation between *g* and test anxiety we multiplied −.28 (the average of the meta-analytic correlations) by .64 (a conservative estimate of the proportion of variance due to *g*). This results in an estimated value of $r_{ga} = -.18$. Thus, we used that value, in addition to $r_{ga} = 0.0$, for the simulation.

2.2.2. g-test familiarity relationship (r_{gf})

Evidence concerning the correlation between *g* and T_F is relative sparse. Nonetheless, we can make some reasonable estimates. First, to the extent that individual differences in *g* reflect the ability to learn and profit from experience, it is plausible that *g* and T_F will become increasingly positively correlated with increasing test experience. For example, Kulik, Kulik, and Bangert (1984) found smaller practice effects for a lower ability subgroup compared to medium or high ability subgroups, suggesting that high ability individuals benefit more from a single practice test than lower ability individuals (cf. te Nijenhuis et al., 2007). Meta-analytic results also suggest a small positive correlation between these variables. Hembree (1988) reported the effect of test-skills training on actual test performance to be an *r* equivalent of .13, and Ergene (2003) reported an *r* equivalent of .21. Please note, we are not suggesting that practice effects are *g*-loaded (see, Lievens, Reeve, & Heggestad, 2007; te Nijenhuis et al., 2007). Rather, we are simply noting that *g* tends to be correlated with learning outcomes in all domains, we thus assume that in the context of test-skills training, *g* would also be related to skill/knowledge acquisition. Thus, if we assume an average correlation of .17 between some form of method-specific knowledge acquisition and test performance, the estimated r_{gf} would be .11 (i.e., .17 multiplied by .64, which is the proportion of test variance due to *g*). Thus, for the simulation, two possible values were chosen: $r_{gf} = 0.0$ and $r_{gf} = .11$.

2.2.3. Criterion-test anxiety relationship (r_{Ya})

With respect to the relation between test anxiety and criteria normally associated with ability test scores, we can posit that it is likely to vary between zero and some moderate negative value. For example, empirical research clearly shows that scores on measures of test anxiety (almost all of which measure only the debilitating aspect of anxiety) do in fact negatively correlate with performance outcomes. Hembree's (1988) meta-analysis showed test anxiety had an estimated $\rho = -.29$ with school grades, and found that the worry component of test anxiety was negatively related to achievement test performance (estimated $\rho = -.31$). For non-test criteria such as job performance, it is unclear whether there would be any correlation. On the one hand, to the extent that one's test anxiety is specific to standardized tests, we are likely to see no relation to non-test criteria. On the other hand, to the extent that test anxiety is simply a manifestation of a more general anxiety trait (e.g., general anxiety, fear of evaluation in general), we might expect to again see negative correlations between test anxiety scores and scores on evaluative criteria. Thus, the relation between test anxiety and performance criteria likely varies between no relation and some moderately negative value such as that shown in Hembree's meta-analysis. As such, for the purpose of the simulation we chose values of $r_{Ya} = 0.0$ and $r_{Ya} = -.29$.

2.2.4. Criterion-test familiarity relationship (r_{Yf})

There is little research on the relation between test familiarity (with respect to the predictor test) and criterion performance. Intuitively though, it is difficult to imagine why familiarity with the predictor test would enhance one's performance on the variety of criteria predicted by ability tests. Even for a criterion such as GPA (which would be based partly on exams), it is difficult to imagine a correlation much higher than $r = .15$. Our reasoning here is based on the results of two meta-analyses examining the effect of test-specific skills training on test performance; Hembree (1988) reported an *r* equivalent of .13 whereas Ergene (2003) reported an *r* equivalent of .21. These provide the best existing estimates of the size of the correlation between training for a specific test and subsequent performance on that test. It is difficult to imagine any reason why such test-specific training would be more strongly related to any given criterion variable, especially if it were not another test (e.g., supervisor ratings on a job, teacher evaluations of written work). As such, we arbitrarily chose a value of $r_{Yf} = .15$ as the upper value for the relation between predictor-test-specific knowledge and any criteria. We

[3] For ease of expression, the subscripts on the correlations have been simplified by dropping the true score (i.e., *T*) notation. These should not be confused with the typical notation where the single subscripts would indicate observed variables. To help identify the difference, we have used lowercase letters on the subscripts of correlations. For example, "$_a$" should be read as T_A. Thus, the notation r_{af}, for example, reflects the correlation between anxiety true scores and test familiarity true scores. For purpose of the simulation, the criterion, Y, is the observed score. Thus, we use a capitalized Y in the subscripts throughout.

believe this is a reasonable (in fact, likely a quite liberal) value as it is slightly lower than the average estimate of the relation between test-specific training and test performance. In most cases, the correlation is likely to be null. Thus, the values chosen were: r_{Yf}=0.0 and r_{Yf}=.15.

2.2.5. Test anxiety–test familiarity relationship (r_{af})

Both the deficits model of test anxiety (e.g., Tobias, 1985) and the interference model of test anxiety (e.g., Wine, 1971) posit a negative correlation between test anxiety and test familiarity. The deficits model posits that test anxiety decreases as one acquires test-specific knowledge or otherwise increases his or her readiness. The interference model suggests that anxiety may decrease as the person becomes more accustomed to the situation and is thus less distracted by unfamiliarity with the testing situation. Consistent with these suggestions, Hembree (1988) reported a mean effect size of − .54 for systematic desensitization programs (i.e., repeatedly exposing a highly anxious student to the test lead to a reduction in test anxiety). Similarly, Hembree (1988) reported a significant reduction in test anxiety after participating in "test-wiseness" training programs with an *r* equivalent of − .27. As a result, regardless of which theory of test anxiety to which one subscribes, both theory and empirical evidence suggest that T_A is inversely related to some form of test familiarity. Because we could not find any other reasonable estimates of this relationship, we chose to use Hembree's *r* equivalent of − .27. Thus, two possible values were chosen: r_{af}=0.0 and r_{af}=− .27.

2.2.6. Variance in test anxiety ($\sigma^2_{T_A}$) and Variance in test familiarity ($\sigma^2_{T_F}$)

Defining the specific amount of true score variance due to anxiety or test familiarity was not possible from the existing literature. Additionally, it is to be expected that the different tests will have higher or lower g-saturations (i.e., more or less of the observed score variance is due to *g*). Nonetheless, to provide a simulation that reflects a realistic situation, we chose to consider IQ scores. Based on Jensen's (1998) work, it can be estimated that about 64% of the variance in observed IQ scores is due to '*g*.' As such, 36% of the variance would be due to error and other systematic (i.e., true score) factors. Thus, for a baseline condition we set σ^2_g=1.0, and the variances for $\sigma^2_{T_A}$ and $\sigma^2_{T_F}$ equal to .28 (i.e., .18/.64=.28). In this case, '*g*' accounts for approximately 64% of the observed score variance while *A* and *F* each account for approximately 18% of the observed score variance each (assuming that all of the covariance terms are 0.0). For a set of contrasting values, we chose to simulate a condition in which the variance due to g was lower and the variance due to anxiety and familiarity was relatively large. Thus, we set σ^2_g=1.0, and each of the variances for $\sigma^2_{T_A}$ and $\sigma^2_{T_F}$ equal to .50. In this case, '*g*' accounts for approximately 50% of the observed score variance while *A* and *F* each account for approximately 25% of the observed score variance each (assuming that all of the covariance terms are 0.0). Note, to minimize the complexity of the results, we did not include conditions where $\sigma^2_{T_A}$ and $\sigma^2_{T_F}$ differed from each other (i.e., we did not cross these two factors).

2.3. Data generation

A total of 48,000 samples were generated, with 750 samples (*N*=200 in each) generated in each of the 64 conditions: 2 (r_{Ya})× 2 (r_{Yf})×2 (r_{ga})×2 (r_{gf})×2 (r_{af})×2($\sigma^2_{T_A}$ and $\sigma^2_{T_F}$). More specifically, we started each condition by constructing a 6 X 6 correlation matrix consisting of the following correlations: r_{Yg}, r_{Ya}, r_{Yf}, r_{ga}, r_{gf}, r_{af}. The correlation r_{Yg}, representing the true CRV of general mental ability, was set at .50 for all conditions. The other correlations represent the first five independent variables. To derive the raw data for each sample, each of these correlation matrices was subjected to a Cholesky decomposition 750 times using the matrix facility of SPSS (SPSS, 1999). The Cholesky decomposition option in the SPSS matrix facility is a procedure that provides data for a specified number of simulated participants consistent with a particular correlation matrix (SPSS, 1999). Thus, for each of the 750 samples within each condition, we generated scores for the criterion variable (*Y*), *g*, anxiety (*A*), and test familiarity (*F*) for each of the 200 simulated individuals. Each of these variables was distributed uniformly with a mean of 0 and a standard deviation of 1. As noted, the *A* and *F* variables were then multiplied by either .28 or .5 (depending on condition) to reduce their variance relative to *g*. Thus, in the end, we had *K*=48,000 samples (750 samples × 64 conditions) and a total *N*=9,600,000.

2.4. Dependent variable

The dependent variable was the average observed correlation between the criterion scores and the observed ability test score (*X*). To calculate the observed ability test scores, we created a new variable, *X*, by summing each individual's *g*, *A* and *F* scores. The correlation between *X* and *Y* was then calculated for each sample. Thus, we had 48,000 observed correlations (r_{XY}).

3. Results

The means of the 750 CRV coefficients computed within each of the 64 conditions are presented in Table 1. Across all

Table 1
Average observed correlation between observed test score (X) and criterion (Y)

r_{ga}	r_{af}	r_{gf}	r_{Ya}=0.0		r_{Ya}=−.29	
			r_{Yf}=0.0	r_{Yf}=$^{+}$.15	r_{Yf}=0.0	r_{Yf}=$^{+}$.15
When σ^2_g=1.0, $\sigma^2_{T_A}$=.28, and $\sigma^2_{T_F}$=.28						
r_{ga}=0.0	r_{af}=0.0	r_{gf}=0.0	.463	.507	.386	.427
		r_{gf}=$^{+}$.11	.454	.492	.372	.416
	r_{af}=−.27	r_{gf}=0.0	.473	.513	.395	.437
		r_{gf}=$^{+}$.11	.461	.501	.383	.421
r_{ga}=−.18	r_{af}=0.0	r_{gf}=0.0	.487	.528	.407	.444
		r_{gf}=$^{+}$.11	.468	.509	.391	.437
	r_{af}=−.27	r_{gf}=0.0	.498	.537	.407	.450
		r_{gf}=$^{+}$.11	.483	.517	.399	.438
When σ^2_g=1.0, $\sigma^2_{T_A}$=.5, and $\sigma^2_{T_F}$=.5						
r_{ga}=0.0	r_{af}=0.0	r_{gf}=0.0	.406	.467	.290	.350
		r_{gf}=$^{+}$.11	.393	.452	.280	.337
	r_{af}=−.27	r_{gf}=0.0	.428	.491	.304	.368
		r_{gf}=$^{+}$.11	.411	.473	.292	.353
r_{ga}=−.18	r_{af}=0.0	r_{gf}=0.0	.435	.498	.308	.373
		r_{gf}=$^{+}$.11	.417	.480	.296	.360
	r_{af}=−.27	r_{gf}=0.0	.460	.527	.325	.394
		r_{gf}=$^{+}$.11	.439	.503	.311	.378

Note. The true correlation between *g* and *Y* was set at ρ=.50. The subscripts are as follows: 'g'=general cognitive ability; 'A'=test anxiety; 'F'=test familiarity.

64 conditions, the mean observed CRV coefficient is .42 (S.D. = .07; range .280–.537). On average, the observed CRV was 85% (S.D. = 14%; range 56%–107%) of the rho value. Again, recall that we eliminated error variance from the simulation so that these correlations are equivalent to being disattenuated for unreliability.

First, and perhaps most important, in only 6 of the 64 conditions did the mean observed CRV overestimate the true relation between *g* and the criterion variable. In all other conditions, the mean observed CRV was lesser than or equal to the true correlation of $\rho = .50$. Inflated CRVs were only observed in the scenarios where anxiety is unrelated to the criterion and where familiarity is related to the criterion. In no other scenario did the observed CRV overestimate the actual rho value. Recall from the discussion above, we believe the number of real life situations where test familiarity is related to criterion performance is quite limited. Thus, these results indicate that, in most real life situations assuming reasonable values for all the other parameters, contaminating factors such as anxiety and familiarity will not lead to upward biases of CRV. Said differently, bias hypotheses that claim these factors produce falsely high CRV appear to be inconsistent with our model.

Nonetheless, the results show that observed CRV can vary widely, with values ranging from $r = .280$ to $r = .537$. Considering that we did not explicitly model random error, these results clearly demonstrate that test anxiety and test familiarity can potentially have a serious moderating effect on observed CRVs. As our results show, the magnitude of this effect depends on the specific values used for each of the parameters in Eq. (9). Assuming we used reasonable estimates in this simulation, these results suggest that the observed CRV of ability tests can be significantly reduced in some situations, while remaining largely intact in other situations. Said differently, the simulations results confirm that CRV can be significantly underestimated in some situations.

A couple of counter-intuitive findings are also apparent. The results suggest that regardless of whether anxiety is correlated with the criterion measure, the observed validity of ability tests is higher (i.e., closer to the true correlation between *g* and the criterion) to the extent that *g* and anxiety are (negatively) correlated. In either the upper or lower section of Table 1, consider the analogous observed correlations between the conditions when anxiety is correlated with *g* compared to when it is not. When *g* and anxiety are correlated (i.e., the second and fourth quarters of the table), the observed validity coefficients (average $r = .43$) are consistently higher than the analogous condition where *g* and anxiety are uncorrelated (i.e., the rows in the first and third quarters of the tables; average $r = .41$). An explanation for these results becomes evident through an inspection of Eq. (9). Under conditions in which *g* and test anxiety are negatively correlated, this negative covariance term reduces the sum of the denominator.

Test familiarity, in contrast, had the opposite effect. In all cases, increasing the correlation between *g* and test familiarity decreased the CRV of the observed score, even when test familiarity was positively correlated with the criterion. However, it should also be noted that the CRV *increases* to the extent that test anxiety and test familiarity are negatively correlated.

A full evaluation of Table 1 suggests that effects of two parameters are particularly important; specifically, the correlation between anxiety and the criterion (r_{Ya}), and test familiarity and the criterion (r_{Yf}). The two effects involving relationships with the criterion variable are apparent by comparing values across the columns in Table 1. For example, when comparing the columns of results where $r_{Ya} = 0$ to the columns of results where $r_{Ya} = -.29$ (i.e., the two left side columns compared to the two right side columns) noticeable differences can be seen. The values shown in the left side columns are consistently larger than the corresponding values (in the same row) in the columns on the right side of the tables by an average of .10. Differences in CRVs are also apparent when comparing values (on the same row) in the columns where $r_{Yf} = 0$ to the values in the columns where $r_{Yf} = {}^{+}.15$. On average the values in these columns differ by .06. By way of contrast, the differences between CRVs are not as noticeable when making comparisons across rows within a column (i.e., moving between conditions varying parameters other than those parameters involving the relation between the criterion and anxiety or familiarity). The average difference in CRVs between conditions represented by rows is only .01. For example, the average difference in CRVs when $r_{ga} = 0$ and $r_{ga} = -.18$ is .01.

It is important to note that the specific values we obtained are obviously a function of the specific values we used in the simulation. It is possible to argue that other values could have been used. Indeed, this is one of our main points; we strongly encourage researchers to consider what are likely to be reasonable values *in their specific situation*, and apply those values to the formulas we have derived above. In this spirit, we also conducted a series of sensitivity analyses to examine the degree to which changes in some of the key estimates used would impact the pattern and substantive nature of our results. First, we ran sensitivity analyses using a higher rho (rho = .75 instead of the original .50). For this sensitivity analysis, we reran 26 conditions chosen somewhat randomly, but ensuring to cover the full range of conditions. Although the actual observed CRVs are higher, the percent of rho reflected by observed r_{xy} was basically the same. For many conditions, the percentages were exactly the same. That is, changing the value of rho had no effect on the degree to which the observed r_{xy} estimates rho. For the others, the changes were small fluctuations (small increases or decreases of 3% or so). More importantly, the overall pattern of results was the same; the conditions that lead to the lowest observed CRV (relative to the other conditions) still yielded the lowest and those that yielded the highest still yielded the highest. Said differently, the substantive conclusions regarding the impact of anxiety and familiarity were the same. (These results are not presented but are available from the first author).

Second, we re-ran the simulation changing the amount of variance in observed scores due to anxiety and familiarity. To this end, we ran all the conditions where both *A* and *F* had originally been set at .5 to be equal to *g*; that is, *g*, anxiety and familiarity each contributed 33% of the variance to the observed scores. Though an admittedly extreme case, the effect was predictable. That is, by decreasing the variance due to *g*, we essentially increased the variance in X due to anxiety and familiarity. Not surprising, the percent of rho captured by observed CRV decreased accordingly. However, the pattern of results was again the same. The conditions that gave the lowest observed CRV still gave the lowest; the ones that gave the highest, still gave the highest. The generally constant decrease in CRV led to fewer cases where CRV was overestimated, but the overall *pattern* of observed CRVs was the

same as the results shown in Table 1. (A detailed presentation of these additional analyses is available from the first author).

4. Discussion

Recall the primary purpose of the current paper was to use Classical Test Theory formulas to elucidate how test anxiety and test familiarity influence observed scores on cognitive ability tests, the variance in those scores across a sample of test takers, and most importantly, the CRV of those scores. With respect to the issue of CRV, our Eq. (9) shows how the non-ability factors of test anxiety and test familiarity will impact the observed correlation between observed ability test scores and a criterion variable; in particular, this equation reveals several additional terms that must be considered. The subsequent simulation was presented to help elucidate the implications of these additional parameters. Although the specific pattern of results from the simulation is somewhat complex, we believe they demonstrate several important implications of Eq. (9).

First, according to our model, the contaminating factors of test anxiety and test familiarity can have a significant negative impact on the observed CRV of cognitive ability tests. In fact, in the simulation we found that the observed CRV coefficient between ability test scores and a criterion measure was higher than the true CRV coefficient only in the unlikely scenarios where test-specific familiarity was correlated with criterion performance. As such, 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.

Of course, finding that the CRV is degraded in most cases is not unexpected; adding variables to a composite which is less correlated with the criterion than the initial variable (*g* in this case) will generally reduce the overall CRV. However, Eq. (9) shows, and the simulation confirms, that the degree of impact of anxiety and familiarity on observed CRV coefficients is complex and highly variable. Although the average effect is a reduction in the CRV, the results do show a large range of possibilities including a situation (though likely rare) in which the observed CRV is actually larger than that due to *g* alone. This underscores a main point of this paper: a true understanding of the influence of test anxiety and test familiarity on CRV coefficients requires the careful consideration of a number of factors as they exist in a specific situation. The exact impact will depend on the precise association among a number of parameters, the relative amount of change in those parameters across situations, and on the nature of the criterion measure itself. Hence, researchers need to consider what the most likely or reasonable values are in their situation and use the formulas to estimate the potential impact on observed CRV. The purpose of the simulation was to give a sense of the degree to which various combinations might influence CRV; it is not intended to be an exhaustive explication of all possibilities.

Third, the results of the simulation of variance clearly show that the most important determinants of the effect of test anxiety and test familiarity on CRV are the relationships between these factors and *criterion* scores. We believe this is a critically important finding given that there has been very little attention paid to these parameters. A vast majority of the existing research, theorizing, and commentaries on the impact of non-cognitive factors on the CRV of ability tests has focused exclusively on the relationships between these factors and scores on the predictor. Our results suggest that much more attention needs to be paid to the relationships between these factors and scores on the criterion variables. As just indicated, it is possible that in specific circumstances the inclusion of these non-ability factors might actually enhance observed CRV. While we acknowledge these circumstances are likely rare, the point should not be dismissed: Understanding how so-called "biasing" factors in cognitive ability tests impact the validity and utility of tests requires serious consideration of the criterion to be predicted.

4.1. Implications for future research

To conduct the simulation study, we needed to identify reasonable estimates for the seven focal parameters in Eq. (9). Although we believe we used the best available estimates, it became obvious that there was little to no empirical evidence regarding some of those parameters. We believe this potential limitation of our simulation underscores a particularly important implication of the current paper. Despite a substantial literature on the issue of test bias, test anxiety, and test familiarity, there are key gaps in our knowledge which are preventing a comprehensive understanding of how and when non-ability factors might negatively impact the use of ability test scores. The derivation of a psychometric model (i.e., Eq. (9)) allowed us to identify all the pieces of information that are needed. Thus, we call for researchers to explicitly state the psychometric model underlying their hypotheses concerning the effects of non-ability factors on ability test scores and their CRV. Second, a clear implication is that more research is needed. Specifically, we make a call for research that focuses on estimating, in real world settings, the values of key parameters identified in Eq. (9). Continued estimation of the correlation between test anxiety and observed test scores would appear to be of little value as multiple meta-analyses (e.g., Ackerman & Heggestad, 1997; Hembree, 1988) have provided good evidence for that particular parameter. Needed, however, is research examining the relationships between anxiety and the *g*-factor (as opposed to observed test scores), the components of test familiarity and the *g*-factor, and in particular, the relation between these non-ability constructs and key criterion variables.

Second, we note that Eq. (9) can also be useful as a psychometric model for understanding other test related issues, such as the impact of anxiety reduction programs, retesting, or test-skills training on the CRV of test scores. Eq. (9) makes it clear that simply hypothesizing a mean decrease or increase in test anxiety or test familiarity is insufficient. Rather, to predict the impact on CRV one needs to know how these activities influence all of the parameters in Eq. (9). Take the issue of retesting as an example. Existing arguments regarding the nature of retest effects (i.e., a systematic increase in mean scores) often cite a decrease in debilitating factors such as test anxiety or increases in facilitating factors such as test familiarity (e.g., Lievens, Buyse, & Sackett, 2005). Assuming for the moment these arguments have merit, understanding and predicting how retesting will impact test scores and their CRV requires one to have knowledge of all of the parameters in

Eq. (9) at both initial testing and retesting, or at least an idea of how parameters will change from the initial test to the retest. However, as noted previously, the current literature does not provide sufficient information to make all of these estimates with any reasonable degree of certainty. Thus again, we believe this demonstrates the utility of explicitly stating the psychometric model underlying conceptual arguments. By doing so, researchers can gain a better sense of what information is needed to fully understand a phenomenon and can make more precise estimates regarding the impact of interventions or extraneous factors on CRV. In the current paper, we have used Classical Test Theory as a foundation for our model. This is not the only possible basis for such research. Indeed, we encourage researchers to conduct similar work using other complementary frameworks such as latent trait models (e.g., factor analytic frameworks, item-response theory). We believe future endeavors such as these would continue to strengthen and enhance our theoretical understanding of how these various factors exactly influence psychometric and consequential validity of ability tests (Messick, 1989).

Finally, we would like acknowledge a number of concerns raised by a reviewer regarding or models. First, it is important to again emphasize that the statistical concept of "true score" is not the same thing as the concept of a latent variable or a psychological construct. As we noted above, from a technical standpoint, "true score" is defined as a composite of all systematic sources of variance. Thus, true score can be defined as the mean score for an individual if he or she were to complete the assessment across a large number of identical testing situations (i.e., under the same testing conditions, in the same physical and psychological state). Therefore, it is important to note that in places where we make analogies between true scores and latent variables, we are technically in error and have obscured the distinction between these concepts. Second, we acknowledge that Classical Test Theory defines CRV as the correlation between observed test scores (which is driven by shared true score variance). In our discussion, we have used the concept of a true CRV (e.g., a rho value) as reflecting the observable correlation among true score components only. Again, this is blending of CTT and latent variable frameworks which lead to some statements that are not technically correct. We acknowledge that some may find the use of CTT in general, and making analogies between CTT and latent variable frameworks, potentially confusing. However, we believe that CTT is satisfactory for our purpose which is to make salient the point that researchers need to consider additional parameters in attempts to understand how non-ability factors might influence test scores and observed CRV.

4.2. Conclusions

The primary purpose of this paper was to articulate a model that can be used to understand the likely impact of test anxiety and test familiarity on the CRV of ability tests. Using the history of differential psychology as a guide, we believe that by (re)embracing general measurement theory and using it as a foundation, research on the applied assessment of individual differences and the attendant issue of bias due to non-ability factors will be more scientifically profitable. With such a theory in hand, empirical researchers should be more apt to develop a coherent body of research by systematically investigating each parameter in Eq. (9). In doing so, we are much more likely to achieve a fundamental understanding of how these factors influence observed scores on cognitive ability tests and their corresponding CRV.

References

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

Anastasi, A. (1981). Coaching, test sophistication and developed abilities. *American Psychologist*, *36*, 1086–1093.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth, TX: Harcourt Brace Jovanovich.

Ergene, T. (2003). Effective interventions on test anxiety reduction: A meta-analysis. *School Psychology International*, *24*, 313–328.

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.

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin*, *95*, 179–188.

Kulik, J. A., Kulik, C. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal*, *21*, 435–447.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, *127*, 162–181.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology*, *58*, 981–1007.

Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting in selection settings. *Journal of Applied Psychology*, *92*, 1672–1682.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13–103)., 3d ed. New York: American Council of Education.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, *8*, 287–312.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262–274.

SPSS (1999). *SPSS 10.0 syntax reference guide.* Chicago, IL: SPSS 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.

Tobias, S. (1985). Test anxiety: Interference, defective skills and cognitive capacity. *Educational Psychologist*, *20*, 135–142.

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

Zeidner, M. (1995). Personality trait correlates of intelligence. In D. H. Saklofske, & M. Zeidner (Eds.), *International handbook of personality and intelligence* (pp. 299–319). New York, NY: Plenum Press.