

Research Report

2018-3

Keeping Your Cool: Does Test Anxiety Bias Performance on the ACT?

JEFFREY STEEDLE, PHD

Validity Threat

Debilitating test anxiety is a general threat to validity **if** it biases assessment scores. This study indicated that ACT scores were **not** biased by anxiety.

ACT.org



ACT[®]

Acknowledgements

The author acknowledges Dr. Peter Halpin of New York University for expert guidance regarding structural equation modeling.

Authors

Jeffrey T. Steedle, PhD

Jeffrey Steedle is a senior research scientist in Validity and Efficacy Research specializing in postsecondary outcomes research and validity evidence for ACT's workforce assessment programs.

Table of Contents

INTRODUCTION	1
PERSPECTIVES	1
TEST ANXIETY	1
THE INTERFERENCE AND DEFICIT HYPOTHESES.....	2
DIFFERENTIAL ANXIETY.....	3
ANXIETY IN ADMISSIONS TESTING.....	3
SEM STUDIES OF TEST ANXIETY	4
PRESENT STUDY	6
METHOD	7
DATA COLLECTION AND SAMPLE.....	7
MEASURES.....	7
THE ACT.....	7
TEST-PREPARATION QUESTIONNAIRE.....	7
DATA FILTERING	8
ANALYSIS.....	9
RESULTS	10
SAMPLE DEMOGRAPHICS.....	10
DESCRIPTIVE ANALYSES	11
ACT SCORES.....	12
QUESTIONNAIRE ITEMS.....	12
QUESTIONNAIRE ITEMS FOR DEMOGRAPHIC GROUPS.....	16
STRUCTURAL EQUATION MODELING	16
MODEL DEVELOPMENT.....	16
MODEL-DATA FIT.....	16
PARAMETER ESTIMATES.....	19
SENSITIVITY OF RESULTS TO SAMPLE.....	19
DIFFERENTIAL ANXIETY.....	20
DISCUSSION	20
LIMITATIONS.....	21
PRACTICAL IMPLICATIONS	22
CONCLUSIONS.....	24

Table of Figures

FIGURE 1. PROPOSED STRUCTURAL EQUATION MODEL5

Table of Tables

TABLE 1. PRO SAMPLE DEMOGRAPHICS 11

TABLE 2. ACT SCORE DESCRIPTIVE STATISTICS 12

TABLE 3. DESCRIPTIVE STATISTICS FOR YES/NO QUESTIONNAIRE ITEMS AND ASSOCIATIONS WITH OUTCOME 13

TABLE 4. CORRELATION MATRIX FOR QUESTIONNAIRE ITEMS..... 15

TABLE 5. MODEL FIT 17

TABLE 6. STANDARDIZED SEM PARAMETER ESTIMATES..... 19

Abstract

Debilitating test anxiety is a general threat to validity if it biases assessment scores. Moreover, if bias differs between demographic groups, anxiety also raises concerns about test fairness. This study applied structural equation modeling to investigate possible measurement bias due to anxiety on the ACT[®] assessment and relationships among contextual factors, anxiety, and estimated proficiency. Results indicated that ACT scores were not biased by anxiety—in general and for female, male, minority, and White examinees. Anxiety was primarily related to examinees' feelings of being unprepared for the test. Health factors such as hours of sleep and eating breakfast were related to reduced anxiety and greater test performance. Results support practical guidance for students, parents, and counselors who want to minimize the emotional and physiological symptoms of test anxiety, even if it is not likely to affect test performance.

Keeping Your Cool: Does Test Anxiety Bias Performance on the ACT?

Jeffrey T. Steedle, PhD

Introduction

Examinees are said to exhibit test anxiety if they experience adverse cognitive and physiological responses to testing situations they view as threatening (Spielberger & Vagg, 1995). A large body of prior research indicates that test anxiety is negatively associated with assessment performance (Seipp, 1991; Hembree, 1988). Given this association, students who experience anxiety while being evaluated may not demonstrate the full extent of their academic achievement or cognitive ability. Thus, test anxiety is a potential source of construct-irrelevant variance threatening the validity of test score interpretations for intended uses. Moreover, test anxiety may affect certain groups more than others, which makes it a threat to assessment fairness. For example, female students (Chapell et al., 2005; Zeidner, 1990), students of lower socioeconomic status (Putwain, 2008), minority students (Hembree, 1988), and students who speak English as a second language (Hodge, McCormick, & Elliot, 1997) have reported greater test anxiety.

In the interests of validity and fairness, this study sought to evaluate whether test anxiety systematically biased scores on a college admissions test and to broaden understanding of contextual factors associated with test anxiety and performance. For this study, a questionnaire about test preparation and anxiety was administered to a sample of examinees after they took the ACT® assessment (ACT, 2017). Questionnaire data were analyzed to examine how factors related to testing

environment, test preparation, and physical health were associated with test anxiety and performance. Results indicate whether anxiety biased admissions test results in general and whether the effect varied by gender or ethnicity. Of practical value, results also support empirically-based guidance for minimizing test anxiety, which should be of interest to examinees, parents, teachers, counselors, and test publishers.

Perspectives

Several large meta-analyses have established the negative relationship between test anxiety and academic performance (Hembree, 1988; Seipp, 1991; Richardson, Abraham, & Bond, 2012). This section provides a brief introduction to test anxiety, reviews test anxiety research related to differential anxiety and admissions tests, and examines prior research that employed methodology similar to the current study.

Test Anxiety

Dusek (1980) defined test anxiety as “an unpleasant feeling or emotional state that has psychological and behavioral concomitants, and that is experienced in formal testing or other evaluative situations” (p. 88). Sarason and Stoops (1978) described sufferers of test anxiety succinctly as “persons for whom tests are noxious experiences” (p. 107). Test anxiety arises when examinees view a testing situation as threatening (Spielberger & Vagg, 1995), but not all anxiety related to testing is harmful. Indeed, some examinees describe anxiety as helping them focus on the assessment task

(Chamberlain, Daly, & Spalding, 2011). Mentions of test anxiety typically refer to debilitating test anxiety rather than facilitating test anxiety, where debilitating and facilitating test anxieties prompt task-irrelevant and task-directed behaviors, respectively (Alpert & Haber, 1960).

The focus of this study is debilitating test anxiety, which Liebert and Morris (1967) hypothesized as having two components: worry and emotionality. Worry entails examinees' concerns about performance, and emotionality describes physiological responses to the threatening situation (e.g., perspiration and elevated heart rate). In a similar way, more recent research describes the cognitive and affective components of test anxiety. The only notable difference is that the cognitive components extend beyond worry to include other mental activities such as task-irrelevant thinking (Hodapp & Benson, 1997). One meta-analysis estimated the correlation between test anxiety and academic performance to be $-.21$ (Seipp, 1991), though the cognitive aspects of test anxiety generally correlate more strongly with test scores than the affective components (Hembree, 1988). Researchers commonly distinguish between *state* anxiety caused by an immediately threatening testing situation and *trait* anxiety, which is a general propensity to experience anxiety in a variety of evaluation contexts (Spielberger & Vagg, 1995).

The interference and deficit

hypotheses. There are two prevailing explanations for the association between test anxiety and performance. The interference hypothesis posits that anxiety interferes with task-relevant behaviors (e.g., knowledge retrieval) because attention is divided (Wine, 1971). For example, task-irrelevant thoughts—about poor performance, about others' performance, about the examiner—were more common for test-anxious adults (Sarason & Stoops, 1978). College students reporting higher

test anxiety were more susceptible to distraction during testing (Keogh, Bond, French, Richards, & Davis, 2004), and the attention of test-anxious children appeared to be divided between task-relevant and task-irrelevant stimuli (Dusek, Mergler, & Kermis, 1976). Research findings are consistent with the interference hypothesis when estimates of academic achievement appear to be biased by test anxiety and unbiased in the absence of test anxiety. Along those lines, a common form of support for the interference hypothesis comes from research showing the effectiveness of behavioral and cognitive interventions that reduce test anxiety, thereby improving performance (Hembree, 1988). Such findings are also taken as evidence that anxiety causes low performance, rather than some other causal relationship (Hill & Wigfield, 1984).

In contrast, the deficit hypothesis contends that there is no causal relationship between test anxiety and performance. Rather, their association can be explained by a third variable (Tobias, 1979). Specifically, some examinees have deficits in their knowledge and skills, which cause poor test performance and arouse anxiety because testing makes examinees acutely aware of their deficits. If true, test anxiety should be unrelated to performance after controlling for ability (i.e., anxiety does not bias performance). Consistent with the deficit hypothesis, the most common reason college students reported for experiencing test anxiety was lack of preparation (Bonaccio & Reeve, 2010). Indeed, test-anxious students appear to encode and organize knowledge in less effective ways, which may lead to deficits (Cassady, 2004; Birenbaum & Pinku, 1997). Some studies support the deficit hypothesis by showing that associations between anxiety and performance are stronger when anxiety is measured after the assessment (Zeidner, 1991; Sommer & Arendasy, 2014; Seipp, 1991). That is, perhaps anxiety builds up for less able students

as they take the test or they use anxiety to rationalize poor performance. Other studies, described in greater detail below, indicate that anxiety does not relate significantly to assessment performance while also accounting for ability in a structural equation model (Reeve & Bonaccio, 2008; Sommer & Arendasy, 2016).

Differential anxiety. If different examinee groups experience anxiety to different degrees and if test anxiety interferes with examinees' abilities to demonstrate their knowledge and skills, then test anxiety is a threat to assessment fairness. That is, test anxiety potentially compromises the comparability of test score interpretations for different examinees and examinee groups. However, if differences in test anxiety do not manifest in differences in performance when controlling for examinee ability, as predicted by the deficit hypothesis, test anxiety may not present a challenge to assessment fairness.

In a large number of studies, females reported higher average anxiety than males (Hembree, 1988), and studies have detected a stronger relationship between anxiety and grade point average for females (Chapell et al., 2005). Yet, meta-analytic evidence indicates no differential relationship between anxiety and test performance for males and females (Seipp, 1991). Other studies have detected weak negative associations between anxiety and socioeconomic status (Putwain, 2007; 2008) and greater average test anxiety among Black and Hispanic students compared to White students (Hembree, 1988). Likewise, another study revealed that the students least likely to report anxiety on a college admissions test were White, male, and native English speakers (Hodge, McCormick, & Elliot, 1997). However, other research has detected no differences in test anxiety by socioeconomic status or ethnicity (e.g., Zeidner, 1990). Research is lacking on the strength of the relationship

between anxiety and test performance for different ethnic and socioeconomic groups.

Anxiety in admissions testing. Test anxiety tends to be greater when tests are used for evaluation—as in college admissions testing—rather than formative or developmental purposes (Reeve, Bonaccio, & Charles, 2008; Hembree, 1988). Prior research suggests that test anxiety may arise from fear of failure, fear of being looked down upon, feelings of insufficiency (Kilmen, 2015), recognition of the importance of performing well (Selkirk, Bouchey, & Eccles, 2011), and perceptions of test difficulty (Hong, 1999). These factors are all potentially relevant to college admissions tests since performance is connected to college acceptance, personal prestige, and financial support to attend college.

The phenomenon of high stakes inducing test anxiety is explained by expectancy-value theory, which posits that test anxiety is determined by an examinee's expectation of performing well and perceived value of an assessment (Wigfield & Eccles, 2000). In one study, for example, sixth- and seventh-grade students who expected to perform poorly in math and English but highly valued success reported the greatest anxiety (Selkirk, Bouchey, & Eccles, 2011). Likewise, the two most common reasons for test anxiety among college students were lack of preparation and the consequences of test performance on course grades and grade-point average (Bonaccio & Reeve, 2010).

In the context of college admissions testing, it is expected that most examinees place high value on the assessment because performance has consequences for college acceptance and financial aid. When examinees perceive a test to be difficult, anxiety tends to be higher (Hong, 1999; Hembree, 1988) and expectations of

performing well are likely to be lower. Moreover, expectations likely vary according to examinees' performance goals, ability levels, and psycho-social attributes such as academic buoyancy (Putwain, Daly, Chamberlain, & Sadreddini, 2015) and academic self-efficacy (Nie, Lau, & Liau, 2011), both of which correlate negatively with test anxiety.

With high value placed on the assessment, potentially low expectations of performing well, and test anxiety being relatively common among high school students (Hembree, 1988), college admissions testing should be a productive context for studying test anxiety. One study, for example, detected correlations around $-.35$ between test anxiety and SAT scores, and the anxiety-performance relationship was stronger for the SAT than for a low-stakes, non-evaluative assessment (Cassady, 2004). Correlations similar in magnitude were observed between cognitive test anxiety and ACT performance (von der Embse & Witmer, 2014). In that study, three types of anxiety collectively accounted for 2% to 7% of the variance in ACT scores beyond that accounted for by demographics and high school grade-point average.

Like the SAT and ACT studies, correlations in the range of $-.35$ to $-.40$ were observed in a study of test anxiety and performance on the GRE—the admissions test for many graduate school programs (Powers, 1986). In a randomized trial, the correlation between trait test anxiety and performance on a college admissions test in Israel was approximately $-.20$ when anxiety was measured before the test, but it was closer to $-.40$ when measured after the test (Zeidner, 1991). That result was taken as evidence supporting the deficit hypothesis since the experience of taking the assessment influenced self-reported trait anxiety, which should be stable over time. A study of performance on a medical school admissions

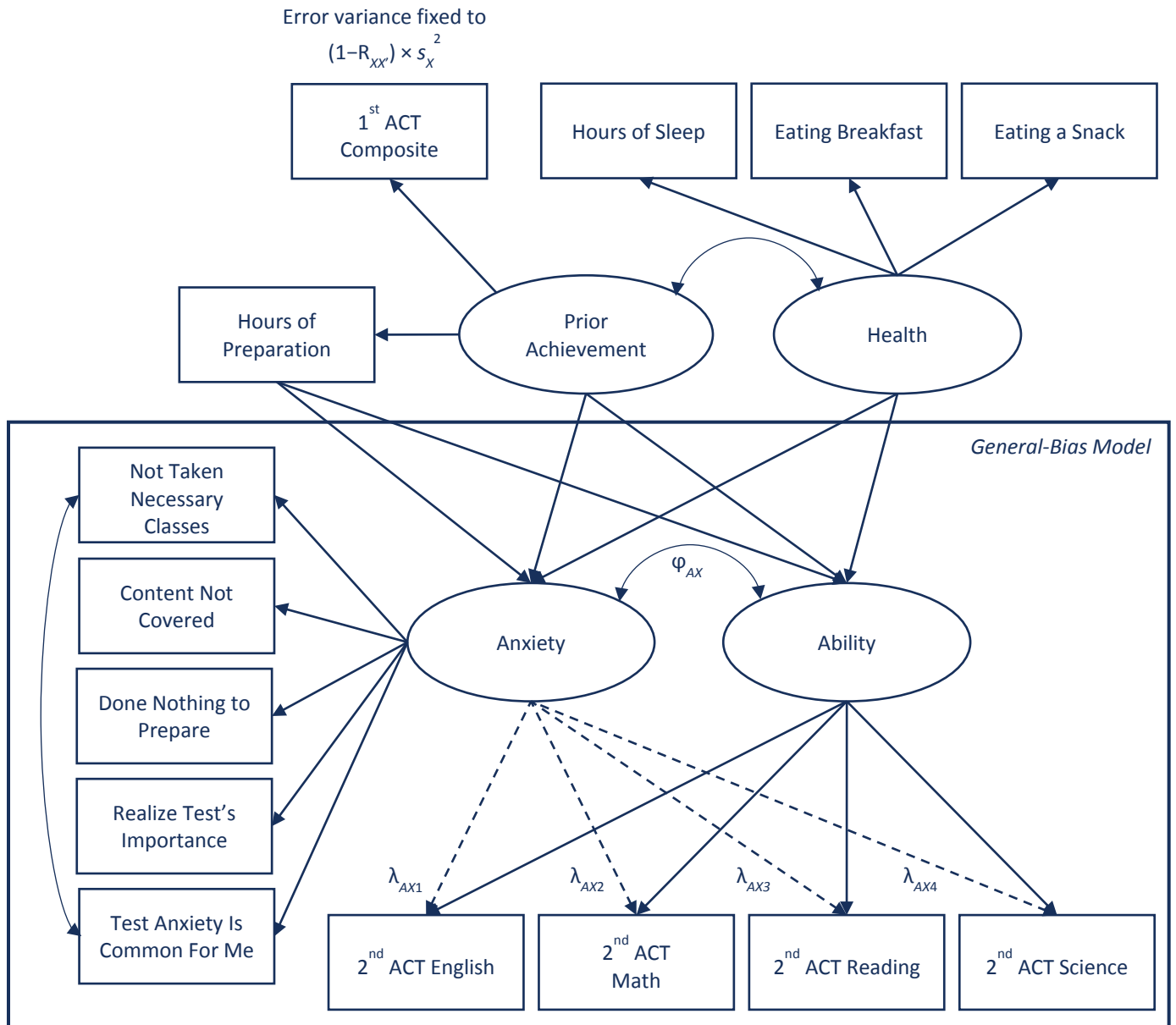
tests reached a similar conclusion because estimates of ability were not biased by self-reported trait anxiety measured after the assessment (Sommer & Arendasy, 2016).

SEM Studies of Test Anxiety

Like some prior test anxiety research, this study employed structural equation modeling (SEM) to examine whether anxiety biased observed assessment performance while controlling for examinee ability. In most of those studies, anxiety was treated as a latent variable manifested by responses to a test anxiety inventory, and ability was a latent variable manifested by observed test performance. Following Halpin, da-Silva, and De Boeck (2014), the latent correlation between anxiety and ability is referred to as the *deficit effect* because its presence supports the deficit hypothesis by explaining observed associations between test performance and self-reported anxiety. *Interference effects* are cross-loadings from latent anxiety to observed performance. Non-zero interference effects are consistent with measurement bias due to test anxiety.

The lower portion of Figure 1 represents the *general-bias* model, with the deficit effect labeled ϕ_{AX} and the interference effects labeled λ_{AX1} through λ_{AX4} . In SEM analyses of test anxiety, the fit of the general-bias model is compared to the *no-bias* model (with the interference effects constrained to zero) to test whether construct-irrelevant test anxiety biases observed test performance. The interference hypothesis is supported if the general-bias model fits the data better than the no-bias model and the interference effects are significantly less than zero. The deficit hypothesis is supported if the no-bias model fits the data as well as the general-bias model and the interference effects are non-significant. Such analyses have been called studies of measurement invariance because they test

Figure 1. Proposed structural equation model.



whether “groups differ in the way the measurement of a psychological construct...is related to that construct” (Wicherts, Dolan, & Hessen, 2005, p. 698). In this context, the “groups” are students of differing test anxiety, the “measurement” is observed test performance, and the “construct” is academic achievement. Halpin, da-Silva, and De Boeck (2014) presented this general approach to studying test anxiety and proved that model identification (and therefore unambiguous

interpretation of parameter estimates) required fitting a *restricted* general-bias model with some interference effects constrained to be equal. However, that approach only works if the loadings of observed performance on latent ability are significantly different from one another (i.e., they are not tau-equivalent). That hypothesis can be tested by fitting a restricted no-bias model with all such loadings set equal. The condition is met if the no-bias model fits the data significantly better than the restricted

no-bias model, and it is therefore safe to interpret the restricted general-bias model parameter estimates as a unique solution. At this point, model-data fit of the restricted general-bias and no-bias models may be compared using methods for nested models.

Halpin and his colleagues (2014) also provided an example to illustrate their proposed approach. They analyzed data from linguistics and natural science assessments administered to 382 high school students. State task-irrelevant thinking was measured using questions about distracting thoughts experienced by examinees. There were no stakes attached to performance, and most examinees reported low anxiety, yet the analyses detected significant interference effects for some items (as large as $-.36$) and no significant latent correlation between anxiety and ability (i.e., no deficit effect). Thus, results from this example analysis appeared to support the interference hypothesis.

Reeve and Bonaccio (2008) investigated possible bias due to test anxiety using a multiple-indicator multiple-cause model (Muthén, 1989) in which anxiety was treated as an observed (not latent) variable. Though examinees took five cognitive ability tests in a low-stakes setting, many of them still reported moderate to high anxiety on a 10-item scale measuring state worry. SEM results were consistent with the deficit hypothesis because anxiety did not apparently bias test performance. Specifically, none of the interference effects were statistically significant, anxiety correlated significantly with general intelligence ($r = -.45$), and model-data fit was as good or better for the model with no interference effects.

Sommer and Arendasy (2014; 2016) applied the Halpin, da-Silva, and De Boeck (2014) analysis framework in a series of two studies. In the first,

411 psychology students reported trait anxiety, and then completed four cognitive ability tests. State anxiety was recorded near the beginning of the tests and after completing the tests. Across analyses, the general-bias models fit no better than the no-bias models, trait anxiety was significantly related to cognitive ability and general intelligence (g) latent traits, and interference effects were non-significant. Additional analyses revealed that pre-test state anxiety was not significantly related to ability, but post-test state anxiety was significantly correlated with g , verbal fluency, and algebra word problem solving. The second study further corroborated the deficit hypothesis. In that study, 1,768 medical school applicants took high-stakes admissions tests measuring biology, chemistry, physics, and mathematics achievement and then reported trait worry, task-irrelevant thinking, and emotionality. For all tests and types of anxiety, the no-bias model fit as well as the general-bias model, anxiety significantly correlated with ability, and the interference effects were negligible in magnitude and non-significant.

Present Study

The current study supplements the research literature by investigating measurement bias due to test anxiety in a common, authentic, and important assessment context: college admissions testing. For this study, college-bound high school students took the ACT, and they self-reported test anxiety afterward by answering questions related to test preparation, the test's importance, and propensity to experience test anxiety.

This study applied the analysis framework of Halpin, da-Silva, and De Boeck (2014), but it also extended their model with additional contextual variables possibly related to test anxiety and estimated proficiency. Those variables included hours of test preparation,

prior achievement, testing environment, and health factors like hours of sleep and eating breakfast. In prior research, test anxiety tended to be lower for high-achieving students, and it was unrelated to testing in a familiar or unfamiliar location (Hembree, 1988). Thus, prior achievement was expected to correlate negatively with anxiety, and testing environment variables were not expected to correlate with anxiety. As for the other factors, it seemed plausible that test anxiety would be lower for students who prepare for the test and make healthy choices. However, those predictions were speculative due to lack of prior research. This study attempted to fill that knowledge gap and provide empirically supported recommendations for test takers. Another unique aspect of this study was the application of SEM to address questions of test fairness by comparing models fit with data from different demographic groups. In all, this study addressed the following research questions:

1. Does test anxiety bias ACT performance for college-bound examinees?
2. How do contextual factors relate to anxiety and estimated ability on the ACT?
3. Do relationships between anxiety and ACT performance differ between demographic groups?

Empirical evidence addressing these questions will help improve understanding of contextual factors associated with test anxiety and the potential consequences of anxiety in college admissions testing for different student groups. Moreover, results could have practical implications for examinees, parents, and counselors in terms of minimizing test anxiety.

Method

Data Collection and Sample

The questionnaire analyzed in this study was sent to 76,000 students who took the ACT twice. Initially, the questionnaire was sent to a random sample of 38,000 students who took the ACT for the first time in spring 2012 and the second time in October 2012. In the second wave of data collection, the questionnaire was sent to a random sample of 38,000 students who tested in fall 2012 and in April 2013. All students in the samples indicated that they participated in some form of test preparation for their second tests. Each sampled student received a custom URL via email to the online questionnaire 2–3 days after taking the ACT for the second time. Examinees would not have received their ACT scores at the time of responding to the questionnaire. The initial data set included 8,472 respondents, which reflected a response rate of 9.5%.

Measures

The ACT. The ACT test measures high school academic achievement in four subject areas (English, math, reading, and science), and it is used primarily for college admissions (ACT, 2017). All ACT scores are reported on a 1–36 score scale. Some analyses in this study used the ACT Composite score, which is the average of the four subject-area tests. In this study, students' first ACT Composite score served as a measure of prior achievement, and second ACT scores served as outcome variables. Gender, ethnicity, and self-reported high school grade-point average were obtained from ACT registration data.

Test-preparation questionnaire.

The test-preparation questionnaire asked about the number of hours spent on 11 types of test preparation (e.g., practice tests, online-test

preparation, workbooks, software, test-preparation courses, and tutors). Respondents selected the number of hours spent on each type of test preparation: 0, 1–5, 6–10, 11–55, 16–20, or more than 20 hours. To estimate total hours of test preparation, the ordinal responses were transformed to 0, 3, 8, 13, 18, and 23 hours, respectively, and summed.

Next, the questionnaire focused on testing context and anxiety. The first of those items asked about hours of sleep the night before taking the test. In analyses, ordinal hours of sleep responses were transformed to numeric values (1–3 = 2 hours, 4–6 = 5 hours, 7–9 = 8 hours, 10 or more = 11 hours). Next, the questionnaire presented a series of yes/no items asking whether the student ate breakfast, ate a snack during the break time, knew anyone in the test room, and took the test in a familiar place. Respondents who tested in an unfamiliar location were also asked whether they had trouble finding the test site and whether that experience negatively affected test performance.

All respondents were asked, “While taking the test, were you so stressed or anxious that you believe it negatively affected your performance on the test?” Respondents who answered ‘yes’ were presented a series of checkboxes listing six possible reasons for their anxiety: (a) I had not yet taken the class(es) necessary for doing well on one or more areas of the test, (b) Some areas of the test had not been covered at all or had not been covered adequately in my high school classes, (c) I realized I had not done anything to prepare myself for taking this type of test, (d) I suddenly realized how important the test results were to my future, (e) It is not uncommon for me to experience test anxiety, and (f) Other. The most common “other” reason was needing more time to complete the assessment. Other examples included illness or injury, extracurricular activities or jobs, and

restatements of preceding items (e.g., the importance of the test and not getting enough sleep).

Before proceeding, it is important to acknowledge several limitations of the questionnaire. First, the questionnaire was not developed for the specific purposes of this study, so limitations imposed by the questionnaire were unavoidable. For example, additional contextual factors might have been studied (e.g., motivation and peer pressure), but the questionnaire did not include any relevant items. The major limitation of the anxiety items is that they were not written with a particular theory or type of test anxiety in mind. Most of the anxiety items dealt with preparation for the ACT or thoughts about the current assessment situation, so they were considered indicators of state test anxiety. The only notable exception was the item about commonly experiencing test anxiety, which better aligned with trait anxiety. Finally, measurement timing is known to affect self-reported test anxiety (e.g., Sommer & Arendasy, 2016), so results might have differed if the questionnaire was administered before the test rather than 2–3 days afterward. However, delivering the questionnaire to a select sample before the test would have threatened the standardization of the assessment, and there was no mechanism for administering the questionnaire on testing day. Despite these limitations, this study still offered a useful replication of prior research in a new assessment context, thereby providing theoretical and practical insights about debilitating test anxiety.

Data Filtering

The initial data set included 8,472 students who opted into the questionnaire and actually responded to some items. Effort was made to omit data for students that apparently exhibited insufficient effort responding (Huang, Curran, Keeney, Poposki, & DeShon, 2012). That

is, some students generated illogical or improbable patterns of responses that may have reflected careless, inattentive, or random responding. Such behavior is generally thought to introduce measurement error that deflates criterion-related validity coefficients, reduces reliability, and distorts factor structure (Huang, Bowling, Liu, & Li, 2015). Error variance caused by insufficient effort responding can also reduce power in statistical analyses like multiple regression (Maniaci & Rogge, 2014).

For example, students were removed if they selected the same response (other than 0 hours) to most of the test preparation items. That is, it would be exceptionally unlikely for a student to participate in 10 or 11 types of preparation, let alone spend the same amount of time on each activity. Students were also removed if they provided an illogical pattern of responses to items about the testing location (e.g., saying they took the test at their own high schools and in an unfamiliar location) or if they endorsed all six reasons for experiencing test anxiety.

The cutoffs for these filtering procedures were intentionally conservative (i.e., additional suspicious data might have been deleted) to avoid removing students with atypical but valid response patterns. The reduction of 456 students (5.4%) was consistent with prior research in terms of the prevalence of insufficient effort responding (Meade & Craig, 2012). Though the consequences of insufficient effort responding are typically small (Johnson, 2005; Huang et al., 2012), such filtering is a reasonable precaution to protect against the influence of unsound data.

The final step in data preparation was to remove students who responded 'yes' to the item asking whether stress or anxiety negatively affected performance the first time they took the ACT (or did not respond to the item), which left 3,732 students in the data set.

The reason for this filtering was to make first ACT scores more dependable measures of prior achievement by eliminating the possible effect of debilitating anxiety. That is, if first ACT scores reflected debilitating anxiety, this could have influenced the estimated relationships among first ACT scores, second ACT scores, and anxiety experienced while taking the ACT for the second time. Results of a sensitivity analysis are provided to determine whether this filtering procedure affected results or conclusions.

Analysis

ACT scores and questionnaire data were analyzed with SEM implemented via the *lavaan* package in R (Rosseel, 2012). On account of the data including categorical variables, parameters were estimated using diagonally weighted least squares. To include students with incomplete data, missing questionnaire responses were filled in via multiple imputations with 20 iterations. This process represented uncertainty in the imputation process by replacing missing data with a distribution of possible, acceptable values (Rubin, 1986). Rubin's (1987) rules were used to combine coefficient estimates and standard errors across multiple imputations.

In all, four structural equation models were fit to the data. First, the restricted general-bias and no-bias models described by Halpin and his colleagues (2014) were fit. Model identification for the restricted general-bias model was achieved by restricting the interference effects for ACT English and reading scores to be equal (i.e., $\lambda_{AX1} = \lambda_{AX3}$ in Figure 1). This restriction was chosen because English and reading were considered the most related subject areas. Note that prior studies analyzed item scores as observed performance, which was straightforward since all examinees took the same items. For the ACT, however, numerous forms were administered

simultaneously, so equated scale scores were analyzed as observed performance.

Next, the restricted general-bias and no-bias models were fit with additional variables regressed on anxiety and ability (see entirety of Figure 1). These models included hours spent on test preparation, first ACT Composite score as a measure of prior achievement, and a latent variable for health. In Figure 1, prior achievement is shown as a latent trait with a single indicator (first ACT Composite), which was possible because the error variance of ACT Composite scores was known. When fitting the model, the error variance was fixed to $1 - R_{xx}$ times the observed variance, where R_{xx} is the reliability of ACT Composite scores (.97; ACT, 2017). This approach avoids the treatment of first ACT Composite scores as perfectly reliable (Kline, 2010).

To account for temporal precedence, the model included directional relationships from preparation, prior achievement, and health toward anxiety and ability. Eating a snack was the only exception, since it would have occurred during testing, but it was kept with the other health factors for simplicity. Hours of preparation was also regressed on first ACT Composite score since earlier performance on the ACT was likely related to an examinee's need and motivation to prepare for taking the test again. Note that several questionnaire items do not appear in Figure 1. As described in the results section, those items were not analyzed, either because they had little variance or because they correlated weakly with other predictors or outcomes. The model included one residual covariance between anxiety items to better account for their observed correlation.

Model-data fit was evaluated using χ^2 and $\chi^2/\text{degrees of freedom}$, which addresses the sensitivity of χ^2 to sample size. Root mean square error of approximation (RMSEA; Steiger

& Lind, 1980) less than 0.05 was treated as an indicator of good model-data fit and 0.08 as the upper bound for acceptable fit (Browne & Cudeck, 1992). In addition, comparative fit index (CFI; Bentler, 1990) values greater than or equal to .95 were considered signs of good fit (Hu & Bentler, 1999). The standardized root mean square residual (SRMR), which reflects the mean absolute difference between observed and predicted correlations among variables, should be less than or equal to .08 (Hu & Bentler, 1999). Fit of the no-bias model was compared to fit of the restricted general-bias model to test whether anxiety biased assessment performance. Specifically, the nested models were compared using significance tests for $\Delta\chi^2$ (Satorra & Bentler, 2001) as well as $\Delta\text{CFI} > .01$ (Cheung & Rensvold, 2002).

To evaluate possible differential measurement bias due to anxiety, the models were fit separately for different demographic groups (female, male, minority, and White). First, the models were fit with data from a single group to investigate possible measurement bias (as in the analyses using all data). Then, models were fit simultaneously for pairs of demographic groups (female and male, minority and White). All parameter estimates were first estimated freely. Then, the models were re-fit with the deficit effect and interference effects constrained to be equal across groups. If the model with freely estimated parameters fit the data significantly better, this was considered evidence of a differential bias effect.

Results

Sample demographics

Descriptive statistics for demographic variables and academic achievement were calculated to examine how the sample in this study compared to a typical group of ACT-tested high school graduates. There were several notable

demographic differences between the study sample and the national ACT-tested high school graduating class (Table 1). Specifically, the sample had a greater percentage of female students, a smaller percentage of African American students, and greater percentages of White and Asian students. The study sample also had higher average high school grade-point averages and higher average ACT Composite scores. These differences likely

reflected selection bias on account of the study sample comprising students who took the ACT twice. Such students tend to be higher achieving and more likely to seek admissions to more selective colleges and universities. There was also possible selection bias due to elective participation in the study. That is, higher achieving students may be more likely to opt into a voluntary questionnaire.

Table 1. Sample Demographics for the ACT-tested High School Graduating Class of 2015 (Sample N = 3,732; Population N = 1,395,418)

		Sample	Population ¹
Gender	Male	39.9%	47.3%
	Female	60.1%	51.6%
Race/Ethnicity	African American	7.6%	14.9%
	White	66.6%	57.5%
	Latino	12.0%	12.6%
	Asian	6.1%	3.1%
High School GPA	0.0–1.9	0.4%	3.8%
	2.0–2.4	2.0%	13.8%
	2.5–2.9	5.9%	24.5%
	3.0–3.4	18.8%	32.5%
	3.5–4.0 or higher	73.0%	25.4%
Mean ACT Composite	First Score	24.4	21.1 ²
	Second Score	25.3	

² The mean ACT Composite score for the population reflects each student's most recent test attempt.

Descriptive Analyses

ACT scores. Mean differences between first and second ACT subject-test scores ranged from 0.6 for math to 1.6 for English on the 1–36 scale (Table 2). All mean differences were statistically significant at the $p < .001$ level according to paired-sample t -tests. Thus, from first to second ACT testing, examinees tended to increase their scores by a small amount.

Corresponding effect sizes ranged from 0.11 to 0.26 standard deviations (using the method for matched groups; Dunlap, Cortina, Vaslow, & Burke, 1996). First and second ACT Composite scores correlated .93 ($p < .001$), so it was known that first ACT scores accounted for much of the variance in second ACT scores. This left little additional unique variance to be accounted for by anxiety, health, and test preparation.

Questionnaire items. Several statistics were estimated to compare ACT Composite scores for those who did and did not affirm the dichotomous questionnaire items (Table 3). First, a paired-sample t -test indicated whether mean differences should be considered significantly different from zero. All but one of the significant t statistics were in the expected direction. That is, students who ate breakfast or ate a snack tended to score higher, and students who affirmed anxiety items tended to score lower. Testing in a familiar location was the exception. One might expect testing in a familiar location to reduce anxiety, thereby improving performance, but results indicated that students testing in an unfamiliar location performed better on average. This result may reflect a selection effect, wherein higher achieving students (e.g., from distant high schools or from smaller private high schools) were more likely to test in an unfamiliar location.

Table 2. ACT Score Descriptive Statistics

ACT Score	Items	1st ACT		2nd ACT		Mean Diff.	Effect Size	Correlation
		Mean	SD	Mean	SD			
English	75	24.0	5.9	25.5	6.1	1.6	0.26	.89
Math	60	24.1	5.1	24.7	5.3	0.6	0.11	.89
Reading	40	25.0	5.9	25.8	6.0	0.8	0.14	.83
Science	40	24.0	4.7	24.8	5.2	0.8	0.16	.78
Composite		24.4	4.8	25.3	5.1	0.9	0.19	.93

All mean differences and correlations were statistically significant at the $p < .001$ level.

Table 3. Descriptive Statistics for Yes/No Questionnaire Items and Associations with Outcome

Category	Item	N	Percentage		Mean 2nd ACT Composite				ANCOVA Main Effect	
			Yes	No	Yes	No	Diff.	t_{paired}	d^f	
Anxiety	Have not taken necessary classes	3,722	4%	96%	21.7	25.5	-3.8	-9.6***	-0.84	-0.61***
	Test content not covered by my classes	3,722	7%	93%	22.2	25.6	-3.4	-11.1***	-0.75	-0.49***
	Done nothing to prepare	3,722	5%	95%	22.7	25.5	-2.8	-7.9***	-0.63	-0.89***
	Suddenly realize test's importance	3,722	10%	90%	23.8	25.5	-1.7	-6.2***	-0.38	-0.37***
	Test anxiety is common for me	3,722	7%	93%	23.5	25.5	-2.0	-6.6***	-0.45	-0.42***
	Other reason for anxiety	3,722	5%	95%	25.7	25.3	0.4	1.2	0.09	-0.61***
Environ	Difficulty finding test site	3,370	1%	99%	22.9	25.4	-2.5	-2.3 [†]	-0.55	-0.35
	Knowing anyone in the test room	3,585	79%	21%	25.4	25.1	0.3	1.3	0.06	0.03
	Testing in a familiar location	3,347	71%	29%	25.1	26.1	-1.0	-5.4***	-0.23	0.08
Health	Eating breakfast	3,626	85%	15%	25.7	23.2	2.6	10.7***	0.56	0.48***
	Eating a snack	3,600	51%	49%	26.0	24.7	1.2	7.4***	0.28	0.18**

[†] $p < .05$, ** $p < .01$, *** $p < .001$

¹ To facilitate comparisons of effect sizes, the standard deviation 4.53 was used as the denominator when calculating all effect sizes. 4.53 was the average ACT Composite score standard deviation between 1991 and 2003 (ACT, 2017).

Next, the mean differences were expressed as effect sizes in standard deviation units. The largest effect sizes (-0.84, -0.75, and -0.63) were observed for the anxiety items relating to preparation: not taking necessary classes, not having content covered in classes, and doing nothing to prepare. Smaller effect sizes of -0.38 and -0.45 were estimated for suddenly realizing the test's importance and commonly experiencing test anxiety, respectively. Difficulty finding the testing site was associated with a 0.55 standard deviation decrease in ACT Composite scores, but only 1% of students reported this problem. Knowing anyone in the testing room was not significantly related to ACT performance, but testing in a familiar location was associated with a .23 standard deviation decrease in performance. Students

who ate breakfast and a snack scored higher than those who did not by 0.56 and 0.28 standard deviations, respectively.

Observed relationships between questionnaire items and ACT scores could reflect their mutual associations with another variable. For example, higher ability students may be more likely to eat breakfast, which gives rise to the observed relationship between eating breakfast and performing well. To address that issue, analysis of covariance (ANCOVA) was used to estimate the main effect of affirming an item while controlling for prior achievement (first ACT Composite score). Interactions between first ACT score and affirming an item were all non-significant, so they were omitted from the ANCOVA models. As shown in Table 3, the

ANCOVA main effects were notably smaller in magnitude than the observed mean differences. This finding indicates that prior achievement accounted for much of the mean differences, which is consistent with the notion that reported anxiety is associated with ability. After controlling for prior achievement, the testing environment variables were no longer significantly related to test performance.

Approximately 75% of students reported getting 7–9 hours of sleep, and 20% reported getting 4–6 hours. There was a weak positive association between hours of sleep and second ACT score (polyserial correlation of .11, $p < .001$). Controlling for prior ACT Composite scores did not affect the strength of this relationship, so hours of sleep may account for a small amount of unique variance in second ACT performance. For the 95% of examinees who reported participating in test preparation, the mean was 20 hours, the standard deviation was 21 hours, and the median was 13 hours. Contrary to intuitive expectations, students who spent more time preparing tended to perform worse on their second ACT attempt ($r = -.12$, $p < .001$). However, when second ACT scores were

regressed on first ACT scores and hours of preparation, both coefficients were positive and statistically significant ($p < .001$). This sign reversal came about because students who performed worse on the first test tended to spend more time preparing for the second test ($r = -.18$, $p < .001$). Regression results indicated that the ACT Composite score would be expected to increase by only 0.013 points for every hour of preparation.

Table 4 shows the correlations among the questionnaire items. The first five anxiety items correlated between .24 and .52, but those items correlated weakly with the “other reason for anxiety” item. Knowing other test takers correlated .24 with testing in a familiar location. Difficulty finding the testing site correlated $-.14$ with testing in a familiar location, but only $-.04$ with knowing other test takers. Students who ate breakfast were more likely to eat a snack, but the relationship was weak ($r = .14$). Students who slept more were slightly less likely to eat breakfast or a snack. Hours of preparation correlated weakly with doing nothing to prepare, not having difficulty finding the test site, and hours of sleep.

Questionnaire items for

demographic groups. Females consistently affirmed to the anxiety items more frequently than males. The largest differences were on the items about realizing the test's importance (12.6% vs. 7.1%) and commonly experiencing test anxiety (9.6% vs. 3.8%). Differences on items about the testing location were all less than two percentage points. Females and males were similarly likely to eat breakfast, but females were more likely to report eating a snack (55.6% vs. 43.7%). Compared to females, males reported an average of 16 minutes more sleep ($t = 3.08, p < .01$) and 1.2 additional hours preparing for the ACT (non-significant). The average ACT Composite score for males was 0.6 points higher than for females ($t = 3.58, p < .001$).

Due to relatively small sample sizes, African American and Hispanic or Latino students were grouped together in analyses. Compared to White students, minority students were consistently more likely to affirm the anxiety items. The largest differences were on the items about doing nothing to prepare (10.6% vs. 3.7%) and realizing the test's importance (15.3% vs. 8.2%). Minority students were about three percentage points less likely to know someone in the testing room and test in a familiar location. They were also less likely to eat breakfast (75.8% vs. 88.7%), but minority and White examinees were equally likely to eat a snack. Minority students got an average of 24 minutes less sleep ($t = 4.35, p < .001$) and reported an average of 0.7 more hours of preparation (non-significant). The average minority ACT Composite score was 4.2 points lower than White examinees ($t = 16.48, p < .001$).

Structural Equation Modeling

Model development. Early iterations of the SEM included latent variables not shown in

Figure 1. At one point, for example, the model included a latent variable for preparation, which was manifested by first ACT score, hours of preparation, and anxiety items about preparation. However, in the model, hours of preparation loaded negatively on the preparation latent variable. This counterintuitive result was likely caused by preparation being driven strongly by first ACT score, which correlated negatively with hours of preparation. For that reason, the model in Figure 1 treats first ACT score as a standalone predictor of anxiety and ability.

In other early models, the items about testing environment (difficulty finding the testing site, knowing anyone in the testing room, and testing in a familiar location) were treated as observed evidence of a latent variable for environment, but that consistently resulted in model convergence problems. The environment items did not work as indicators of anxiety either because they correlated so poorly with the anxiety items. They might have been treated as direct predictors of ACT performance, but descriptive analyses revealed little variance or no effect after controlling for prior achievement (Table 3). For those reasons, the environment items were dropped from the analyses. The only other variable eliminated was the "other reason for anxiety" item. With only 5% of respondents endorsing that item, its correlations with the other five anxiety items were relatively weak, and the model could not adequately account for those correlations (i.e., multiple correlation residuals exceeded .15 in magnitude).

Model-data fit. Table 5 lists model-data fit indices. The χ^2 statistic was used to test the hypothesis that a model perfectly reproduced the observed covariance structure. The exact-fit hypothesis was rejected for every model in this study, but χ^2 is known to be sensitive to large sample size (Kline, 2010). That is, small

Table 5. Model Fit

Model	χ^2	df	$p(\chi^2)$	χ^2/df	RMSEA	RMSEA (upper)	CFI	SRMR	$\Delta\chi^2$	Δdf	$p(\Delta\chi^2)$	ΔCFI
GB1	117.4	22	.000	0.187	0.034	0.040	.993	0.047	—	—	—	—
NB1	127.3	25	.000	0.196	0.033	0.039	.992	0.049	14.90	3	.002	-.001
GB2	449.2	66	.000	0.147	0.039	0.043	.981	0.060	—	—	—	—
NB2	461.7	69	.000	0.149	0.039	0.042	.980	0.061	17.76	3	.001	.000

Note: GB1 = restricted general bias model, NB1 = no-bias model, GB2 = expanded restricted general bias, NB2 = expanded no-bias model

differences between model predictions and observed data can lead to large χ^2 values when sample size is great. Large χ^2 values can also arise because of high correlations among variables (e.g., among ACT scores), which allow for greater model-data discrepancies.

Deviations from multivariate normality are another possible cause, but the Satorra-Bentler statistic calculated in this study controls for non-normality (Satorra & Bentler, 1994).

Differences between observed and predicted correlations were calculated, and none of the correlation residuals exceeded .10 in magnitude. Consistent with that finding, the SRMR was well under .08. By other measures, model-data fit was good for all models. RMSEA and its confidence interval upper bound were less than 0.05, which supported the close-fit hypothesis, and CFI exceeded .95.

To investigate whether the significant χ^2 statistics were caused by large sample size, power analyses (MacCallum, Browne, & Sugawara, 1996) were used to estimate the likelihood of rejecting the null hypothesis of $RMSEA \leq 0.05$ (the close fit hypothesis) with $\alpha = .05$ when the true RMSEA is 0.08 (borderline acceptable fit). With available sample sizes and degrees of freedom, power was 1.0, so the null hypothesis would always be rejected. To

achieve power of .80 for the models with anxiety and ability, sample sizes of 402 and 363 were needed for the restricted general-bias and no-bias models, respectively. The models were refit with random samples of those sizes, and the χ^2 statistics were no longer statistically significant ($p = .50$ and $p = .58$). Results were similar for the models including prior achievement, health, and preparation; the χ^2 p -values were .61 and .58 with sample sizes of 227 and 239. Thus, power analyses supported the notion that significant χ^2 statistics were artifacts of large sample size.

With all models apparently fitting the data well, model-data fit was compared between the restricted general-bias and no-bias models. First, considering the models that included only anxiety and ability (the lower portion of Figure 1), the difference in χ^2 statistics was statistically significant and χ^2/df was higher for the no-bias model, both of which suggested better fit for the restricted general-bias model. However, the difference in CFI was well below .01 and RMSEA was marginally lower for the no-bias model, which suggested similar fit for the two models. Like χ^2 goodness-of-fit statistics, $\Delta\chi^2$ is also sensitive to sample size (Cheung & Rensvold, 2002), so additional power analyses were conducted (MacCallum, Browne, & Cai, 2006).

With random samples of 541 students, which would be required to detect the difference between RMSEA of 0.03 and 0.04 with power of .80 and $\alpha = .05$, $\Delta\chi^2$ was not statistically significant ($p = .30$). Thus, results were interpreted as indicating similarly good fit for the restricted general-bias model and the no-bias model. The same basic pattern in results was observed for the models including health, prior achievement, and hours of preparation: $\Delta\chi^2$ was significant, χ^2/df was slightly higher for the no-bias model, RMSEA was lower for the no-bias model, and ΔCFI was negligible. Again, power analyses suggested that large sample size resulted in significant $\Delta\chi^2$ statistics ($p = .62$ with random samples of 315), so results were interpreted as indicating similarly good fit for the two models. With the no-bias model apparently fitting as well as the restricted general-bias model, model-data fit comparisons indicated that anxiety did not bias test performance.

Parameter estimates. If parameter estimates align with results from model-data fit comparisons, they should reveal non-significant interference effects in the restricted general-bias models. In Table 6, the interference effects are the loadings of ACT English, math, reading, and science on anxiety. In both of the restricted general-bias models, the interference effects all loaded negatively on test scores, but none were significantly different from zero. The loadings for math and science were nearly so in the model including only anxiety and ability ($p < .10$). The deficit effects in the no-bias models were $-.30$ and $-.31$ (both $p < .001$). Overall, results were consistent with the deficit hypothesis since the observed relationship between anxiety and performance was accounted for by the correlation between anxiety and ability. The four ACT subject-area tests had loadings of approximately .85 on the ability latent variable ($p < .001$). In comparison, the non-significant interference effects were

substantially lower in magnitude (approximately $-.25$ in the restricted general-bias model). Of the anxiety items, not taking necessary classes and not covering test content in class had the strongest loadings on the anxiety latent variable. These were followed closely by realizing the test's importance, commonly experiencing test anxiety, and doing nothing to prepare. With a standardized coefficient of .80, eating breakfast had the strongest loading on the health latent variable. Eating a snack and hours of sleep had similar loadings on health (.41 and .32). Health had a latent correlation of .26 with prior achievement ($p < .001$).

The second research question concerned the relationships among contextual factors, anxiety, and estimated ability. Regression results revealed that preparation, health, and prior achievement were positively related to ability, but prior achievement accounted for nearly all of the variance in ability, which left little unique variance to be accounted for by health and preparation. The standardized regression coefficient for the first ACT score was .97 in the no-bias model, and the coefficients for health and hours of preparation were only .10 and .04, respectively (both $p < .001$). The coefficient for hours of preparation (.02) was not statistically significant in the restricted general-bias model. Preparation, health, and prior achievement were all negatively and significantly related to anxiety in regression results. Of those variables, prior achievement had the strongest relationship with anxiety.

Table 6. Standardized SEM Parameter Estimates

Latent Variable	Observed Variable	Type	Basic		Expanded	
			Res. Gen. Bias	No Bias	Res. Gen. Bias	No Bias
Anxiety	Not Taken Classes	Loading	.93***	.93***	.93***	.93***
Anxiety	Areas Not Covered	Loading	.90***	.90***	.90***	.90***
Anxiety	Done Nothing	Loading	.78***	.78***	.80***	.80***
Anxiety	Realize Importance	Loading	.84***	.84***	.83***	.83***
Anxiety	Common Anxiety	Loading	.82***	.82***	.81***	.81***
Anxiety	ACT English	Loading	-.22		-.13	
Anxiety	ACT Math	Loading	-.27		-.19	
Anxiety	ACT Reading	Loading	-.22		-.14	
Anxiety	ACT Science	Loading	-.27		-.19	
Ability	ACT English	Loading	.87***	.88***	.88***	.89***
Ability	ACT Math	Loading	.77***	.83***	.78***	.83***
Ability	ACT Reading	Loading	.80***	.83***	.81***	.83***
Ability	ACT Science	Loading	.84***	.90***	.83***	.88***
Health	Hours Sleep	Loading			.32***	.32***
Health	Breakfast	Loading			.80***	.80***
Health	Snack	Loading			.41***	.41***
Anxiety	Ability	Corr.	-.02	-.30***	.74	-.31***
Health	First ACT Comp.	Corr.			.26***	.26***
Anxiety	Hours Prep.	Regress.			-.13***	-.13***
Anxiety	Health	Regress.			-.13**	-.13**
Anxiety	First ACT Comp.	Regress.			-.24***	-.24***
Ability	Hours Prep.	Regress.			.02	.04***
Ability	Health	Regress.			.08**	.10***
Ability	First ACT Comp.	Regress.			.96***	.97***
Hours Prep.	First ACT Comp.	Regress.			-.16***	-.16***

* $p < .05$, ** $p < .01$, *** $p < .001$

Sensitivity of results to sample.

The SEMs were fit with different data sets to check whether results were sensitive to the use of multiple imputation or to decisions about data filtering. First, the models were re-fit with

the 1,771 respondents having complete data (no imputation). Standardized coefficient estimates were nearly identical (within .02), though smaller sample sizes led to larger standard errors and higher p -values when comparing models using $\Delta\chi^2$. However, significance test

results were consistent, so conclusions would be unchanged. The models were also re-fit with 20 imputed data sets of 8,016 students, including those who reported experiencing debilitating test anxiety the first time they took the ACT. Again, parameter estimates were largely unchanged, but the interference effects were statistically significant ($p < .05$ or $p < .01$). This was apparently due to increased estimation precision afforded by a larger sample size rather than true differences. Indeed, χ^2/df and RMSEA indicated better fit for the no-bias model, and both models accounted for approximately 74% of the variance in observed ACT scores.

Differential anxiety. The final set of analyses addressed the third research question by investigating differential interference effects across demographic groups. In this analysis, only the simpler SEM with anxiety and ability was fit because the more complex model did not converge well for the relatively small group of minority examinees. First, the restricted general-bias and no-bias models were fit separately for female, male, minority, and White students. In the male, female, and minority analyses, $\Delta\chi^2$ was not statistically significant, nor were the interference effects. When analyzing White students only, the restricted general-bias model fit significantly better than the no-bias model ($\Delta\chi^2 = 16.6$, $\Delta df = 3$, $p < .001$; $\Delta CFI = -.001$). Moreover, the interference effects were large (approximately $-.70$) and statistically significant ($p < .05$). Note, however, that $\Delta\chi^2$ was not statistically significant ($p = .20$) when the model was fit with random samples of 541, which would be required to achieve power of $.80$. The deficit effects for minority and White examinees were $-.20$ and $-.29$, respectively.

The next analysis fit separate restricted general-bias models for females and males simultaneously, first with all parameters

estimated freely (as in the preceding analysis), then with the deficit and interference effects set equal. The deficit effects for females and males were both $-.30$, and the difference in model-data fit for the gender analysis was non-significant ($\Delta\chi^2 = 1.06$, $\Delta df = 5$, $p = .949$). The same analysis, when run with data from minority and White examinees, detected significant differences in fit between the models ($\Delta\chi^2 = 13.24$, $\Delta df = 5$, $p < .05$; $\Delta CFI = -.001$). Based on this result and the preceding analyses, there seemed to be stronger evidence of measurement bias for the White examinees, but power analyses again suggested that the significant $\Delta\chi^2$ was an artifact of large sample size. Specifically, the p -value for $\Delta\chi^2$ was $.53$ when the models were refit with random samples of 1,500 (a sample size of only 338 was needed to achieve power of $.80$, but substantial reductions in the minority sample caused convergence problems).

Discussion

This study applied established methods of investigating measurement bias due to test anxiety in the context of college admissions testing. As in several prior studies (Reeve & Bonaccio, 2008; Sommer & Arendasy, 2014; Sommer & Arendasy, 2016), the weight of evidence from this study was consistent with the deficit hypothesis. That is, observed associations between test anxiety and performance were accounted for by the correlation between anxiety and ability. In statistical results, interference effects were non-significant, and model-data fit was similar for the restricted general-bias and no-bias models. In other words, examinees who reported greater test anxiety tended to be lower in ability, which explains why they performed worse on the ACT than examinees reporting lower anxiety. So, in answer to the first research question, results indicate that debilitating test anxiety does not bias performance on the ACT.

The second research question concerned the associations between contextual factors and test anxiety and ability. In the SEMs, anxiety and ability were regressed on prior achievement, hours of preparation, and a latent variable reflecting physical health factors. All three predictors had significant, negative associations with anxiety, but prior achievement was the best predictor of anxiety. That is, students who performed better the first time taking the ACT tended to be less anxious the second time, but not necessarily because they performed better the first time. Indeed, a large number of prior studies also detected a negative correlation between anxiety and ability (Hembree, 1988), which was strongly correlated with prior achievement. That strong correlation was apparent in the regression results for ability. Indeed, the standardized regression coefficient for prior achievement was .97, which left little unique variance to be accounted for by hours of preparation and health. Comparing those two factors, health was relatively important as a predictor of ability. The coefficient for hours of preparation was negligible, which confirms other analyses showing modest ACT score increases associated with additional hours of preparation (Schiel & Valiga, 2014).

Several questionnaire items related to assessment context were omitted from the SEMs, but descriptive analyses still support conclusions about them. Namely, results indicated that testing in a familiar location and knowing someone in the testing room were not useful predictors of test performance and they were uncorrelated with anxiety. These findings corroborated prior studies in which anxiety was unrelated to testing in a familiar or unfamiliar building (Hembree, 1988). That said, examinees who reported performing worse on the ACT because they had difficulty finding the testing site tended to get lower scores. However, the number of affected examinees

was very small, and the difference was not statistically significant when controlling for prior achievement.

The final research question dealt with possible differential effects of test anxiety for different demographic groups. As in much prior research (e.g., Hembree, 1988), females reported greater test anxiety than males, but SEM analyses comparing males and females provided no evidence of differential bias. This finding is consistent with prior research showing no differential relationship between test anxiety and performance for males and females (Seipp, 1991). Also consistent with prior studies (e.g., Hodge, McCormick, & Elliot, 1997), descriptive analyses revealed greater self-reported test anxiety among minority examinees compared to White examinees. Subsequent analyses revealed stronger deficit effects and interference effects for White examinees. That is, White examinees were less likely to report test anxiety, but when they did, there was statistical evidence of measurement bias. Despite the magnitude and significance of the interference effects for White students, the effects of measurement bias were apparently very small since model-data fit was so similar for the no-bias and restricted general-bias models. Specifically, ΔCFI was only $-.001$, the difference in χ^2/df was only $-.003$, and the two models each accounted for approximately 68% of the variance in observed assessment scores. Moreover, refitting the models with smaller samples informed by power analyses resulted in non-significant differences in χ^2 statistics.

Limitations

The specific conditions of a test anxiety study limit the generalization of results. This study utilized data from a sample of test takers that were likely on track to attend college. This was apparent from their relatively high average ACT score, high average grade-point average, and

the fact that they took the ACT twice—presumably to improve their scores for college applications. Results and conclusions might have differed with data from a sample representing the full variance in ability of high school students. Likewise, results might have differed with another measure of anxiety or with different timing of the anxiety measurement. As noted previously, the anxiety questionnaire items were not developed to focus on a particular type of test anxiety, which limits the specificity of conclusions from this study (e.g., state vs. trait anxiety, cognitive vs. affective aspects of anxiety).

The anxiety latent trait in this study primarily reflected feelings of being unprepared for the test, which would be expected to correlate with ability and performance in the manner described by the deficit hypothesis. Since lack of preparation is a common explanation for anxiety (Bonaccio & Reeve, 2010), it is possible that other studies with similar results also detected the effect of anxiety caused by lack of preparation, even if they did not measure it directly. Though results of this study may not generalize broadly, they were consistent with several other studies, and they apply directly to a testing context that affects millions of students each year.

Overall, results were consistent with the deficit hypothesis, but they do not prove the deficit hypothesis or disprove the interference hypothesis for college admissions testing.

Indeed, the deficit hypothesis and the interference hypothesis should not be treated as a dichotomy because deficit and interference effects could be acting simultaneously. In statistical terms, the anxiety-ability correlation may not fully explain the association between anxiety and observed test performance. Interference effects in the main analysis were not significantly different from zero, but they were all negative, which is

consistent with the notion that test anxiety exerts a negative bias on test performance. In this study, that bias was negligible, but it could be greater in other assessment contexts.

When sample size is low, a study may lack the power needed to detect differences in model-data fit and significant interference effects. For example, with only 185 examinees, Reeve and Bonaccio (2008) detected no significant differences in model-data fit between the general-bias and no-bias models, and interference effects as large as $-.29$ were not statistically significant. With much larger sample sizes, the current study faced different challenges. Nearly every test involving χ^2 or $\Delta\chi^2$ was statistically significant, yet most measures of model-data fit indicated acceptable or similar fit. The interference effects, which were approximately $-.25$ or $-.16$ depending on the model, might have been statistically significant with greater sample size, and this could have affected conclusions drawn from the study. Of course, large sample sizes are generally desirable, but this study highlights some challenges they pose for interpreting results from null-hypothesis significance testing. Power analyses proved useful for evaluating whether large χ^2 and $\Delta\chi^2$ statistics reflected large sample size rather than poor fit, and future studies should continue this practice as needed.

Practical Implications

Recent investigations suggest that the prevalence of test anxiety may be increasing (McDonald, 2001). At the same time, concerns about test anxiety have grown, in part due to increasing pressure on students to perform well on accountability tests (Cizek & Burg, 2006). Rightly so, these concerns focus primarily on students' well-being rather than assessment validity. That is, school districts mainly want to prevent anxiety and associated

physical symptoms such as headaches, nausea, and poor sleep and emotional reactions like worry, withdrawal, and frustration (Heiser et al., 2015).

Similar reactions may also be elicited by high-stakes college admissions tests, which are increasingly being used for 11th-grade accountability testing. Even if it does not cause measurement bias, debilitating test anxiety is real, and so are the unpleasant symptoms. Regardless of the effect on validity, minimizing test anxiety would promote students' physical and emotional well-being, and it could reduce negative perceptions of achievement testing. To that end, this section provides practical recommendations for minimizing test anxiety based on empirical evidence.

First and foremost, examinees should take classes covering the content measured by the test. In this study, the anxiety latent variable largely reflected students' self-assessments of their lack of preparation, which is a common reason for experiencing test anxiety (Bonaccio & Reeve, 2010). Given that hours of test preparation was very weakly associated with estimated ability, the most valuable test preparation may occur in classrooms. That said, some preparation for a specific college admissions test is recommended to at least become familiar with the test's format, content, and length. Suddenly realizing the test's importance was another explanation for test anxiety, so counselors can possibly help examinees come to terms with the test's importance well in advance of testing day. This could also provide motivation to better prepare for the test.

Most examinees have no choice about the environment in which they take a college admissions tests, so they cannot control whether they test in a familiar location or with familiar people in the room. Fortunately, this study indicated that those factors were unrelated to test performance and anxiety when controlling for prior achievement. One recommendation is to make sure that examinees can easily locate the testing center.

This study suggests some straightforward recommendations regarding physical health factors. Namely, examinees should get adequate sleep the night before testing, eat breakfast the day of testing, and bring a snack to eat during break time. In SEM results, the health latent variable was negatively associated with anxiety and positively associated with ability. Though the regression coefficients for health were small in magnitude, they revealed possible double benefits of good health: lower anxiety and higher estimated ability.

Though not supported by this study, test administrators and educational institutions may consider other actions to reduce test anxiety. For example, keeping the testing environment free of distraction and providing appropriate memory supports have both proven effective at reducing test anxiety, as have certain behavioral or cognitive interventions (Hembree, 1988). Other efforts could focus on student development of positive psychosocial attributes that are negatively associated with test anxiety such as academic buoyancy (Putwain, Daly, Chamberlain, & Sadreddini, 2015) and academic self-efficacy (Nie, Lau, & Liao, 2011).

Conclusions

The major empirical contribution of this study was providing evidence that anxiety does not induce measurement bias in college admissions testing—in general and for certain demographic groups. Results were consistent with the hypothesis that ability explains the observed relationship between anxiety and performance. Though the expected effect of minimizing debilitating test anxiety on assessment validity would be negligible, efforts to reduce anxiety could still be beneficial for examinees' well-being. The best way to reduce anxiety is to ensure that students feel well prepared by taking high school courses that cover test content, but other results from this study translate into practical recommendations that could be implemented by most examinees with little effort and at low cost (e.g., get adequate sleep, eat breakfast, consider the importance of the test beforehand, and know the location of the testing site). Test publishers can acknowledge that assessments may seem threatening at times, but there are steps students can take to ready themselves for the challenges of college admissions testing.

References

- ACT. (2017). *Technical manual: The ACT*. Iowa City, IA: ACT. Retrieved from https://www.act.org/content/dam/act/unsecured/documents/ACT_Technical_Manual.pdf
- Alpert, R., & Haber, R. (1960). Anxiety in academic achievement situations. *Journal of Abnormal and Social Psychology, 61*(2), 207–215. doi:10.1037/h0045464
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. doi:10.1037/0033-2909.107.2.238
- Birenbaum, M., & Pinku, P. (1997). Effects of test anxiety, information organization, and testing situation on performance on two test formats. *Contemporary Educational Psychology, 22*(1), 23–38. doi:10.1006/ceps.1997.0923
- Bonaccio, S., & Reeve, C. L. (2010). The nature and relative importance of students' perceptions of the sources of test anxiety. *Learning and Individual Differences, 20*(6), 617–625. doi:10.1016/j.lindif.2010.09.007
- Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research, 21*(2), 230–258. doi:10.1177/0049124192021002005
- Cassady, J. C. (2004). The impact of cognitive test anxiety on text comprehension and recall in the absence of external evaluative pressure. *Applied Cognitive Psychology, 18*(3), 311–325. doi:10.1002/acp.968.
- Chamberlain, S., Daly, A. L., & Spalding, V. (2011). The fear factor: Students' experiences of test anxiety when taking A-level examinations. *Pastoral Care in Education, 29*(3), 193–205. doi:10.1080/02643944.2011.599856
- Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A., & McCann, N. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology, 97*(2), 268–274. doi:10.1037/0022-0663.97.2.268
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255. doi:10.1207/S15328007SEM0902_5
- Cizek, G. J., & Burg, S. S. (2006). *Addressing test anxiety in a high-stakes environment: Strategies for classrooms and schools*. Thousand Oaks, CA: Corwin.
- Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1*(2), 170–177. doi:10.1037/1082-989X.1.2.170
- Dusek, J. B. (1980). The development of test anxiety in children. In I. G. Sarason (Ed.), *Test anxiety: Theory, research and application* (pp. 87–110). Hillsdale, NJ: Erlbaum.

- Dusek, J. B., Mergler, N. L., & Kermis, M. D. (1976). Attention, encoding, and information processing in low- and high-test-anxious children. *Child Development, 47*(1), 201–207. doi:10.1111/1467-8624.ep12432731
- Halpin, P. F., da-Silva, C., & De Boeck, P. (2014). A confirmatory factor analysis approach to test anxiety. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(3), 455–467. doi:10.1080/10705511.2014.915377
- Heiser, P., Simidian, G., Albert, D., Garruto, J., Catucci, D., Faustino, P., . . . Caci, K. (2015). *Anxious for success: High anxiety in New York's Schools*. Latham, NY: New York State School Board Association & New York Association of School Psychologists. Retrieved from http://www.nyssba.org/clientuploads/nyssba_pdf/Test_Anxiety_Report.pdf
- Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*(1), 47–77. doi:10.3102/00346543058001047
- Hill, K. T., & Wigfield, A. (1984). Test anxiety: A major educational problem and what can be done about it. *Elementary School Journal, 85*, 105–126. doi:10.1086/461395
- Hodapp, V., & Benson, J. (1997). The multidimensionality of test anxiety: A test of different models. *Anxiety, Stress, & Coping, 10*(3), 219–244. doi:10.1080/10615809708249302
- Hodge, G. M., McCormick, J., & Elliot, R. (1997). Examination-induced distress in a public examination at the completion of secondary schooling. *British Journal of Educational Psychology, 67*(2), 185–197. doi:10.1111/j.2044-8279.1997.tb01236.x
- Hong, E. (1999). Test anxiety, perceived test difficulty, and test performance: Temporal patterns of their effects. *Learning and Individual Differences, 11*(4), 51–69. doi:10.1016/S1041-6080(99)80012-0
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. doi:10.1080/10705519909540118
- Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology, 30*(2), 299–311. doi:10.1007/s10869-014-9357-6
- Huang, J. L., Curran, P. G., Keeney, J., Potoski, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27*(1), 99–114. doi:10.1007/s10869-011-9231-8
- Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality, 39*(1), 103–129. doi:10.1016/j.jrp.2004.09.009
- Keogh, E., Bond, F. W., French, C. C., Richards, A., & Davis, R. E. (2004). Test anxiety, susceptibility to distraction and examination performance. *Anxiety, Stress, and Coping, 17*(3), 241–252. doi:10.1080/10615300410001703472

- Kilmen, S. (2015). Why do the tests make us anxious? *International Journal of Social Sciences and Education*, 5(4), 604–612.
- Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.
- Liebert, R., & Morris, L. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports*, 20(3), 975–978. doi:10.2466/pr0.1967.20.3.975
- MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, 11(1), 19–35. doi:10.1037/1082-989X.11.1.19
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1(2), 130–149. doi:10.1037/1082-989X.1.2.130
- Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, 48(1), 61–83. doi:10.1016/j.jrp.2013.09.008
- McDonald, A. S. (2001). The prevalence and effects of test anxiety in school children. *Educational Psychology*, 21(1), 89–101. doi:10.1080/01443410020019867
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17(3), 437–455. doi:10.1037/a0028085
- Muthén, B. O. (1989). Using item specific instructional information in achievement modeling. *Psychometrika*, 54(3), 385–396. doi:10.1007/BF02294624
- Nie, Y., Lau, S., & Liao, A. K. (2011). Role of academic self-efficacy in moderating the relation between task importance and test anxiety. *Learning & Individual Differences*, 21(6), 736–741. doi:10.1016/j.lindif.2011.09.005
- Powers, D. E. (1986). *Test anxiety and the GRE general test*. (ETS Research Report 86-45). Princeton, NJ: Educational Testing Service. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/j.2330-8516.1986.tb00200.x/epdf>
- Putwain, D. W. (2007). Test anxiety in UK schoolchildren: Prevalence and demographic patterns. *British Journal of Educational Psychology*, 77(3), 579–593. doi:10.1348/000709906x161704
- Putwain, D. W. (2008). Test anxiety and GCSE performance: The effect of gender and socio-economic background. *Educational Psychology in Practice*, 24(4), 319–334. doi:10.1080/02667360802488765

- Putwain, D. W., Daly, A. L., Chamberlain, S., & Sadreddini, S. (2015). Academically buoyant students are less anxious about and perform better in high-stakes examinations. *British Journal of Educational Psychology, 85*(3), 247–263. doi:10.1111/bjep.12068
- Reeve, C. L., & Bonaccio, S. (2008). Does test anxiety induce measurement bias in cognitive ability tests? *Intelligence, 36*(6), 526–538. doi:10.1016/j.intell.2007.11.003
- Reeve, C. L., Bonaccio, S., & Charles, J. E. (2008). A policy-capturing study of the contextual antecedents of test anxiety. *Personality & Individual Differences, 45*(3), 243–248. doi:10.1016/j.paid.2008.04.006
- Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*(2), 353–387. doi:10.1037/a0026838
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. doi:10.18637/jss.v048.i02
- Rubin, D. B. (1986). Basis ideas of multiple imputation for nonresponse. *Survey Methodology, 12*(1), 37–47. Retrieved from <http://www.statcan.gc.ca/pub/12-001-x/1986001/article/14439-eng.pdf>
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.
- Sarason, I. G., & Stoops, R. (1978). Test anxiety and the passage of time. *Journal of Consulting and Clinical Psychology, 46*(1), 102–109. doi:10.1037/0022-006X.46.1.102
- Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors on covariance structure analysis. In A. von Eye, & C. C. Clogg (Eds.), *Latent variable analysis* (pp. 399–419). Thousand Oaks, CA: Sage.
- Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*(4), 507–514. doi:10.1007/BF02296192
- Schiel, J. L., & Valiga, M. J. (2014). *How the amount of time spent on test preparation relates to ACT Scores*. (ACT Information Brief 2014-7). Iowa City, IA: ACT. Retrieved from <http://www.act.org/content/dam/act/unsecured/documents/Info-Brief-2014-7-rev1.pdf>
- Seipp, B. (1991). Anxiety and academic performance: A meta-analysis of findings. *Anxiety Research, 4*(1), 27–41. doi:10.1080/08917779108248762
- Selkirk, L. C., Bouchey, H. A., & Eccles, J. S. (2011). Interactions Among Domain-Specific Expectancies, Values, and Gender: Predictors of Test Anxiety During Early Adolescence. *Journal of Early Adolescence, 31*(3), 361–389. doi:10.1177/0272431610363156
- Sommer, M., & Arendasy, M. (2014). Comparing different explanations of the effect of test anxiety on respondents' test scores. *Intelligence, 42*(1), 115–127. doi:10.1016/j.intell.2013.11.003

- Sommer, M., & Arendasy, M. E. (2016). Does trait test anxiety compromise the measurement fairness of high-stakes scholastic achievement tests? *Learning and Individual Differences, 50*(1), 1–10.
doi:10.1016/j.lindif.2016.06.030
- Spielberger, C. D., & Vagg, P. R. (1995). Test anxiety: A transactional process model. In C. D. Spielberger, & P. R. Vagg (Eds.), *Test anxiety: Theory, assessment and treatment* (pp. 3–14). Bristol, UK: Taylor & Francis.
- Steiger, J. H., & Lind, J. M. (1980). Statistically based tests for the number of common factors. Paper presented at the Annual Meeting of the Psychometric Society, Iowa City, IA.
- Tobias, S. (1979). Anxiety research in educational psychology. *Journal of Educational Psychology, 71*(5), 573–582. doi:10.1037/0022-0663.71.5.573
- von der Embse, N. P., & Witmer, S. E. (2014). High-stakes accountability: Student Anxiety and Large-Scale Testing. *Journal of Applied Psychology, 30*(2), 132–156. doi:10.1080/15377903.2014.888529 .
- Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology, 89*(5), 696–716. doi:10.1037/0022-3514.89.5.696
- Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68–81. doi:10.1006/ceps.1999.1015
- Wine, J. (1971). Test anxiety and direction of attention. *Psychological Bulletin, 76*(2), 92–104.
doi:10.1037/h0031332
- Zeidner, M. (1990). Does test anxiety bias scholastic aptitude test performance by gender and sociocultural group? *Journal of Personality Assessment, 55*(1/2), 145–161. doi:1080/00223891.1990.9674054
- Zeidner, M. (1991). Test anxiety and aptitude test performance in an actual college admissions testing situation: Temporal considerations. *Personality and Individual Differences, 12*(2), 101–109.
doi:10.1016/0191-8869(91)90092-P

ACT is an independent, nonprofit organization that provides assessment, research, information, and program management services in the broad areas of education and workforce development. Each year, we serve millions of people in high schools, colleges, professional associations, businesses, and government agencies, nationally and internationally. Though designed to meet a wide array of needs, all ACT programs and services have one guiding purpose—helping people achieve education and workplace success.



[ACT.org/research](https://act.org/research)

Copyright © 2018 by ACT, Inc. All rights reserved.