ABSTRACT

GASPERSON, SEAN MORGAN. The Effects of Item-by-Item Feedback during a Computer Adaptive Test. (Under the direction of Lori Foster Thompson.)

Decisions involved in designing high-stakes tests include whether to give examinees feedback and, if it is given, how to display the feedback to them. Researchers have yet to examine the effects of item response theory-related feedback on state test anxiety and examinees' perceptions of fairness when said feedback is given on an item-by-item basis – such as during a computer adaptive test. To address these gaps in the literature, this experiment examined the effects of feedback on state test anxiety and perceived test fairness, accounting for individual differences in feedback acceptance, exam performance, and performance goal orientation. These effects were examined across three modes of item-by-item feedback (i.e., text feedback, graphical feedback, and text/graphical feedback together) during a computer adaptive test in an online sample (N = 338). A three-way interaction occurred between text and graphical feedback, test performance, and prove performance goal orientation. As expected, among those who received feedback, there was a significant, positive relationship between feedback acceptance and perceived test fairness. Limitations and future directions are discussed.

The Effects of Item-by-Item Feedback during a Computer Adaptive Test

by
Sean Morgan Gasperson

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Psychology

Raleigh, North Carolina

2014

APPROVED BY:

Dr. Lori Foster Thompson Committee Chair	Dr. Adam W. Meade
Dr. Mark A. Wilson	Dr. S. Bartholomew Craig

BIOGRAPHY

Sean Gasperson was born in High Point, North Carolina. He graduated from T. Wingate Andrews High School in 2004, then attended Southern Wesleyan University from 2004-2008, where he received a Bachelor of Arts degree in Psychology. He received his Master's degree in General/Theoretic Psychology (Industrial/Organizational Psychology concentration) from East Carolina University in 2010. He began studying I/O Psychology at N.C. State in Fall 2010.

ACKNOWLEDGEMENTS

I would like to thank many people for their support during this journey. First, I would like to acknowledge my committee chair, Dr. Lori Foster Thompson, and my committee members, Drs. Adam Meade, Bart Craig, and Mark Wilson. Your advice, support, humor, and candor have been tremendous assets to me, both personally and professionally. Thank you all for the knowledge, skills, abilities, and other characteristics you have engendered in me to be the best scientist-practitioner that I can be, so that I may "spiral up" in the cycle of high performance.

This project and pathway would have been impossible without the love of my family and friends. My parents, George and Margaret, have loved me unconditionally and have helped me keep my name off of the board ever since I wanted to wear my Superman thing to school. My brother, Jesse, and sister-in-law, Carrie, have been the type of parents that I aspire to be. My ECU advisor, Mark Bowler, has been a valuable mentor and friend. Mark, thank you for keeping me in the in-group! To my cohorts – thank you all for keeping this arduous process enjoyable! I will always cherish your friendships.

Finally, and certainly not least, I would like to thank my wife, Jenna. You have endured the hardships of this journey as much as I have. Thank you for your love and understanding through it all. Your partnership in this life is worth a million doctorates. I am the luckiest man alive to have you by my side. Next to you is my favorite place to be.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

Many factors can affect test takers in high-stakes testing environments such as pre-employment testing or university entrance exams. Each year, decision makers and psychometricians spend a substantial amount of money selecting and developing tests for administration to applicants (Harper, 2008). Test design is vital to consider due to the potential ramifications of (a) examinees responding to test characteristics in different ways, and (b) examinees having adverse reactions to aspects of the test. Among the decisions involved in designing a test are whether to give examinees feedback, when to give feedback to them, and how to give feedback to them. These decisions could affect examinees' anxiety levels during the test as well as how fair they perceive the test to be. However, there is a dearth of literature devoted to this topic, particularly in the context of computer adaptive testing (CAT). The purpose of this study is to assess the effect of various modes of item-by-item feedback on examinees' state test anxiety and perceptions of test fairness during a computer adaptive test in a high-stakes testing situation.

Test Feedback

Research exploring the effects of giving feedback to test takers dates back to the early 1900s (e.g., Arps, 1920; Judd, 1905). Much of the early research focused on "knowledge of results," or providing test takers information about the outcome of the test (Kluger & DeNisi, 1996). More recently, the definition of feedback interventions (FIs) has been expanded to encompass any relevant information about the performance of a task, including specific feedback about *how* one performs the task. Kluger and DeNisi (1996) defined FIs as "actions taken by (an) external agent(s) to provide information regarding some aspect(s) of one's task

performance" (p. 255). Giving examinees feedback typically refers to providing them with some sort of report of their performance at the end of the test. However, researchers have explored the extent to which more complex characteristics of the feedback (e.g., content, frequency, mode, and source) affect outcomes such as performance on a particular task or job (for reviews, see Alvero, Bucklin, & Austin, 2008 and Balcazar, Hopkins, & Suarez, 1985).

Two dimensions on which feedback can vary are the timing of the feedback (i.e., when it is administered) and the mode of feedback (i.e., how the feedback is displayed to the test taker). Feedback is typically given at the end of the test or later, though some research has been devoted to assessing the effect of giving feedback after examinees complete each item (e.g., Butler, Karpicke, & Roediger, 2007; Clark, Fox, & Schneider, 1998). This research, though practical, is lacking in two ways. First, the content of the feedback is limited to whether examinees answered the item correctly or incorrectly. Second, the research has been mainly focused on examinees' preference for the timing of feedback and how delayed versus immediate feedback affects performance on the task (e.g., Butler et al., 2007). Other outcomes, such as examinees' test anxiety and perceptions of the test's fairness, should be explored.

Feedback can also vary by how it is given or displayed to the examinees. In conducting a literature review on the effect of various forms of feedback, Alvero et al. (2008) identified four primary methods of providing feedback: text, verbal, graphical, and mechanical (e.g., via video). Comparing text and graphical feedback, 50% of the studies that reported providing text feedback alone reported that the feedback was effective (i.e., the feedback produced the predicted effect), as did 50% of the studies assessing graphical

feedback effects. Text feedback that was accompanied by graphical feedback, however, was effective in 86% of the studies. Thus, research suggests that coupling text feedback with graphical feedback is more effective. Like feedback timing, however, research on feedback mode has primarily focused on performance- and satisfaction-based outcomes.

Computer adaptive testing and item response theory (IRT) provide a valuable platform to address shortcomings in the feedback literature. When examinees take computer adaptive tests, questions of varying difficulty levels are selected from a test bank based on the examinees' estimated level of an underlying trait (i.e., theta). The exam ends when a predetermined criterion is met, such as an examinee receiving a maximum number of items or when the standard error of their estimated theta reaches a predetermined value. Whereas traditional paper-based tests or non-adaptive computer-based tests do not have the capability to report real-time ability estimates, computer adaptive tests do. For example, one of the basic computer adaptive test demonstrations on the Concerto testing system developed by The Psychometrics Centre at the University of Cambridge displays IRT-based feedback graphically to test takers (University of Cambridge, 2014). This feedback includes theta estimates, item difficulty levels, test information, and item characteristic curves, and is a noted advantage of using the testing system. Reporting some or all of this information to examinees during the test is an option for test administrators who employ computer adaptive testing. Empirical information about the effects of this type of feedback in a high-stakes testing situation is lacking, and is needed to guide decisions about how the portrayal of this information affects test takers, particularly as related to their levels of anxiety and perceived fairness. Furthermore, research has yet to examine whether known text and/or graphical

feedback effects extend to tests that provide said feedback on an item-by-item basis. This study contributes to the literature by being the first to assess the effects of IRT-specific feedback in this way (i.e., item-by-item) during this type of test (i.e., CAT). Feedback could act as a method of focusing examinees' attention toward aspects of their performance. Alternatively, by presenting additional stimuli, it could distract them from the test. Research is needed to assess its effects on outcomes such as state test anxiety and perceived test fairness.

State Test Anxiety

Test anxiety has been operationalized as emotional, physiological, and behavioral responses to the potential consequences of negative outcomes on a test or exam (Zeidner, 1998). Test anxiety is induced when test takers perceive that the demands of the test situation exceed their intellectual, motivational, and/or social capabilities (Sarason & Sarason, 1990). Research on the topic has addressed a variety of concerns, including interventions for test anxiety (e.g., Ergene, 2003), personal and situational predictors of test anxiety (e.g., Putwain, Woods, & Symes, 2010), its relationship with various constructs such as learned helplessness (e.g., Akca, 2011), and perhaps most notably, its relationship with task performance (see Alam, 2013; Cohen, Ben-Zur, & Rosenfield, 2008; In'nami, 2006). Test anxiety can generally be divided into state test anxiety and trait test anxiety. State test anxiety refers to an anxious, temporary condition specific to a particular testing situation, while trait test anxiety refers to a more stable tendency to have anxiety across any testing situation (Hong, 1998).

Much research has been devoted to the effect of feedback on state test anxiety, often with the primary purpose of seeking to reduce test takers' anxiety (e.g., von der Embse,

Barterian, & Segool, 2013). Feedback has been thought to reduce test anxiety by eliminating ambiguity that may be present about performance on the tasks at hand (Hansen, 1974). Furthermore, because test takers will, in general, answer more items correctly than incorrectly, there is more opportunity for positive feedback, thus reducing apprehension about one's performance (DiBattista & Gosse, 2006). In an effort to assess the viability of item feedback in reducing state test anxiety, some researchers have found that contrary to what was hypothesized, feedback increased state test anxiety (Wise, Plake, Eastman, Boettcher, & Lukin, 1986). Conversely, other researchers have shown that providing feedback to examinees has resulted in significant reductions in state test anxiety over time (e.g., Hansen, 1974). Other research has reported no relationship between immediate feedback and test anxiety (Clark et al., 1998; DiBattista & Gosse, 2006). Thus, though researchers have explored the relationship between feedback and state test anxiety, there appears to be little consensus as to what effect, if any, it has. These inconsistencies in research can be evidence of a moderating effect, such that feedback's effects on state test anxiety may be contingent on a third variable. How well examinees perform on the test could serve as this moderator. High and low performers on tests may react differently to feedback due to the differing amounts of positive reinforcement they receive. Though performance's moderating influence on feedback's effects during an exam has not been investigated, the effect of providing end-of-exam performance reports on anxiety has. Bradshaw and Gaudry (1968) found that experiencing failure can result in higher state test anxiety, and experiencing success can result in lower state test anxiety. Participants were randomly assigned to fail or succeed on the test, and received performance feedback at the end of testing. Those who

received feedback of poor performance reported higher test anxiety, and those who received feedback of high performance reported lower test anxiety. This research, however, did not assess whether higher and lower performers react differently to real-time feedback.

Immediate feedback (i.e., item-by-item) inherently gives examinees an indication of test performance. As individuals take a test, immediate feedback makes them aware of the discrepancy (or lack thereof) between their current performance and their desired performance (Kluger & DeNisi, 1996). For better performers, this discrepancy will be lower. For poorer performers, it will be higher. Furthermore, feedback that gives examinees a more detailed item-by-item account of their performance and how their abilities are being estimated heightens examinees' awareness of that discrepancy. The typical alternative, giving examinees their score at the end of the test, does not have the capacity to influence state test anxiety throughout the testing situation because examinees may not be completely certain how many items they answered correctly or incorrectly. Thus, for poorer performers, detailed item-by-item feedback should lead to higher levels of state test anxiety. For better performers, detailed item-by-item feedback should lead to lower levels of test anxiety. Although some research has been devoted to assessing the main effect of feedback on state test anxiety (e.g., DiBattista & Gosse, 2006; Hansen, 1974), the moderating effect of performance on the relationship between feedback and state test anxiety during a test that provides ongoing item-level feedback throughout the test has not been examined.

> ***Hypothesis 1***: Test performance will moderate the effect of item-by-item feedback on state test anxiety, with feedback reducing anxiety for better performers and increasing anxiety for poorer performers.

As prior research has suggested that the way feedback information is presented to people can influence reactions to it (Alvero et al., 2008), this and all subsequent hypotheses will be tested across three modes of feedback: text feedback, graphical feedback, and text and graphical feedback together.

There may be types of individuals for whom the moderating effect of test performance is especially likely. In particular, examinees oriented toward performance goals may be particularly sensitive to feedback conveying performance failures and successes. As an overall construct, goal orientation refers to how people react to opportunities to demonstrate or develop ability (Dweck, 1986). Goal orientation can be divided into learning goal orientation (i.e., seeking to master and acquire new skills) and performance goal orientation (i.e., seeking opportunities to demonstrate or validate one's competence; Dweck & Leggett, 1988). Vandewalle (1997) further differentiated performance goal orientation into *prove* performance goal orientation (PPGO) and *avoid* performance goal orientation (APGO). PPGO is one's tendency to seek opportunities that may demonstrate competence, while APGO is one's tendency to avoid situations that may disprove competence. In a meta-analysis examining the relationship of goal orientation dimensions and proximal and distal outcomes, Payne, Youngcourt, and Beaubien (2007) found that individuals with high levels of APGO and PPGO were likely to have high levels of state test anxiety. As item-by-item feedback provides immediate knowledge of results, individuals who are high in performance goal orientation (both avoid and prove) and receive feedback indicating that they are performing poorly on the test should exhibit especially high levels of state test anxiety.

Hypothesis 2: There will be a three-way interaction between feedback, test performance, and performance goal orientation on state test anxiety, such that performance's moderating influence on the effect of feedback is more pronounced for those more oriented toward avoid performance (H2a) and prove performance (H2b) goals.

Perceived Test Fairness

Moving beyond state test anxiety, feedback's influence over other outcome variables is also important to consider. Perceptions of a test's fairness, in particular, warrant consideration. Perceptions of test fairness can be understood by examining the organizational justice literature. Organizational justice has been a popular area of research throughout recent years. It refers to "people's perceptions of fairness in organizations along with their associated behavioral, cognitive, and emotional reactions" (Greenberg, 2011, p. 271). Three primary types of organizational justice have been identified – distributive justice, procedural justice, and interactional justice (Bierhoff, Cohen, & Greenberg, 1986; Bies & Moag, 1986; Folger & Greenberg, 1985). Procedural justice (i.e., the perceived fairness of the processes that are used when organizational decisions are made; Folger & Greenberg, 1985) is thought to be composed of ten "rules" which, when violated, lead to perceptions of injustice (Gilliland, 1993; Leventhal, 1980). As pre-employment testing is included in the organizational decision-making process (more specifically, the hiring process), much research about procedural justice focuses on perceptions of test fairness.

One of the procedural justice rules is feedback. Feedback can be given in various forms, including from others and from the test itself. Research has supported the notion that

feedback during the testing process can affect perceived test fairness. Specifically, research has indicated that the mere presence of feedback during the testing process can result in higher evaluations of fairness (Lounsbury, Bobrow, & Jensen, 1989). Furthermore, other research has shown a significant effect of feedback on post-feedback fairness perceptions (van Vianen, Taris, Scholten, & Schinkel, 2004). Feedback, when given throughout the test, provides information to examinees regarding how they are performing and how their final scores are calculated. Particular to computer adaptive testing, item-by-item feedback can facilitate examinees' understanding of how their scores are calculated by providing information about the scoring process. In essence, rather than simply providing a final score after testing, information such as real-time ability (i.e., theta) estimates, the precision with which the examinees' ability is being estimated (i.e., standard error of measurement), and the difficulty of the items can be given to the examinee, providing transparency and clarity about the methods of item response theory. This transparency is a defining characteristic of procedural justice – namely, facilitating examinees' knowledge about the how the tests are selected, constructed, and used (Gilliland, 1993).

Though feedback can be an important factor in the evaluation of fairness perceptions, there may be some conditions under which feedback is especially likely to raise fairness perceptions. Two potential moderators worth exploring are performance on the test and feedback acceptance. How well an examinee performs can affect the relationship between feedback and perceived test fairness. As previously noted, examinees who receive feedback are more likely to perceive the test as fair. Furthermore, research has demonstrated that better performers are more likely to perceive the test as fair (Bauer, Maertz, Dolen & Campion,

1998; Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993). Researchers have attributed this relationship to self-serving biases (Chan, Schmitt, Jennings, Clause, & Delbridge, 1998). That is, poorer performing examinees are more likely than better performing examinees to evaluate a test as unfair in order to maintain a positive view of themselves by means of threat-reduction or ego-enhancement. As examinees receive detailed item-by-item feedback throughout the test, any ambiguity regarding their performance will be reduced. For better performers, item-by-item feedback should reinforce the higher perceptions of fairness that they hold about the test, while for poorer performers, item-by-item feedback should reinforce the lower perceptions of fairness that they hold about the test.

Though previous research has examined the main effects of feedback and test performance on perceived test fairness, there is a dearth of research exploring their interactive effects. It is possible that providing real-time information about the test and examinees' test performance, combined with how well one performs on the test, could affect perceptions of test fairness.

> ***Hypothesis 3****:* Test performance will moderate the degree to which item-by-item feedback affects perceived test fairness, such that feedback will increase perceived test fairness for better performers and decrease perceived test fairness for poorer performers.

Another variable worthy of examination is feedback acceptance. Some research suggests that the examinee must accept the feedback in order for it to be effective (e.g., Nease, Mudgett, & Quiñones, 1999). Feedback acceptance refers to the extent to which the feedback recipients believe that the feedback regarding their performance is accurate (Ilgen,

Fisher, & Taylor, 1979). Research has previously examined the moderating effect of feedback acceptance on the relationship between feedback and outcomes such as satisfaction with the test and fairness perceptions. Tonidandel, Quiñones, and Adams (2002) found that when participants accepted test feedback and had accurate perceptions of their performance, there was a positive relationship between feedback and perceptions of test fairness. When participants did not accept test feedback and had inaccurate perceptions of their performance, there was no relationship between feedback and fairness perceptions. Thus, it appears that examinees may evaluate whether the performance feedback aligns with their perceptions of their own performance when deciding whether to accept the feedback. Though previous research (e.g., Tonidandel et al., 2002) provides some evidence of the relationship of feedback acceptance and perceived test fairness, the feedback was only given at the end of the test. While end-of-test feedback is informative, research has not provided a thorough examination of real-time feedback as it relates to perceived test fairness.

Hypothesis 4*:* Among individuals who receive item-by-item feedback, there will be a positive relationship between feedback acceptance and perceived test fairness.

Method

Participants

Participation for this study was solicited via Amazon Mechanical Turk (MTurk). MTurk is a crowdsourcing website through which researchers can pay participants to complete various *Human Intelligence Tasks* (HITs). Research has consistently demonstrated its viability compared to other sources of data (Azzam & Jacobson, 2013; Behrend, Sharek, Meade, & Wiebe, 2011; Casler, Bickel, & Hackett, 2013; Goodman, Cryder, & Cheema,

2013). This study's sample was restricted to individuals who speak English as a first language and who reside in the United States. This is due to the fact that some items in the test require a nuanced understanding of the English language.

Design

This study employed a between-groups design. Participants were randomly assigned to one of four conditions that varied according to the mode of feedback that they received: control (no feedback), text feedback only, graphical feedback only, or text and graphical feedback. The mode of feedback is the independent variable, with the four levels described above. There were two dependent variables – state test anxiety and perceived test fairness. Test performance and goal orientation were tested as moderators, and a correlational relationship between feedback acceptance and perceived test fairness was assessed.

Procedure

This study involved the administration of a high-stakes test measuring general reasoning ability. A high-stakes test was desired to allow for generalizability of this study's results to high-stakes situations such as employment selection or entrance to academic programs. In particular, practice items for the Miller Analogies Test (MAT) were used to measure general reasoning. The MAT has been used as an entrance exam for graduate programs as well as a pre-employment selection tool (Kuncel, Hezlett, & Ones, 2004). A meta-analysis exploring the utility of the MAT relative to other common exams and its relationship to certain outcomes revealed significant correlations with both forms of the Graduate Record Examination (i.e., verbal and quantitative), general ability and reasoning measures, graduate grade point average, and job performance (Kuncel et al., 2004).

In order to develop the computer adaptive test needed to test the study hypotheses, preliminary data were collected to estimate the item parameters. This necessitated a “phase 1” of data collection for exam development purposes prior to collecting “phase 2” data for hypothesis testing. For phase 1, participants were given one of two subsets of practice items for the MAT. The test was divided into subsets due to the fact that there are 150 total MAT practice items. Administering all 150 items to each participant would take a substantial amount of time, and would introduce the possibility of test fatigue. Therefore, a common equating method was used to avoid this problem. This entailed giving two smaller subsets to different groups of participants, with a few items that are common to both subsets. Each subset contained a total of 85 items. Sixty five of these items were unique to each subset, while 20 items were common across the two subsets. These common items served as the anchor items to link the two subsets. The item characteristics of the anchor items were used to adjust the item characteristics of the unique items so that all 150 reasoning items were placed on a common metric. The item parameters of all the reasoning items were used to determine which items were given throughout the subsequent computer adaptive test. To ensure that the content of the two subsets was as identical to the other as possible, the area of knowledge that each reasoning item measures were labeled (e.g., arts, math, science), and the items were divided such that each area of knowledge was represented equally across the two subsets. Procedurally, participants were first presented with an informed consent form after volunteering to participate. Participants answered items that assessed various demographic variables such as age and gender, and then proceeded to complete a subset of the MAT items. Participants were randomly assigned to receive one of the two subsets. Participants for phase

1 were compensated $0.75. A total of 488 participants signed up to complete phase 1. Participants' responses were discarded if they did not complete the entire test, spent fewer than ten minutes taking the test, or spent greater than 120 minutes taking the test. This left 335 usable responses for phase 1.

Before the item subsets were linked, dimensionality and reliability were assessed. It is inappropriate to equate two test forms that measure different constructs and that are not equally reliable (Dorans & Holland, 2000). Thus, an exploratory factor analysis (EFA) was conducted on each item subset, and the factor structure of each test was compared. Concurrently, the reliability of each measure was assessed. Items that displayed low item total correlations and poor factor loadings were discarded. The factor structures and reliability estimates of both forms were adequately similar. Thus, it was appropriate to combine the forms into a single test. After combining the forms, items with discriminations lower than .5 were discarded. In total, ten analogy items were discarded, leaving 140 final items that were entered into the test bank for the CAT.

Phase 2 of data collection was executed to test this study's hypotheses. Phase 2 involved giving participants the computer adaptive test that contains the reasoning items, followed by the measures of this study's dependent and moderator variables. Participants were presented with an informed consent form after volunteering to participate. If they agreed to participate in the study, they were presented with instructions that evoked a high-stakes testing situation. This informed them that the three examinees who score the highest on the reasoning test would receive Amazon.com gift cards ($35 for first, $25 for second, $15 for third). Participants for phase 2 were compensated $1.00, with the possibility of

receiving additional incentives. A total of 380 participants signed up to complete this phase. Participants' responses were discarded if they had incomplete data or spent fewer than two minutes on the assessments of the dependent and moderator variables examined in this study. This left 338 usable responses for phase 2.

Participants were randomly assigned to one of the four feedback conditions. Each condition varied with regard to whether and how the feedback information was displayed on the screen after answering each item. The control condition contained neither the text nor graphical feedback; that is, it provided no information about examinees' estimated theta, etc. The three feedback conditions displayed examinees' estimated theta, the standard error of their theta estimate, and the item difficulty numerically after each item, accompanied by a brief description of how this information could be interpreted. These terms (e.g., theta estimate, item difficulty, etc.) were described in a way that could be understood by laypersons, and were shown on every screen throughout the test. The text-only feedback condition displayed the numbers corresponding with these terms without any graphs. For the graphical feedback condition, three separate line graphs displaying examinees' estimated theta, the precision of their theta estimate, and the item difficulty were displayed, accompanied with verbal explanations of these terms. Examinees could track their progress and real time score by looking at the text and/or graph. The combined text and graphical feedback condition contained all of the information that was displayed in the text and graphical conditions. The difference between the text and graphical (combined) condition and the graphical condition is that the graphical condition did not display numerical values of theta, standard error of measurement, and item difficulty, as the text condition did—only

graphical depictions of them. An example of each condition can be found in Appendix A. Before participants completed the computer adaptive test, they were given a brief tutorial to familiarize them with the terminology and feedback they would receive during the test. Participants in the control condition were only given a tutorial that related to the analogy items. Participants in the feedback conditions were given the analogy item tutorial as well as a feedback tutorial that provided pictorial examples accompanied with explanations of the feedback they would receive during the test. In the feedback conditions, participants were required to answer three questions about the tutorial to ensure that they understood it.

The CAT began with an item that corresponded to a theta level of zero. Each subsequent item was selected by giving examinees an item that they have not previously taken that provides the most information at their current estimated theta level. The test terminated when the standard error for the examinees' estimated theta reached a predefined value (i.e., no substantial additional information was being gained). Scores were estimated via the expected a posteriori (EAP) method. After the test terminated, examinees were directed to another web page where they completed subsequent measures assessing the moderator and outcome measures of interest.

Measures

Demographics (8 items). Eight items were administered in phases 1 and 2 of data collection to assess participants' age, gender, race, work experience, country of origin, native language, country of residence, and education. These data were collected to provide a description of the samples used in the study. These and all other items are listed in the proposal document in Appendix B.

General reasoning (150 items). One hundred and fifty practice items for the Miller Analogies Test (MAT) formed the basis for phase 1 and 2 data collection. These items were developed to be used by individuals who are preparing to take the MAT (Study Mode, 2014). The MAT is a series of analogies which requires examinees to infer the relationship between two words and apply that relationship to another set of words. Examinees must select a word that completes the word pair that is most similar to the given word pair. An example item is "(____) : PUCCINI :: SCULPTURE : OPERA." The response options for this item are "A. Cellini, B. Rembrandt, C. Wagner, D. Petrarch."

The items that assessed the dependent and moderator variables were given only during phase 2, and were presented with Likert-type response scales, with scores ranging from 1 (*strongly disagree*) to 7 (*strongly agree*).

State test anxiety (8 items, $\alpha = .93$). State test anxiety was measured via a modified eight-item measure from Hong and Karstensson (2002). Four items measure state worry, and four items measure state emotionality. An example item of state worry is "During the test I got so nervous that I forgot facts I really knew." An example item of state emotionality is "During the test, I felt very tense." The scale was modified by removing context-specific language from one item (i.e., "in the course") and replacing it with more general language (i.e., "on the test").

Goal orientation (8 items). Vandewalle's (1997) performance goal orientation measure was used. This measures prove performance goal orientation (4 items, $\alpha = .81$) and avoid performance goal orientation (4 items, $\alpha = .87$). An example item for prove performance goal orientation is "I'm concerned with showing that I can perform better than

my coworkers." An example item for avoid performance goal orientation is "I would avoid taking on a new task if there was a chance that I would appear rather incompetent to others."

Perceived test fairness (8 items, α = .84). Perceptions of test fairness were assessed via an eight-item measure adapted from Tonidandel et al. (2002). An example item is "Under the circumstances, the test was fair." The test was modified by removing context-specific language in one item (i.e., "knowledge of general psychology") and replacing it with more general language (i.e., "general reasoning").

Feedback acceptance (4 items, α = .87). A four-item scale taken from Tonidandel et al. (2002) measured feedback acceptance. An example item is "I do not believe that the feedback is accurate."

Manipulation check (4 items, α = .94). Four items were given to examinees to assess whether the experimental manipulations (i.e., the feedback conditions) were salient to examinees. That is, these items were given to ensure that during the CAT, examinees who received feedback were aware that they were receiving it. These items included: "Each time I answered a test question, I received feedback about how I was doing on the test," "I always knew how well I was doing while taking the test because of the feedback display," "Feedback was provided during the test to show my score," and "During the test, there was feedback that gave me updates about my test score."

Attention check (3 items). Three items were used to evaluate whether examinees were responding carelessly. An example item is "Please answer strongly agree."

Results

After participants were discarded based on the criteria of incomplete data and time spent on dependent and moderator variables, attention check items were used to check for careless respondents. No remaining participant in any condition answered more than one attention check incorrectly, so none were discarded based on this criterion. After the data were cleaned and reliabilities were estimated, skewness and kurtosis of all variables were checked to assess whether any assumptions of normality would be violated. All values were determined to be acceptable. Next, a one-way ANOVA was employed to assess the salience of the experimental manipulations to the examinees. Results indicated that participants in the text feedback condition (M = 5.69, SD = .93), the graph condition (M = 5.94, SD = 1.12), and the text and graph condition (M = 6.00, SD = .87) were significantly more likely than those in the control condition (M = 2.15) to indicate that they had received feedback (F (4, 345) = 245.82, $p < .001$). Thus, it was clear that the feedback manipulation was salient to participants. Descriptive statistics and correlations among variables were calculated (Table 1). Descriptive statistics for dependent and moderator variables were calculated broken down by study condition (Table 2), and demographic variables were calculated by phase. As seen in Table 3, each sample provided a gender-balanced representation of working adults, which is a strength of the study and provides support for the generalizability of the results.

For hypothesis 1, linear regressions were used to assess the moderating influence of test performance on the effect of feedback on state test anxiety with a two-way interaction between feedback and test performance (Tables 4-6). Three separate regressions were conducted – one for each form of feedback (i.e., text, graphical, and text/graphical together).

Separate regressions were necessary because feedback mode was dummy coded. Each regression compared one mode of feedback to the control condition. Though the overall models were significant, the interactions were not. Thus, hypothesis 1 was not supported. As seen by the significant beta values, test performance was the main contributor to the significant overall models.

Hypothesis 2 was tested via linear regressions (Tables 7-12). Six three-way interactions between feedback, test performance, and performance goal orientation (three modes of feedback per type of goal orientation) were assessed. Of the six interactions tested, one was significant. The significant interaction occurred between prove performance goal orientation, test performance, and text and graphical feedback. Thus, hypothesis 2 was partially supported. The significant interaction is graphed in Figure 1. As seen in Figure 1, when text and graphical feedback were present, the magnitude of the differences in state test anxiety between higher and lower performers did not depend on their level of PPGO. When there was no text and graphical feedback present, the magnitude of the differences in state test anxiety between higher and lower performers was greater for participants who were higher in PPGO than lower in PPGO. This significant three-way interaction can be seen in Table 12. The main effects of both APGO and PPGO were the primary contributors to the significance of the overall models, as shown by their beta weights.

Linear regressions were also used to test hypothesis 3 (Tables 13-15). Hypothesis 3 assessed the moderating influence of test performance on the effect of feedback on perceived test fairness with a two-way interaction between feedback and test performance. Again, three

interactions were tested – one for each form of feedback. The interactions were not significant. Thus, hypothesis 3 was not supported.

Hypothesis 4 assessed the relationship between feedback acceptance and perceived test fairness among those who received feedback via a correlation. This correlation was significant ($r = .72$, $p < .001$). Thus, hypothesis 4 was supported.

Discussion

CAT systems have the capability to provide item-by-item feedback to examinees that contains more information than just performance-related information. In some cases, this is being done. However, little is understood about how item-by-item feedback can affect test takers. The purpose of this study was to assess the effect of various modes of IRT-specific item-by-item feedback on examinees' state test anxiety and perceptions of test fairness during a computer adaptive test in a high-stakes testing situation. Specifically, interactions were tested to assess whether feedback differentially affected examinees depending on their standing on other variables. Prior theory and research provided a sound basis for examining these effects, suggesting that certain variables, such as test performance and performance goal orientation, could determine how feedback affected examinees. In this study, however, there was little evidence to suggest that feedback mode had any effect on test takers, or that performance and performance goal orientation determined the nature of this effect. There was evidence to support the relationship between feedback acceptance and perceived test fairness. That is, those who viewed the feedback as accurate were more likely to perceive the entire test as fair. This could add value to Gilliland's taxonomy of organizational justice. Specifically, simply providing feedback to examinees may not be enough to affect

perceptions of fairness – one must evaluate the content and method of feedback to ensure that examinees will perceive it as accurate.

Upon further examination of the analyses, it is apparent that the primary determinants of the significance of the overall models that predicted state test anxiety were exam performance and both forms of performance goal orientation. Though this wasn't a primary avenue of research in this study, evidence suggests that, independent of any feedback received, poorer performers and those who were more performance-goal oriented were more likely to be anxious during the test. The fact that item-by-item feedback has no effect on the nature of this relationship speaks to the ambiguity of the effect that feedback has had on state test anxiety in prior research. This has been manifest by many differing and, at times, conflicting findings. There is evidence to suggest that moderators exist in this relationship, but based on the results of this research, it is unlikely that performance and performance goal orientation are them.

Regarding perceived fairness, this study's results appear to have mixed support for previous research. In general, providing feedback on a test has resulted in higher evaluations of fairness. In this study, in only one of three conditions (i.e., the text and graphical feedback condition) did the presence of feedback relate to higher perceived test fairness. Perhaps due to the fact that there was simply more feedback present, and thus more detail and clarity, examinees were more confident in the fairness of their evaluation in this condition.

Though there was some support for this research, the overall narrative is that feedback did not have a notable effect on test takers. This could have occurred for a number of reasons. Primarily, it is possible that, contrary to the theory behind this research, feedback

served to *distract* participants from their current versus desired performance, rather than heightening their awareness to it. A basic assumption of human performance theory (HPT) is the idea of limited cognitive resources—that when people complete tasks, they draw from a finite amount of cognitive resources (Draycott & Kline, 1996). In this study, requiring participants to pay attention to many sources of feedback that displays information with which they may not have been familiar prior to this study, coupled with the task of completing an ability test, could have required the use of too many cognitive resources for the feedback to truly be effective. To assess whether this may have been the case, a post hoc exploratory ANOVA comparing performance on the CAT for each feedback group was conducted. This analysis revealed that significant differences existed between groups (F (3, 335) = 4.33, $p < .01$). Post hoc comparisons showed that participants in the control condition (M = 0.38, SD = 1.01) scored significantly higher on the CAT than the participants in the text and graphical condition (M = -0.11, SD = 0.80). Though this alone is insufficient evidence for a deleterious effect of feedback, it raises the possibility that the cognitive resources spent analyzing the item-by-item feedback hindered participants' attention to the ability test.

Practical Implications

These results could have implications for researchers who explore the effects of feedback and designers of high-stakes tests. Though there is prior evidence that suggests moderators exist in the relationships between feedback and state test anxiety and feedback and perceived test fairness, the there is no significant evidence in this study that suggests that they are test performance and performance goal orientation. Furthermore, this research indicates that, except for the effect of text and graphical feedback on perceived test fairness,

no main effect of feedback existed. Thus, when designing high-stakes tests and considering examinees' state test anxiety and perceptions of test fairness, IRT-specific item-by-item feedback does not appear to be an effective intervention.

The follow-up tests of the effects of feedback on performance yielded interesting results – namely, that item-by-item feedback presented in graph and text format simultaneously can hinder examinees' performance. Though additional research is needed, these results indicate that examinees who receive item-by-item feedback could be disadvantaged relative to those who do not receive feedback. This is an important matter, considering the criticality of whether exams function similarly across examinees. Prior research has revealed mixed results about the effect of feedback on performance (Kluger & DeNisi, 1996). In this case, it is unclear whether the deleterious effect of feedback was related to the content, complexity, or timing of the feedback. However, test makers and administrators who provide item-by-item feedback to examinees should consider the results of this study.

Limitations and Future Directions

This study was not without limitations. First, due to the complexity of the test, the study limited participation to those who currently reside in the United States and speak English as a primary language. As research has shown that people from different cultures can react differently to performance feedback (Earley, Gibson, & Chen, 1999), future research should explore the extent to which these results remain constant across various cultures.

This study also limited the modes of feedback to text and graphical feedback. However, research has indicated that feedback can be given in many other forms, including

verbal and via video (Alvero et al., 2008). Thus, future studies should assess the extent to which other modes of feedback are effective during testing. Third, though state test anxiety has typically been measured at the end of exams, it could be beneficial to assess anxiety levels on an item-by-item basis. This may serve to provide a more detailed picture of state test anxiety throughout the testing process. This could take the form of a mobile application or in-text Likert-type rating scales. Finally, this study heavily relied on an online unproctored ability test. Online exam security has proven to be a contentious issue, and research has demonstrated that examinees may cheat on unproctored online exams (Harmon, Lambrinos & Kennedy, 2008). Though steps were taken to deter participants from cheating, there was no definitive way of knowing whether participants utilized other resources during the test.

Conclusion

Among the decisions to be made when administering a high-stakes examination are 1) whether to give feedback to examinees, 2) when to give it to them, and 3) what information to include in the feedback. Though many of the hypothesized interactions were not significant in this research study, the dialogue about the effect of feedback should continue. Continuing to explore the effect of aspects of the testing situation can only serve to improve the testing experience for examinees and administrators.

REFERENCES

Akca, F. (2011). The relationship between test anxiety and learned helplessness. *Social Behavior and Personality: An International Journal, 39*, 101-111.

Alam, M. M. (2013). A study of test anxiety, self-esteem, and academic performance among adolescents. *IUP Journal of Organizational Behavior, 12*, 33-43.

Alvero, A. M., Bucklin, B. R., & Austin, J. (2008). An objective review of the effectiveness and essential characteristics of performance feedback in organizational settings (1985-1998). *Journal of Organizational Behavior Management, 21*, 3-29.

Arps, G. F. (1920). Work with knowledge of results versus work without knowledge of results. *Psychological Monographs, 28*, 1-41.

Azzam, T., & Jacobson, M. R. (2013). Finding a comparison group: Is online crowdsourcing a viable option? *American Journal of Evaluation, 34*, 372-384.

Balcazar, F., Hopkins, B. L., & Suarez, Y. (1985). A critical, objective review of performance feedback. *Journal of Organizational Behavior Management, 7*, 65-89.

Bauer, T. N., Maertz, Jr., C. P., Dolen, M. R., & Campion, M. A. (1998). Longitudinal assessment of applicant reactions to employment testing and test outcome feedback. *Journal of Applied Psychology, 83*, 892-903.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods, 43*, 800-813.

Bierhoff, H. W., Cohen, R. L., & Greenberg, J. (1986). *Justice in social relations.* New York: Plenum.

Bies, R. J., & Moag, J. F. (1986). Interactional justice: Communication criteria of fairness. In R. J. Lewicki, B. H. Sheppard, & M. H. Bazerman (Eds.), *Research on negotiations in organizations* (Vol. 1, pp. 43–55). Greenwich, CT: JAI Press.

Bradshaw, G. D., & Gaudry, E. (1968). The effects of a single experience of success or failure on test anxiety. *Australian Journal of Psychology, 20*, 219-223.

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple choice tests. *Journal of Experimental Psychology, 13*, 273-281.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior, 29*, 2156-2160.

Chan, D., Schmitt, N., Jennings, D., Clause, C. S., & Delbridge, K. (1998). Applicant perceptions of test fairness: Integrating justice and self-serving bias perspectives. *International Journal of Selection and Assessment, 6*, 232-239.

Clark, J. W., Fox, P. A., & Schneider, H. G. (1998). Feedback, test anxiety, and performance in a college course. *Psychological Reports, 82*, 203-208.

Cohen, M., Ben-Zur, H., & Rosenfield, M. J. (2008). Sense of coherence, coping strategies, and test anxiety as predictors of test performance among college students. *International Journal of Stress Management, 15*, 289-303.

DiBattista, D., & Gosse, L. (2006). Test anxiety and the immediate feedback assessment technique. *The Journal of Experimental Education, 74*, 311-327.

Dorans, N. J., & Holland, P. W. (2000). *Population invariance and the equatability of tests: Basic theory and the linear case* (RR-00-19). Washington, D. C.: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-00-19-Dorans.pdf

Draycott, S. G., & Kline, P. (1996). Validation of the AGARD STRES Battery of Performance Tests. *Human Factors, 38*, 347-361.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41*, 1040-1048.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95*, 256-273.

Earley, P. C., Gibson, C. B., & Chen, C. C. (1999) "How did I do?" versus "How did we do?": Cultural contrasts of performance feedback use and self-efficacy. *Journal of Cross-Cultural Psychology, 30*, 594-619.

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

Folger, R., & Greenberg, J. (1985). Procedural justice: An interpretive analysis of personnel systems. *Research in Personnel and Human Resources Management, 3*, 141-183.

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review, 18*, 694-734.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of mechanical Turk samples. *Journal of Behavioral Decision Making, 26*, 213-224.

Greenberg, J. (2011). Organizational justice: The dynamics of fairness in the workplace. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology, Vol 3: Maintaining, expanding, and contracting the organization* (pp. 271-327). Washington, DC US: American Psychological Association. doi:10.1037/12171-008

Harmon, O. R., Lambrinos, J., & Kennedy, P. (2008). Are online exams an invitation to cheat? *Journal of Economic Education, 39*, 116-125.

Harper, A. (2008). Psychometric tests are now a multi-million-pound business: What lies behind a coach's decision to use them? *International Journal of Evidence Based Coaching and Mentoring, 2*, 40-51.

Hansen, J. B. (1974). Effects of feedback, learner control, and cognitive abilities on state anxiety and performance in a computer-assisted instruction task. *Journal of Educational Psychology, 66*, 247-254.

Hong, E. (1998). Differential stability of individual differences in state and trait test anxiety. *Learning and Individual Differences, 10*, 51-69.

Hong, E. & Karstensson, L. (2002). Antecedents of state test anxiety. *Contemporary Educational Psychology, 27*, 348-367.

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback seeking on behavior in organizations. *Journal of Applied Psychology, 64*, 349-371.

In'nami, Y. (2006). The effects of test anxiety on listening test performance. *System: An International Journal of Educational Technology and Applied Linguistics, 34*, 317-340.

Judd, C. H. (1905). Practice without knowledge of results. *Psychological Review Monographs, 2*, 185-198.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254-284.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86*, 148-161.

Leventhal, G. S. (1980). What should be done with equity theory? New approaches to the study of fairness in social relationships. In K. J. Gergen, M. S. Greenberg, & R. H. Willis (Eds.), *Social Exchange: Advances in Theory and Research:* 27-55. New York, Plenum.

Lounsbury, J. W., Bobrow, W., & Jensen, J. B. (1989). Attitudes toward employment testing: Scale development, correlates, and "known-group" validation. *Professional Psychology: Research and Practice, 20*, 340-349.

Nease, A. A., Mudgett, B. O., & Quiñones, M. A. (1999). Relationships among feedback sign, self-efficacy, and acceptance of performance feedback. *Journal of Applied Psychology, 84*, 806-814.

Payne, S. C., Youngcourt, S. S., & Beaubien, J. M. (2007). A meta-analytic examination of the goal orientation nomological net. *Journal of Applied Psychology, 92*, 128-150.

Putwain, D. W., Woods, K. A., & Symes, W. (2010). Personal and situational predictors of test anxiety of students in post-compulsory education. *British Journal of Educational Psychology, 80*, 137-160.

Sarason, I. G., & Sarason, B. R. (1990). Test anxiety. In H. Leitenberg (Ed.), *Handbook of social and evaluative anxiety* (pp. 475-496). New York: Plenum.

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

Study Mode, LLC (2014). *Majortests.com Miller Analogies Test Prep*. Retrieved from http://www.majortests.com/mat/miller-analogies-test-practice.php.

Tonidandel, S., Quiñones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology, 87*, 320-332.

University of Cambridge (2014). Concerto platform for the development of on-line adaptive tests. Retrieved from http://www.psychometrics.cam.ac.uk/newconcerto.

Vandewalle, D. (1997). Development and validation of a work domain goal orientation instrument. *Educational and Psychological Measurement, 57*, 995-1015.

van Vianen, A. E. M., Taris, R., Scholten, E., & Schinkel, S. (2004). Perceived fairness in personnel selection: Determinants and outcomes in different stages of the assessment procedure. *International Journal of Selection and Assessment, 12*, 149-159.

von der Embse, N., Barterian, J., & Segool, N. (2013). Test anxiety interventions for children and adolescents: A systematic review of treatment studies from 2000-2010. *Psychology in the Schools, 50*, 57-71.

Wise, S. L., Plake, B. S., Eastman, L. A., Boettcher, L. L., & Lukin, M. E. (1986). The effects of item feedback and examinee control on test performance and anxiety in a computer-administered test. *Computers in Human Behavior, 2*, 21-29.

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

Table 1

Descriptive Statistics and Correlations among Study Variables

Variable	M	*SD*	1	2	3	4	5	6
1. General Reasoning	0.13	0.90	-					
2. PPGO	4.51	1.24	-0.04	(.81)				
3. APGO	3.79	1.43	-0.07	0.39**	(.87)			
4. Feedback Acceptance[a]	4.75	1.23	-0.07	0.14*	0.03	(.87)		
5. State Test Anxiety	3.29	1.45	-0.21**	0.42**	0.51**	0.09	(.93)	
6. Perceived Test Fairness	4.50	1.07	0.06	0.02	-0.13*	0.72**	-0.16**	(.84)

N = 358, **p* < .05, ***p* < .01, coefficient alpha reliabilities on diagonal. PPGO = Prove Performance Goal Orientation, APGO = Avoid Performance Goal Orientation. [a] *N* = 258 (only assessed among participants receiving feedback).

Table 2

Means and Standard Deviations of Moderator and Outcome Variables by Study Condition

Variable	Condition			
	Control[a]	Text[b]	Graph[c]	Text + Graph[d]
General Reasoning	0.39 (1.01)	0.09 (0.79)	0.13 (0.94)	-0.11 (0.80)
State Test Anxiety	3.14 (1.45)	3.46 (1.51)	3.13 (1.43)	3.40 (1.43)
PPGO	4.38 (1.33)	4.34 (1.34)	4.76 (1.15)	4.52 (1.12)
APGO	3.78 (1.47)	3.67 (1.45)	3.88 (1.32)	3.80 (1.45)
Feedback Acceptance	NA	4.51 (1.31)	4.77 (1.25)	4.94 (1.09)
Perceived Test Fairness	4.32 (1.09)	4.44 (1.10)	4.51 (1.09)	4.70 (1.00)

Note. [a]N = 92. [b]N = 85. [c]N = 84. [d]N = 86.

Table 3

Demographic Variables by Phase

Variable	Phase 1	Phase 2
Average age (*SD*)	38.01 (12.54)	34.55 (11.48)
Gender		
Male	140 (47.3%)	196 (54.7%)
Female	155 (52.4%)	162 (45.3%)
Ethnicity		
African American	20 (6.8%)	22 (6.1%)
Asian American	15 (5.1%)	42 (11.7%)
Caucasian	245 (82.8%)	271 (75.7%)
Hispanic	11 (3.7%)	11 (3.1%)
Native American	2 (0.7%)	6 (1.7%)
Other	3 (1.0%)	6 (1.7%)
Average hours worked per week (*SD*)	27.36 (18.51)	25.53 (18.13)
Education		
Some high school	2 (0.7%)	1 (0.3%)
High school	40 (13.5%)	39 (10.9%)
Some college	103 (34.8%)	117 (32.7%)
College degree	115 (38.9%)	156 (43.6%)
Master's	30 (10.1%)	42 (11.7%)
Doctorate	6 (2.0%)	3 (0.8%)

Table 4

Regression Analysis Predicting State Test Anxiety from Text Feedback and Test Performance

Variable	*B*	*SE B*	*β*
Text Feedback	0.20	0.18	0.06
Test Performance	-0.35	0.10	-0.22**
Feedback x Performance	0.09	0.22	0.02
R^2	0.05		
F	5.66**		

**$p < .01$.

Table 5

Regression Analysis Predicting State Test Anxiety from Graphical Feedback and Test Performance

Variable	*B*	*SE B*	*β*
Graphical Feedback	-0.23	0.18	-0.07
Test Performance	-0.26	0.10	-0.16**
Feedback x Performance	-0.28	0.19	-0.09
R^2	0.06		
F	6.61**		

**$p < .01$.

Table 6

Regression Analysis Predicting State Test Anxiety from Text and Graphical Feedback and Test Performance

Variable	*B*	*SE B*	*β*
Text+Graph Feedback	0.13	0.18	0.04
Test Performance	-0.35	0.10	-0.22**
Feedback x Performance	0.11	0.22	0.03
R^2	0.05		
F	5.36**		

**$p < .01$.

Table 7

Regression Analysis Predicting State Test Anxiety from Text Feedback, Test Performance, and Avoid Performance Goal Orientation

Variable	*B*	*SE B*	*β*
Text Feedback	0.50	0.43	0.15
Test Performance	0.45	0.23	0.28
APGO	0.56	0.06	0.54**
Feedback x Performance	-0.69	0.55	-0.19
Feedback x APGO	-0.05	0.11	-0.06
Performance x APGO	-0.20	0.06	-0.49**
Feedback x Performance x APGO	0.21	0.14	0.22
R^2	0.33		
F	23.03**		

**$p < .01$. APGO = Avoid Performance Goal Orientation

Table 8

Regression Analysis Predicting State Test Anxiety from Graphical Feedback, Test Performance, and Avoid Performance Goal Orientation

Variable	*B*	*SE B*	*β*
Graphical Feedback	-0.33	0.48	-0.10
Test Performance	0.19	0.25	0.12
APGO	0.54	0.05	0.53**
Feedback x Performance	0.52	0.46	0.17
Feedback x APGO	-0.00	0.12	-0.00
Performance x APGO	-0.11	0.06	-0.28
Feedback x Performance x APGO	-0.19	0.11	-0.24
R^2	0.33		
F	23.56**		

**$p < .01$. APGO = Avoid Performance Goal Orientation

Table 9

Regression Analysis Predicting State Test Anxiety from Text and Graphical Feedback, Test Performance, and Avoid Performance Goal Orientation

Variable	*B*	*SE B*	*β*
Text+Graphical Feedback	-0.18	0.45	-0.05
Test Performance	0.28	0.23	0.18
APGO	0.52	0.06	0.51**
Feedback x Performance	0.27	0.56	0.07
Feedback x APGO	0.07	0.11	0.09
Performance x APGO	-0.15	0.06	-0.39**
Feedback x Performance x APGO	-0.03	0.13	-0.04
R^2	0.32		
F	21.94**		

**$p < .01$. APGO = Avoid Performance Goal Orientation

Table 10

Regression Analysis Predicting State Test Anxiety from Text Feedback, Test Performance, and Prove Performance Goal Orientation

Variable	*B*	*SE B*	*β*
Text Feedback	0.38	0.57	0.11
Test Performance	0.14	0.32	0.09
PPGO	0.50	0.07	0.43**
Feedback x Performance	-0.13	0.80	-0.03
Feedback x PPGO	-0.02	0.12	-0.03
Performance x PPGO	-0.10	0.07	-0.31
Feedback x Performance x PPGO	0.06	0.17	0.08
R^2	0.23		
F	14.03**		

**$p < .01$. PPGO = Prove Performance Goal Orientation

Table 11

Regression Analysis Predicting State Test Anxiety from Graphical Feedback, Test Performance, and Prove Performance Goal Orientation

Variable	*B*	*SE B*	*β*
Graphical Feedback	0.50	0.68	0.15
Test Performance	0.11	0.32	0.07
PPGO	0.54	0.06	0.46**
Feedback x Performance	0.04	0.80	0.01
Feedback x PPGO	-0.20	0.14	-0.28
Performance x PPGO	-0.08	0.07	-0.24
Feedback x Performance x PPGO	-0.05	0.16	-0.08
R^2	0.24		
F	15.13**		

**$p < .01$. PPGO = Prove Performance Goal Orientation

Table 12

Regression Analysis Predicting State Test Anxiety from Text and Graphical Feedback, Test Performance, and Prove Performance Goal Orientation

Variable	*B*	*SE B*	*β*
Text+Graphical Feedback	-0.68	0.65	-0.20
Test Performance	0.41	0.33	0.25
PPGO	0.46	0.06	0.39**
Feedback x Performance	-1.30	0.72	-0.35
Feedback x PPGO	0.18	0.14	0.25
Performance x PPGO	-0.16	0.07	-0.47*
Feedback x Performance x PPGO	0.30	0.15	0.39*
R^2	0.23		
F	14.32**		

*$p < .05$, **$p < .01$. PPGO = Prove Performance Goal Orientation

Table 13

Regression Analysis Predicting Perceived Test Fairness from Text Feedback and Test Performance

Variable	*B*	*SE B*	*β*
Text Feedback	-0.09	0.14	-0.04
Test Performance	0.03	0.07	0.02
Feedback x Performance	0.22	0.17	0.08
R^2	0.01		
F	1.10		

Table 14

Regression Analysis Predicting Perceived Test Fairness from Graphical Feedback and Test Performance

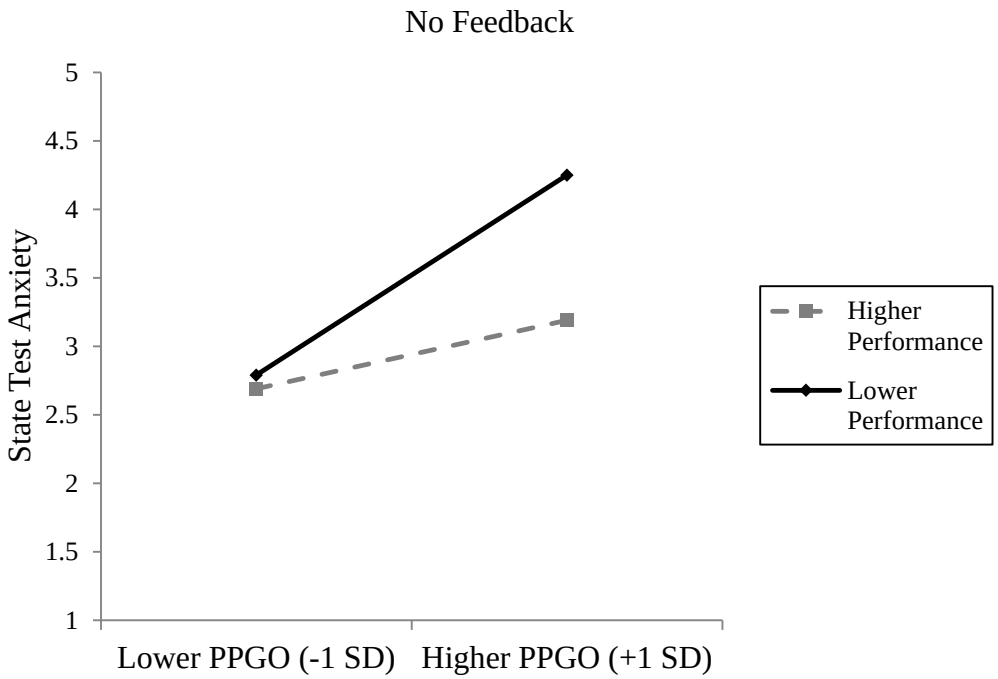
Variable	*B*	*SE B*	*β*
Graphical Feedback	0.04	0.14	0.02
Test Performance	0.11	0.08	0.09
Feedback x Performance	-0.15	0.15	-0.07
R^2	0.01		
F	0.76		

Table 15

Regression Analysis Predicting Perceived Test Fairness from Text and Graphical Feedback and Test Performance

Variable	*B*	*SE B*	*β*
Text+Graph Feedback	0.32	0.14	0.13*
Test Performance	0.05	0.07	0.05
Feedback x Performance	0.20	0.16	0.08
R^2	0.02		
F	2.67*		

*$p < .05$



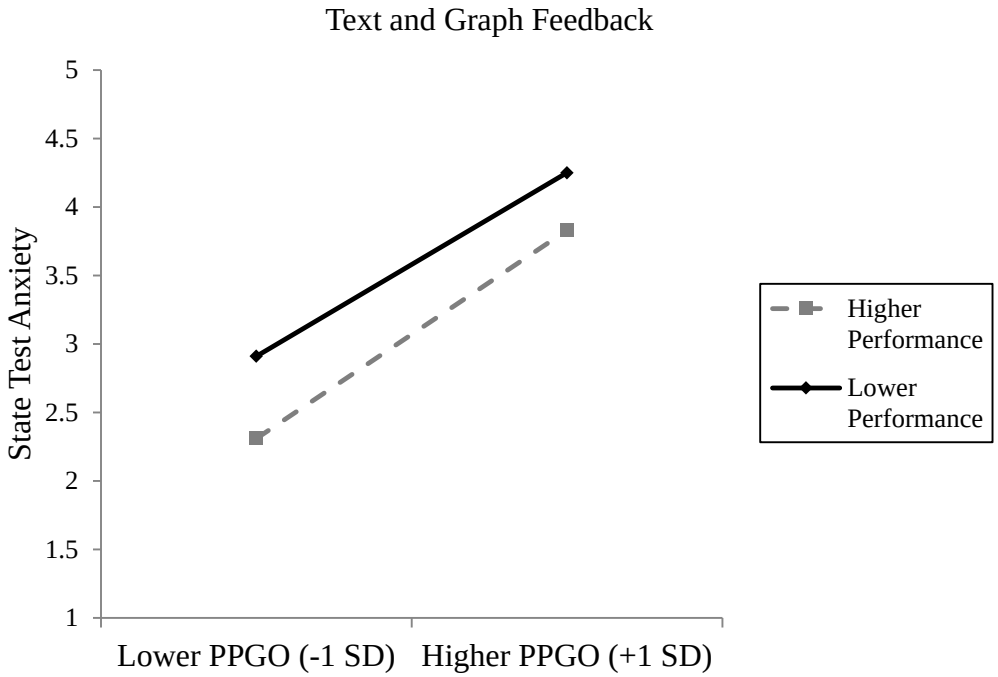


Figure 1. Three-way interaction between feedback (text and graphical), performance, and prove performance goal orientation (PPGO).

APPENDICES

Appendix A

Depictions of Feedback Displays

Text Feedback

DIVIDEND : STOCKHOLDER :: (____) : AUTHOR

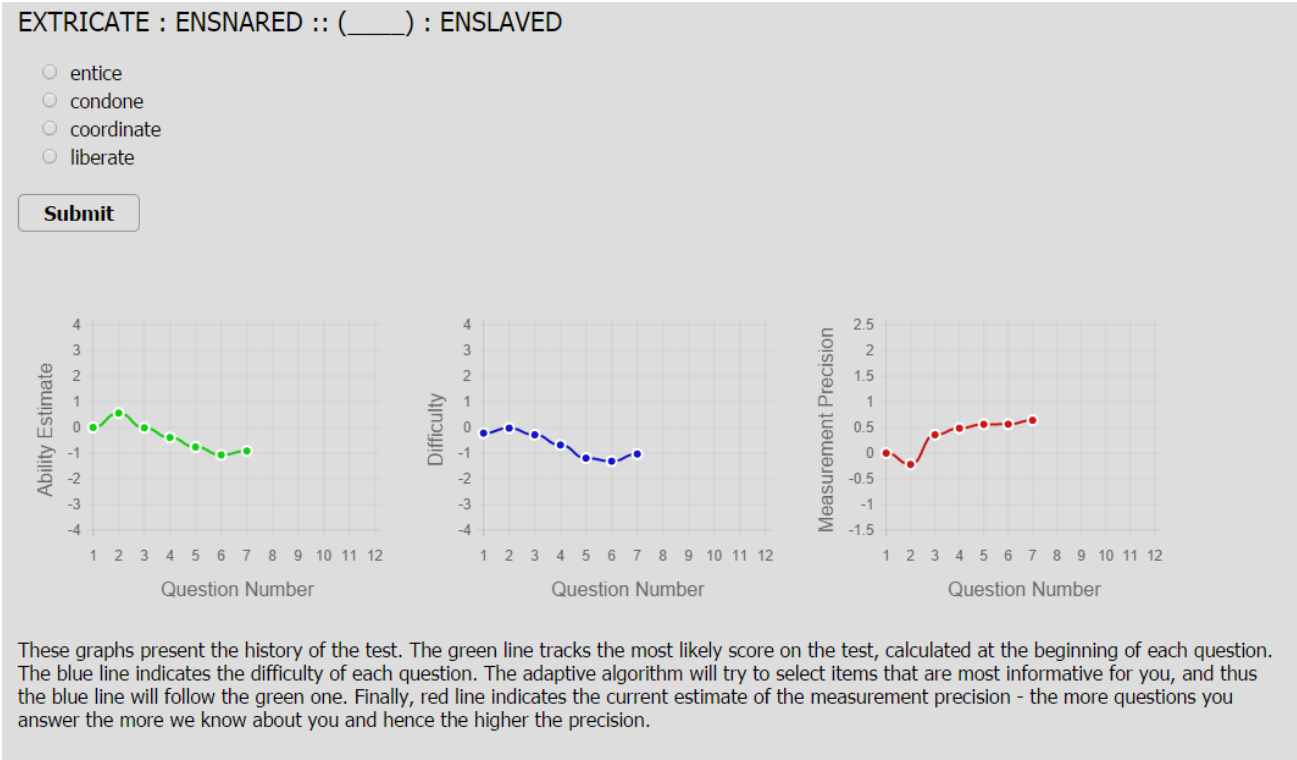
- patent
- royalty
- wage
- interest

Submit

Item #	Ability Estimate	Measurement Precision	Difficulty
5	-1.069	0.563	-1.315
4	-0.761	0.561	-1.195
3	-0.392	0.484	-0.683
2	-0.015	0.357	-0.286
1	0.56	-0.22	-0.028

The feedback in the top right table presents the history of the test. The ability estimate tracks the most likely score on the test, calculated at the beginning of each question. The difficulty denotes how hard each question is. The adaptive algorithm will try to select items that are most informative for you, and thus the difficulty will follow the ability estimate. Finally, the measurement precision indicates the current estimate of the how precise your score calculation is - the more questions you answer, the more we know about you and hence the higher the precision.

Graphical Feedback



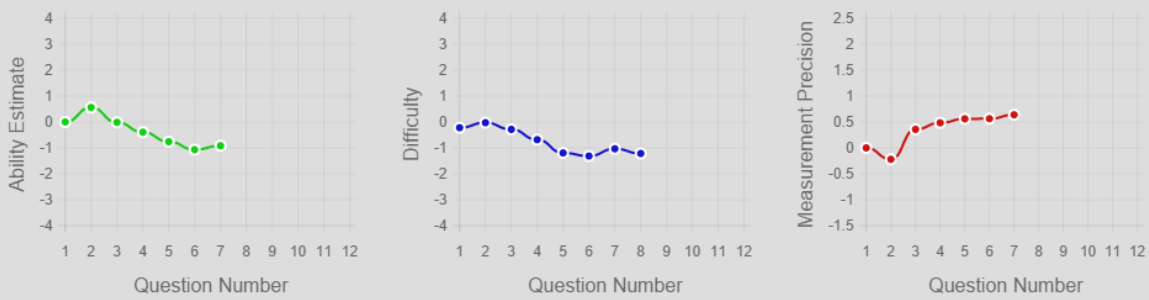
Text and Graphical Feedback

BOAST : LANGUAGE :: SWAGGER : (____)

- anger
- gait
- sight
- wealth

Submit

Item #	Ability Estimate	Measurement Precision	Difficulty
6	-0.917	0.639	-1.035
5	-1.069	0.563	-1.315
4	-0.761	0.561	-1.195
3	-0.392	0.484	-0.683
2	-0.015	0.357	-0.286
1	0.56	-0.22	-0.028



These graphs present the history of the test. The green line tracks the most likely score on the test, calculated at the beginning of each question. The blue line indicates the difficulty of each question. The adaptive algorithm will try to select items that are most informative for you, and thus the blue line will follow the green one. Finally, red line indicates the current estimate of the measurement precision - the more questions you answer the more we know about you and hence the higher the precision.

Appendix B

DISSERTATION PROPOSAL:

ABSTRACT

GASPERSON, SEAN MORGAN. The effects of item-by-item feedback during a computer adaptive test. (Under the direction of Lori Foster Thompson.)

Many factors can affect test takers in high-stakes testing environments. Test design is one such factor. Decisions involved in designing tests include whether to give examinees feedback and, if it is given, how to display the feedback to them. Previous research has assessed the effects of feedback, however there is a dearth of research dedicated to assessing feedback's effects on state test anxiety and examinees' perceptions of the test's fairness, and the conditions under which these effects are most pronounced. Furthermore, researchers have yet to examine these effects when feedback is given on an item-by-item basis – such as during a computer adaptive test. To address these gaps in the literature, the proposed study will examine the effects of feedback on state test anxiety and perceived test fairness, accounting for individual differences in feedback acceptance, exam performance, and performance goal orientation. These effects will be examined across three modes of item-by-item feedback (i.e., text feedback, graphical feedback, and text/graphical feedback together) during a computer adaptive test.

The effects of item-by-item feedback during a computer adaptive test

by
Sean Morgan Gasperson

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Psychology

Raleigh, North Carolina

2014

APPROVED BY:

Dr. Adam W. Meade	Dr. Mark A. Wilson	Dr. S. Bartholomew Craig

Dr. Lori Foster Thompson
Committee Chair

LIST OF FIGURES

LIST OF TABLES

Introduction

Many factors can affect test takers in high-stakes testing environments such as pre-employment testing or university entrance exams. Each year, decision makers and psychometricians spend a substantial amount of money selecting and developing tests for administration to applicants (Harper, 2008). Test design is vital to consider due to the potential ramifications of (a) examinees responding to test characteristics in different ways, and (b) examinees having adverse reactions to aspects of the test. Among the decisions involved in designing a test are whether to give examinees feedback, when to give feedback to them, and how to give feedback to them. These decisions could affect examinees' anxiety levels during the test as well as how fair they perceive the test to be. However, there is a dearth of literature devoted to this topic, particularly in the context of computer adaptive testing. The purpose of this study is to assess the effect of various modes of item-by-item feedback on examinees' state test anxiety and perceptions of test fairness during a computer adaptive test in a high-stakes testing situation.

Test Feedback

Research exploring the effects of giving feedback to test takers dates back to the early 1900s (e.g., Arps, 1920; Judd, 1905). Much of the early research focused on "knowledge of results," or providing test takers information about the outcome of the test (Kluger & DeNisi, 1996). More recently, the definition of feedback interventions (FIs) has been expanded to encompass any relevant information about the performance of a task, including specific feedback about *how* one performs the task. Kluger and DeNisi (1996) defined FIs as "actions taken by (an) external agent(s) to provide information regarding some aspect(s) of one's task

performance" (p. 255). Giving examinees feedback typically refers to providing them with some sort of report of their performance at the end of the test. However, researchers have explored the extent to which more complex characteristics of the feedback (e.g., content, frequency, mode, and source) affect outcomes such as performance on a particular task or job (for reviews, see Alvero, Bucklin, & Austin, 2008 and Balcazar, Hopkins, & Suarez, 1985).

Two dimensions on which feedback can vary are the timing of the feedback (i.e., when it is administered) and the mode of feedback (i.e., how the feedback is displayed to the test taker). Feedback is typically given at the end of the test or later, though some research has been devoted to assessing the effect of giving feedback after examinees complete each item (e.g., Butler, Karpicke, & Roediger, 2007; Clark, Fox, & Schneider, 1998). This research, though practical, is lacking in two ways. First, the content of the feedback is limited to whether examinees answered the item correctly or incorrectly. Second, the research has been mainly focused on examinees' preference for the timing of feedback and how delayed versus immediate feedback affects performance on the task (e.g., Butler et al., 2007). Other outcomes, such as examinees' test anxiety and perceptions of the test's fairness, should be explored.

Feedback can also vary by how it is given or displayed to the examinees. In conducting a literature review on the effect of various forms of feedback, Alvero et al. (2008) identified four primary methods of providing feedback: text, verbal, graphical, and mechanical (e.g., via video). Comparing text and graphical feedback, 50% of the studies that reported providing text feedback alone reported that the feedback was effective (i.e., the feedback produced the predicted effect), as did 50% of the studies assessing graphical

feedback effects. Text feedback that was accompanied by graphical feedback, however, was effective in 86% of the studies. Thus, research suggests that coupling text feedback with graphical feedback is more effective. Like feedback timing, however, research on feedback mode has primarily focused on performance- and satisfaction-based outcomes.

Computer adaptive testing and item response theory (IRT) provide a valuable platform to address shortcomings in the feedback literature. When examinees take computer adaptive tests, questions of varying difficulty levels are selected from a test bank based on the examinees' estimated level of an underlying trait (i.e., theta). The exam ends when a predetermined criterion is met, such as an examinee receiving a maximum number of items or when the standard error of their estimated theta reaches a predetermined value. Whereas traditional paper-based tests or non-adaptive computer-based tests do not have the capability to report real-time ability estimates, computer adaptive tests do. For example, one of the basic computer adaptive test demonstrations on the Concerto testing system developed by The Psychometrics Centre at the University of Cambridge displays IRT-based feedback graphically to test takers (University of Cambridge, 2014). This feedback includes theta estimates, item difficulty levels, test information, and item characteristic curves, and is a noted advantage of using the testing system. Reporting some or all of this information to examinees during the test is an option for test administrators who employ computer adaptive testing. Empirical information about the effects of this type of feedback in a high-stakes testing situation is lacking, and is needed to guide decisions about how the portrayal of this information affects test takers, particularly as related to their levels of anxiety and perceived

fairness. Furthermore, research has yet to examine whether known text and/or graphical feedback effects extend to tests that provide said feedback on an item-by-item basis.

State Test Anxiety

Test anxiety has been operationalized as emotional, physiological, and behavioral responses to the potential consequences of negative outcomes on a test or exam (Zeidner, 1998). Test anxiety is induced when test takers perceive that the demands of the test situation exceed their intellectual, motivational, and/or social capabilities (Sarason & Sarason, 1990). Research on the topic has addressed a variety of concerns, including interventions for test anxiety (e.g., Ergene, 2003), personal and situational predictors of test anxiety (e.g., Putwain, Woods, & Symes, 2010), its relationship with various constructs such as learned helplessness (e.g., Akca, 2011), and perhaps most notably, its relationship with task performance (see Alam, 2013; Cohen, Ben-Zur, & Rosenfield, 2008; In'nami, 2006). Test anxiety can generally be divided into state test anxiety and trait test anxiety. State test anxiety refers to an anxious, temporary condition specific to a particular testing situation, while trait test anxiety refers to a more stable tendency to have anxiety across any testing situation (Hong, 1998).

Much research has been devoted to the effect of feedback on state test anxiety, often with the primary purpose of seeking to reduce test takers' anxiety (e.g., von der Embse, Barterian, & Segool, 2013). Feedback has been thought to reduce test anxiety by eliminating ambiguity that may be present about performance on the tasks at hand (Hansen, 1974). Furthermore, because test takers will, in general, answer more items correctly than incorrectly, there is more opportunity for positive feedback, thus reducing apprehension about one's performance (DiBattista & Gosse, 2006). In an effort to assess the viability of

item feedback in reducing state test anxiety, some researchers have found that contrary to what was hypothesized, feedback increased state test anxiety (Wise, Plake, Eastman, Boettcher, & Lukin, 1986). Conversely, other researchers have shown that providing feedback to participants has resulted in significant reductions in state test anxiety over time (e.g., Hansen, 1974). Other research has reported no relationship between immediate feedback and test anxiety (Clark et al., 1998; DiBattista & Gosse, 2006). Thus, though researchers have explored the relationship between feedback and state test anxiety, there appears to be little consensus as to what effect, if any, it has. These inconsistencies in research can be evidence of a moderating effect, such that feedback's effects on state test anxiety may be contingent on a third variable. How well examinees perform on the test could serve as this moderator. High and low performers on tests may react differently to feedback due to the differing amounts of positive reinforcement they receive. Though performance's moderating influence on feedback's effects has not been examined, the effect of providing end-of-exam performance reports on anxiety has. Bradshaw and Gaudry (1968) found that experiencing failure can result in higher state test anxiety, and experiencing success can result in lower state test anxiety. Participants were randomly assigned to fail or succeed on the test, and received performance feedback at the end of testing. Those who received feedback of poor performance reported higher test anxiety, and those who received feedback of high performance reported lower test anxiety. This research, however, did not assess whether higher and lower performers react differently to real-time feedback.

Immediate feedback (i.e., item-by-item) inherently gives examinees an indication of test performance. As individuals take a test, immediate feedback makes them aware of the

discrepancy (or lack thereof) between their current performance and their desired performance (Kluger & DeNisi, 1996). For better performers, this discrepancy will be lower. For poorer performers, it will be higher. Furthermore, feedback that gives examinees a more detailed item-by-item account of their performance and how their abilities are being estimated heightens examinees' awareness of that discrepancy. The typical alternative, giving examinees their score at the end of the test, does not have the capacity to influence state test anxiety throughout the testing situation because examinees may not be completely certain how many items they answered correctly or incorrectly. Thus, for poorer performers, detailed item-by-item feedback should lead to higher levels of state test anxiety. For better performers, detailed item-by-item feedback should lead to lower levels of test anxiety. Although some research has been devoted to assessing the main effect of feedback on state test anxiety (e.g., DiBattista & Gosse, 2006; Hansen, 1974), the moderating effect of performance on the relationship between feedback and state test anxiety during a test that provides ongoing item-level feedback throughout the test has not been examined.

> ***Hypothesis 1***: Test performance will moderate the effect of item-by-item feedback on state test anxiety, with feedback reducing anxiety for better performers and increasing anxiety for poorer performers.

The expected effects for Hypothesis 1 are displayed in Figure 1. As prior research has suggested that the way feedback information is presented to people can influence reactions to it (Alvero et al., 2008), this and all subsequent hypotheses will be tested across three modes of feedback: text feedback, graphical feedback, and text and graphical feedback together.

There may be types of individuals for whom the moderating effect of test performance is especially likely. In particular, examinees oriented toward performance goals may be particularly sensitive to feedback conveying performance failures and successes. As an overall construct, goal orientation refers to how people react to opportunities to demonstrate or develop ability (Dweck, 1986). Goal orientation can be divided into learning goal orientation (i.e., seeking to master and acquire new skills) and performance goal orientation (i.e., seeking opportunities to demonstrate or validate one's competence; Dweck & Leggett, 1988). Vandewalle (1997) further differentiated performance goal orientation into *prove* performance goal orientation and *avoid* performance goal orientation. Prove performance goal orientation is one's tendency to seek opportunities that may demonstrate competence, while avoid performance goal orientation is one's tendency to avoid situations that may disprove competence. In a meta-analysis examining the relationship of goal orientation dimensions and proximal and distal outcomes, Payne, Youngcourt, and Beaubien (2007) found that individuals with high levels of avoid performance goal orientation and prove performance goal orientation were likely to have high levels of state test anxiety. As item-by-item feedback provides immediate knowledge of results, individuals who are high in performance goal orientation (both avoid and prove) and receive feedback indicating that they are performing poorly on the test should exhibit especially high levels of state test anxiety.

Hypothesis 2*:* There will be a three-way interaction between feedback, test performance, and performance goal orientation on state test anxiety, such that performance's moderating influence on the effect of feedback is more pronounced for

those more oriented toward avoid performance (H2a) and prove performance (H2b) goals.

Figure 2 displays the expected effects.

Perceived Test Fairness

Moving beyond state test anxiety, feedback's influence over other outcome variables is also important to consider. Perceptions of a test's fairness, in particular, warrants consideration. Perceptions of test fairness can be understood by examining the organizational justice literature. Organizational justice has been a popular area of research throughout recent years. It refers to "people's perceptions of fairness in organizations along with their associated behavioral, cognitive, and emotional reactions" (Greenberg, 2011, p. 271). Three primary types of organizational justice have been identified – distributive justice, procedural justice, and interactional justice (Bierhoff, Cohen, & Greenberg, 1986; Bies & Moag, 1986; Folger & Greenberg, 1985). Procedural justice (i.e., the perceived fairness of the processes that are used when organizational decisions are made; Folger & Greenberg, 1985) is thought to be composed of ten "rules" which, when violated, lead to perceptions of injustice (Gilliland, 1993; Leventhal, 1980). As pre-employment testing is included in the organizational decision-making process (more specifically, the hiring process), much research about procedural justice focuses on perceptions of test fairness.

One of the procedural justice rules is feedback. Feedback can be given in various forms, including from others and from the test itself. Research has supported the notion that feedback during the testing process can affect perceived test fairness. Specifically, research has indicated that the mere presence of feedback during the testing process can result in

higher evaluations of fairness (Lounsbury, Bobrow, & Jensen, 1989). Furthermore, other research has shown a significant effect of feedback on post-feedback fairness perceptions (van Vianen, Taris, Scholten, & Schinkel, 2004). Feedback, when given throughout the test, provides information to examinees regarding how they are performing and how their final scores are calculated. Particular to computer adaptive testing, item-by-item feedback can facilitate examinees' understanding of how their scores are calculated by providing information about the scoring process. In essence, rather than simply providing a final score after testing, information such as real-time ability (i.e., theta) estimates, the precision with which the examinees' ability is being estimated (i.e., standard error of measurement), and the difficulty of the items can be given to the examinee, providing transparency and clarity about the methods of item response theory. This transparency is a defining characteristic of procedural justice – namely, facilitating examinees' knowledge about the how the tests are selected, constructed, and used (Gilliland, 1993).

Though feedback can be an important factor in the evaluation of fairness perceptions, there may be some conditions under which feedback is especially likely to raise fairness perceptions. Two potential moderators worth exploring are performance on the test and feedback acceptance. How well an examinee performs can affect the relationship between feedback and perceived test fairness. As previously noted, examinees who receive feedback are more likely to perceive the test as fair. Furthermore, research has demonstrated that better performers are more likely to perceive the test as fair (Bauer, Maertz, Dolen & Campion, 1998; Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993). Researchers have attributed this relationship to self-serving biases (Chan, Schmitt, Jennings, Clause, & Delbridge, 1998).

That is, poorer performing examinees are more likely than better performing examinees to evaluate a test as unfair in order to maintain a positive view of themselves by means of threat-reduction or ego-enhancement. As examinees receive detailed item-by-item feedback throughout the test, any ambiguity regarding their performance will be reduced. For better performers, item-by-item feedback should reinforce the higher perceptions of fairness that they hold about the test, while for poorer performers, item-by-item feedback should reinforce the lower perceptions of fairness that they hold about the test.

Though previous research has examined the main effects of feedback and test performance on perceived test fairness, there is a dearth of research exploring their interactive effects. It is possible that providing real-time information about the test and examinees' test performance, combined with how well one performs on the test, could affect perceptions of test fairness.

> ***Hypothesis 3****:* Test performance will moderate the degree to which item-by-item feedback affects perceived test fairness, such that feedback will increase perceived test fairness for better performers and decrease perceived test fairness for poorer performers.

Figure 3 displays the expected effects.

A second potential moderating variable worthy of examination is feedback acceptance. Some research suggests that the examinee must accept the feedback in order for it to be effective (e.g., Nease, Mudgett, & Quinones, 1999). Thus, the extent to which examinees perceive a test as fair may depend on whether they accept feedback regarding their performance on the test. Feedback acceptance refers to the extent to which the feedback

recipients believe that the feedback regarding their performance is accurate (Ilgen, Fisher, & Taylor, 1979). Research has previously examined the moderating effect of feedback acceptance on the relationship between feedback and outcomes such as satisfaction with the test and fairness perceptions. Tonidandel, Quiñones, and Adams (2002) found that when participants accepted test feedback and had accurate perceptions of their performance, there was a positive relationship between feedback and perceptions of test fairness. When participants did not accept test feedback and had inaccurate perceptions of their performance, there was no relationship between feedback and fairness perceptions. Thus, it appears that examinees may evaluate whether the performance feedback aligns with their perceptions of their own performance when deciding whether to accept the feedback. Though previous research (e.g., Tonidandel et al., 2002) provides some evidence of the moderating effect of feedback acceptance on the relationship between feedback and perceived test fairness, the feedback was only given at the end of the test. While end-of-test feedback is informative, research has not provided a thorough examination of real-time feedback as it relates to perceived test fairness. Providing item-by-item feedback of test performance could serve to reduce the discrepancy between perceived and actual performance. How this affects examinees' perceptions of test fairness, however, should depend on the degree to which examinees accept the feedback.

> ***Hypothesis 4****:* Feedback acceptance will moderate the degree to which item-by-item feedback increases perceived test fairness, such that feedback will increase perceived test fairness for examinees with higher feedback acceptance and decrease perceived test fairness for examinees with lower feedback acceptance.

Figure 4 displays the expected effects.

Method

Participants

Participation for this study will be solicited via Amazon Mechanical Turk (MTurk). MTurk is a crowdsourcing website through which researchers can pay participants to complete various *Human Intelligence Tasks* (HITs). Though the data quality of MTurk has been questioned, research has consistently demonstrated its viability compared to other sources of data (Azzam & Jacobson, 2013; Behrend, Sharek, Meade, & Wiebe, 2011; Casler, Bickel, & Hackett, 2013; Goodman, Cryder, & Cheema, 2013).

As described below, there will be two phases of data collection. Participants for phase 1 will be compensated $0.50. Various demographic characteristics of this sample will be tracked and reported, including gender, age, ethnicity, and education level.

Participants for phase 2 will be compensated $0.75. Again, the sample's demographic characteristics will be measured and reported.

This study's sample will be restricted to individuals who speak English as a first language. This is due to the fact that some items in the test (i.e., the Miller Analogies Test practice items described later) require a nuanced understanding of the English language.

Design

This study will employ a between-groups design. Participants will be randomly assigned to one of four conditions that will vary according to the mode of feedback that they receive: control (no feedback), text feedback only, graphical feedback only, or text and graphical feedback. The mode of feedback will be the independent variable, with the four

levels described above. There will be two dependent variables – state test anxiety and perceived test fairness. Test performance, feedback acceptance, and goal orientation will be tested as moderators.

Procedure

This study will involve the administration of a high-stakes test measuring general reasoning ability. In particular, practice items for the Miller Analogies Test (MAT) will be used to measure general reasoning. The MAT has been used as an entrance exam for graduate programs as well as a pre-employment selection tool (Kuncel, Hezlett, & Ones, 2004). A meta-analysis exploring the utility of the MAT relative to other common exams and its relationship to certain outcomes revealed significant correlations with both forms of the Graduate Record Examination (i.e., verbal and quantitative), general ability and reasoning measures, graduate grade point average, and job performance (Kuncel et al., 2004).

In order to develop the computer adaptive test needed to test the study hypotheses, preliminary data will need to be collected to estimate the item parameters. This necessitates a "Phase 1" of data collection for exam development purposes prior to collecting "Phase 2" data for hypothesis testing. Phase 1 will involve giving participants one of two subsets of practice items for the MAT. The test will be divided into subsets due to the fact that there are 150 total MAT practice items. Administering all 150 items to each participant would take a substantial amount of time, and would introduce the possibility of test fatigue. Therefore, a common equating method will be used to avoid this problem. This will entail giving two smaller subsets to different groups of participants, with a few items that are common to both subsets. Each subset will contain a total of 85 items. Sixty five of these items will be unique

to each subset, while 20 items will be common across the two subsets. These common items will serve as the anchor items to link the two subsets. The item characteristics of the anchor items will be used to adjust the item characteristics of the unique items so that all 150 reasoning items will be placed on a common metric. The item parameters of all the reasoning items will be used to determine which items will be given throughout the subsequent computer adaptive test. To ensure that the content of the two subsets is as identical to the other as possible, the area of knowledge that each reasoning item measures will be labeled (e.g., arts, math, science), and the items will be divided such that each area of knowledge is represented equally across the two subsets. Procedurally, participants will first be presented with an informed consent form after volunteering to participate. Participants will answer items that assess various demographic variables such as age and gender, and will then proceed to complete a subset of the MAT items. Participants will be randomly assigned to receive one of the two subsets.

Before the item subsets can be linked, dimensionality and reliability must be assessed. It is inappropriate to equate two test forms that measure different constructs and that are not equally reliable (Dorans & Holland, 2000). Thus, an exploratory factor analysis (EFA) will be conducted on each item subset, and the factor structure of each test will be compared. Concurrently, the reliability of each measure will be assessed. Items that display low item total correlations and poor factor loadings will be discarded. The final items will be entered into the test bank for the CAT.

Phase 2 of data collection will be executed to test this study's hypotheses. Phase 2 will involve giving participants the computer adaptive test that contains the reasoning items,

followed by the measures of this study's dependent and moderator variables. Participants will be presented with an informed consent form after volunteering to participate. If they agree to participate in the study, they will be presented with instructions that evoke a high-stakes testing situation (see Appendix B). This will inform them that the three examinees who score the highest on the reasoning test will each receive a $25 Amazon.com gift card. Once data collection is complete, the gift cards will be distributed on the basis of participants' scores.

Participants will be randomly assigned to one of the four feedback conditions. Each condition will vary with regard to whether and how the feedback information is displayed on the screen after answering each item. The control condition will contain neither the text nor graphical feedback; that is, it will provide no information about examinees' estimated theta, etc. The three feedback conditions will display examinees' estimated theta, the standard error of their theta estimate, and the item difficulty will be displayed numerically after each item, accompanied by a brief description of how this information can be interpreted. These terms (e.g., theta estimate, item difficulty, etc.) will be described in a way that can be understood by laypersons, and will be shown on every screen throughout the test. For example, the following information will appear after answering each item in the three feedback conditions: "The standard error of measurement indicates the precision with which your score is being calculated. As this number decreases, we can be more confident about the accuracy of your score." The text-only feedback condition will display the numbers corresponding with these terms without any graphs. For the graphical feedback condition, a line graph displaying examinees' estimated theta, the standard error of their theta estimate, and the item difficulty will be displayed, accompanied with verbal explanations of these terms. Examinees can track

their progress and real time score by looking at the text and/or graph. The combined text and graphical feedback condition will contain all of the information that is displayed in the text and graphical conditions. The difference between the text and graphical (combined) condition and the graphical condition is that the graphical condition does not display numerical values of theta, standard error of measurement, and item difficulty, as the text condition does—only graphical depictions of them. An example of each condition can be found in Appendix C.

The CAT will begin with an item that corresponds to a theta level of zero. Each subsequent item will be selected by giving examinees an item that they have not previously taken that provides the most information at their current estimated theta level. The test will terminate when the standard error for the examinees' estimated theta reaches a predefined value. After the test has terminated, examinees will be directed to another web page where they will complete subsequent measures assessing the moderator and outcome measures of interest, except for feedback acceptance. Then, examinees will receive a report of their final score on the reasoning test. Scores will be estimated via the expected a posteriori (EAP) method. After final scores are given, examinees will complete the items measuring feedback acceptance (described below). Examinees will be instructed to consider "feedback" as any type of information about their test score that they have received at any point during or after the test administration.

Measures

Demographics (8 items). Eight items will be administered in Phases 1 and 2 of data collection to assess participants' age, gender, race, work experience, country of origin, native

language, country of residence, and education. These data will be collected to provide a description of the samples used in the study. See Appendix A for these and all other items.

General reasoning (150 items, α = X). One hundred and fifty practice items for the Miller Analogies Test (MAT) will form the basis for Phase 1 and 2 data collection. These items were developed to be used by individuals who are preparing to take the MAT (Study Mode, 2014). The MAT is a series of analogies which requires examinees to infer the relationship between two words and apply that relationship to another set of words. Examinees must select a word that completes the word pair that is most similar to the given word pair. An example item is "(____) : PUCCINI :: SCULPTURE : OPERA." The response options for this item are "A. Cellini, B. Rembrandt, C. Wagner, D. Petrarch."

The remaining items will be given only during Phase 2, and will all be presented with Likert-type response scales, with scores ranging from 1 (*strongly disagree*) to 7 (*strongly agree*).

State test anxiety (8 items, α = X). State test anxiety will be measured via a modified eight-item measure from Hong and Karstensson (2002). Four items measure state worry, and four items measure state emotionality. An example item of state worry is "During the test I got so nervous that I forgot facts I really knew." An example item of state emotionality is "During the test, I felt very tense." The scale was modified by removing context-specific language from one item (i.e., "in the course") and replacing it with more general language (i.e., "on the test").

Goal orientation (13 items). Vandewalle's (1997) goal orientation measure will be used. This measures prove performance goal orientation (4 items, α = X) and avoid

performance goal orientation (4 items, α = X). An example item for prove performance goal orientation is "I'm concerned with showing that I can perform better than my coworkers." An example item for avoid performance goal orientation is "I would avoid taking on a new task if there was a chance that I would appear rather incompetent to others."

Perceived test fairness (8 items, α = X). Perceptions of test fairness will be assessed via an eight-item measure adapted from Tonidandel et al. (2002). An example item is "Under the circumstances, the test was fair." The test was modified by removing context-specific language in one item (i.e., "knowledge of general psychology") and replacing it with more general language (i.e., "general reasoning").

Feedback acceptance (4 items, α = X). A four-item scale taken from Tonidandel et al. (2002) will measure feedback acceptance. An example item is "I do not believe that the feedback is accurate." In this context, "feedback" will be defined as any performance information that examinees have received throughout the testing scenario.

Manipulation check (4 items, α = X). Four items will be given to examinees to assess whether the experimental manipulations were salient to examinees. An example item is "Each time I answered a test question, I received feedback about how I was doing on the test."

Attention check (3 items). Three items will be used to evaluate examinees' careless responding. An example item is "Please answer strongly agree."

Proposed Analyses

Multiple linear regressions will be used to test all hypotheses. Feedback will be treated as categorical, thus it will be dummy coded so that the model intercepts can be

interpreted as the average level of the dependent variable for the control condition (i.e., no feedback). All other variables will be treated as continuous. Due to the fact that feedback is dummy coded, there will be three separate interactions tested for each of the hypotheses – one for each of the three feedback display mode conditions. Thus, for each hypothesis, the coefficients returned can be interpreted as the number of unit increases in the dependent variable, on average, corresponding to one unit increase in the moderating variable for that particular mode of feedback.

Hypothesis 1: In order to assess the moderating influence of test performance on the effect of feedback on state test anxiety, a two-way interaction between feedback and test performance will be conducted, with state test anxiety as the dependent variable (DV).

Hypothesis 2: In order to assess hypothesis 2, a three-way interaction between feedback, test performance, and performance goal orientation (prove performance goal orientation for H2a, avoid performance goal orientation for H2b) will be conducted, with state test anxiety as the DV.

Hypothesis 3: To assess test performance's moderating influence on the effect of feedback on perceived test fairness, a two-way interaction between feedback and test performance will be conducted, with perceived test fairness as the DV.

Hypothesis 4: To assess feedback acceptance's moderating influence on the effect of feedback on perceived test fairness, a two-way interaction between feedback and feedback acceptance will be conducted, with perceived test fairness as the DV.

Manipulation check: To assess whether the experimental manipulations (i.e., the item-by-item feedback conditions) were salient to examinees, an analysis of variance

(ANOVA) will be conducted to test group mean differences on the manipulation check items. Post hoc comparisons with Bonferroni corrections will be used. If the experimental group means are significantly greater than the control group mean, the manipulations can be deemed salient to examinees.

References

Akca, F. (2011). The relationship between test anxiety and learned helplessness. *Social Behavior and Personality: An International Journal, 39*, 101-111.

Alam, M. M. (2013). A study of test anxiety, self-esteem, and academic performance among adolescents. *IUP Journal of Organizational Behavior, 12*, 33-43.

Alvero, A. M., Bucklin, B. R., & Austin, J. (2008). An objective review of the effectiveness and essential characteristics of performance feedback in organizational settings (1985-1998). *Journal of Organizational Behavior Management, 21*, 3-29.

Arps, G. F. (1920). Work with knowledge of results versus work without knowledge of results. *Psychological Monographs, 28*, 1-41.

Azzam, T., & Jacobson, M. R. (2013). Finding a comparison group: Is online crowdsourcing a viable option? *American Journal of Evaluation, 34*, 372-384.

Balcazar, F., Hopkins, B. L., & Suarez, Y. (1985). A critical, objective review of performance feedback. *Journal of Organizational Behavior Management, 7*, 65-89.

Bauer, T. N., Maertz, Jr., C. P., Dolen, M. R., & Campion, M. A. (1998). Longitudinal assessment of applicant reactions to employment testing and test outcome feedback. *Journal of Applied Psychology, 83*, 892-903.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods, 43*, 800-813.

Bierhoff, H. W., Cohen, R. L., & Greenberg, J. (1986). *Justice in social relations.* New York: Plenum.

Bies, R. J., & Moag, J. F. (1986). Interactional justice: Communication criteria of fairness. In R. J. Lewicki, B. H. Sheppard, & M. H. Bazerman (Eds.), *Research on negotiations in organizations* (Vol. 1, pp. 43–55). Greenwich, CT: JAI Press.

Bradshaw, G. D., & Gaudry, E. (1968). The effects of a single experience of success or failure on test anxiety. *Australian Journal of Psychology, 20*, 219-223.

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple choice tests. *Journal of Experimental Psychology, 13*, 273-281.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior, 29*, 2156-2160.

Chan, D., Schmitt, N., Jennings, D., Clause, C. S., & Delbridge, K. (1998). Applicant perceptions of test fairness: Integrating justice and self-serving bias perspectives. *International Journal of Selection and Assessment, 6*, 232-239.

Clark, J. W., Fox, P. A., & Schneider, H. G. (1998). Feedback, test anxiety, and performance in a college course. *Psychological Reports, 82*, 203-208.

Cohen, M., Ben-Zur, H., & Rosenfield, M. J. (2008). Sense of coherence, coping strategies, and test anxiety as predictors of test performance among college students. *International Journal of Stress Management, 15*, 289-303.

DiBattista, D., & Gosse, L. (2006). Test anxiety and the immediate feedback assessment technique. *The Journal of Experimental Education, 74*, 311-327.

Dorans, N. J., & Holland, P. W. (2000). *Population invariance and the equatability of tests: Basic theory and the linear case* (RR-00-19). Washington, D. C.: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-00-19-Dorans.pdf

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41*, 1040-1048.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95*, 256-273.

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

Folger, R., & Greenberg, J. (1985). Procedural justice: An interpretive analysis of personnel systems. *Research in Personnel and Human Resources Management, 3*, 141-183.

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review, 18*, 694-734.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of mechanical Turk samples. *Journal of Behavioral Decision Making, 26*, 213-224.

Greenberg, J. (2011). Organizational justice: The dynamics of fairness in the workplace. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology, Vol 3: Maintaining, expanding, and contracting the organization* (pp. 271-327). Washington, DC US: American Psychological Association. doi:10.1037/12171-008

Harper, A. (2008). Psychometric tests are now a multi-million-pound business: What lies behind a coach's decision to use them? *International Journal of Evidence Based Coaching and Mentoring, 2*, 40-51.

Hansen, J. B. (1974). Effects of feedback, learner control, and cognitive abilities on state anxiety and performance in a computer-assisted instruction task. *Journal of Educational Psychology, 66*, 247-254.

Hong, E. (1998). Differential stability of individual differences in state and trait test anxiety. *Learning and Individual Differences, 10*, 51-69.

Hong, E. & Karstensson, L. (2002). Antecedents of state test anxiety. *Contemporary Educational Psychology, 27*, 348-367.

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback seeking on behavior in organizations. *Journal of Applied Psychology, 64*, 349-371.

In'nami, Y. (2006). The effects of test anxiety on listening test performance. *System: An International Journal of Educational Technology and Applied Linguistics, 34*, 317-340.

Judd, C. H. (1905). Practice without knowledge of results. *Psychological Review Monographs, 2*, 185-198.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254-284.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86*, 148-161.

Leventhal, G. S. (1980). What should be done with equity theory? New approaches to the study of fairness in social relationships. In K. J. Gergen, M. S. Greenberg, & R. H. Willis (Eds.), *Social Exchange: Advances in Theory and Research:* 27-55. New York, Plenum.

Lounsbury, J. W., Bobrow, W., & Jensen, J. B. (1989). Attitudes toward employment testing: Scale development, correlates, and "known-group" validation. *Professional Psychology: Research and Practice, 20*, 340-349.

Nease, A. A., Mudgett, B. O., & Quiñones, M. A. (1999). Relationships among feedback sign, self-efficacy, and acceptance of performance feedback. *Journal of Applied Psychology, 84*, 806-814.

Payne, S. C., Youngcourt, S. S., & Beaubien, J. M. (2007). A meta-analytic examination of the goal orientation nomological net. *Journal of Applied Psychology, 92*, 128-150.

Putwain, D. W., Woods, K. A., & Symes, W. (2010). Personal and situational predictors of test anxiety of students in post-compulsory education. *British Journal of Educational Psychology, 80*, 137-160.

Sarason, I. G., & Sarason, B. R. (1990). Test anxiety. In H. Leitenberg (Ed.), *Handbook of social and evaluative anxiety* (pp. 475-496). New York: Plenum.

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

Study Mode, LLC (2014). *Majortests.com Miller Analogies Test Prep*. Retrieved from http://www.majortests.com/mat/miller-analogies-test-practice.php.

Tonidandel, S., Quiñones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology, 87*, 320-332.

University of Cambridge (2014). Concerto platform for the development of on-line adaptive tests. Retrieved from http://www.psychometrics.cam.ac.uk/newconcerto.

Vandewalle, D. (1997). Development and validation of a work domain goal orientation instrument. *Educational and Psychological Measurement, 57*, 995-1015.

van Vianen, A. E. M., Taris, R., Scholten, E., & Schinkel, S. (2004). Perceived fairness in personnel selection: Determinants and outcomes in different stages of the assessment procedure. *International Journal of Selection and Assessment, 12*, 149-159.

von der Embse, N., Barterian, J., & Segool, N. (2013). Test anxiety interventions for children and adolescents: A systematic review of treatment studies from 2000-2010. *Psychology in the Schools, 50*, 57-71.

Wise, S. L., Plake, B. S., Eastman, L. A., Boettcher, L. L., & Lukin, M. E. (1986). The effects of item feedback and examinee control on test performance and anxiety in a computer-administered test. *Computers in Human Behavior, 2*, 21-29.

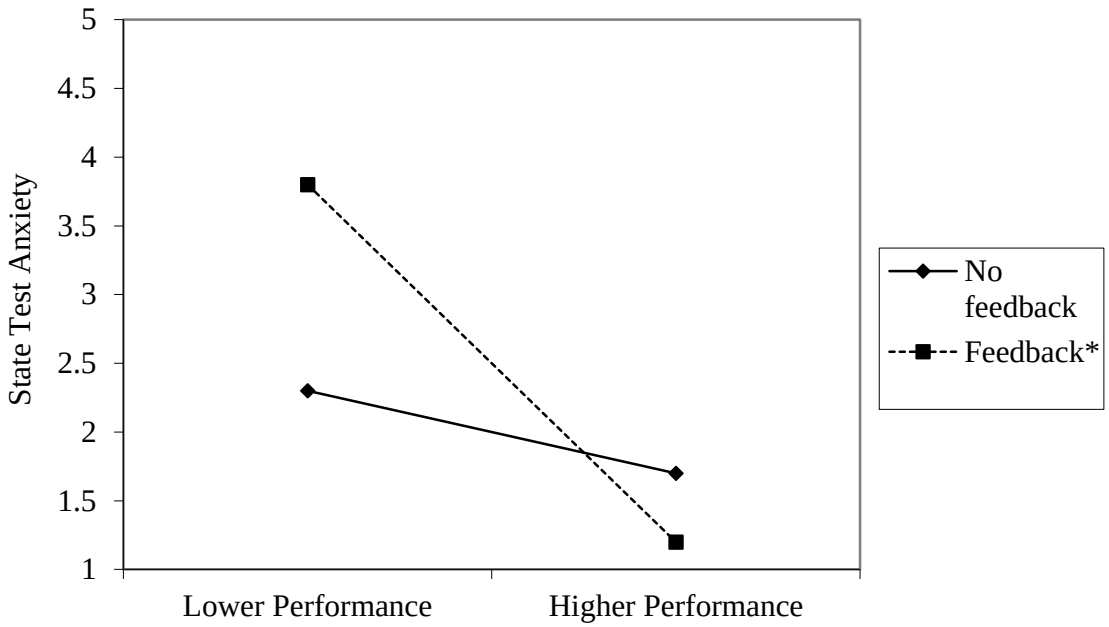
Zeidner, M. (1998). Test anxiety: The state of the art. New York: Plenum Press.

Appendix B, Table 1

Descriptive Statistics and Correlations among the Study Variables

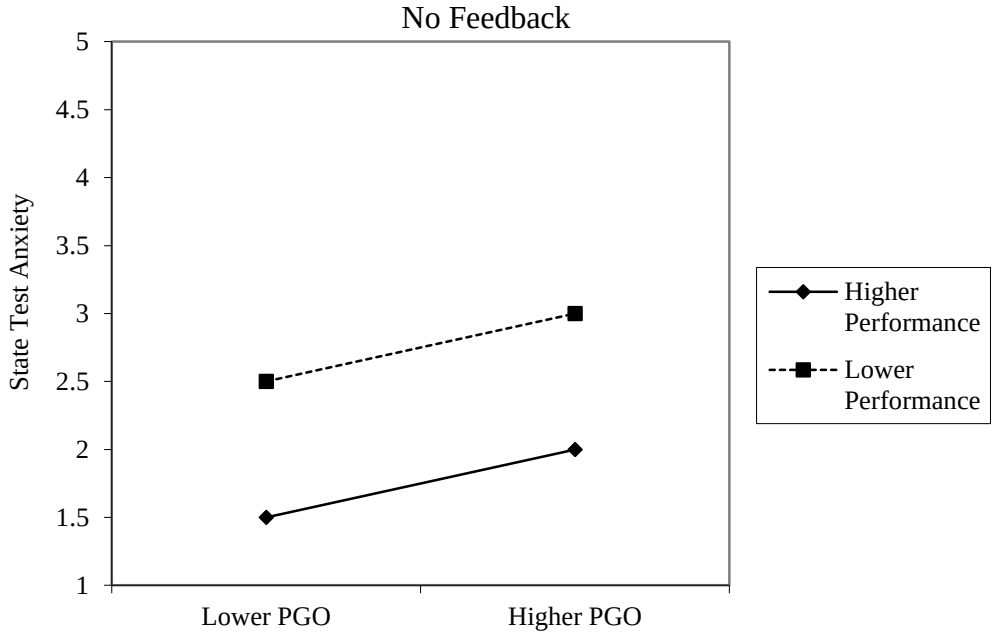
Variable	M	*SD*	1	2	3	4	5	6
1. General Reasoning	X.XX	X.XX	(.XX)					
2. PPGO	X.XX	X.XX	.XX	(.XX)				
3. APGO	X.XX	X.XX	.XX	.XX	(.XX)			
4. Feedback Acceptance	X.XX	X.XX	.XX	.XX	.XX	(.XX)		
5. State Test Anxiety	X.XX	X.XX	.XX	.XX	.XX	.XX	(.XX)	
6. Perceived Test Fairness	X.XX	X.XX	.XX	.XX	.XX	.XX	.XX	(.XX)

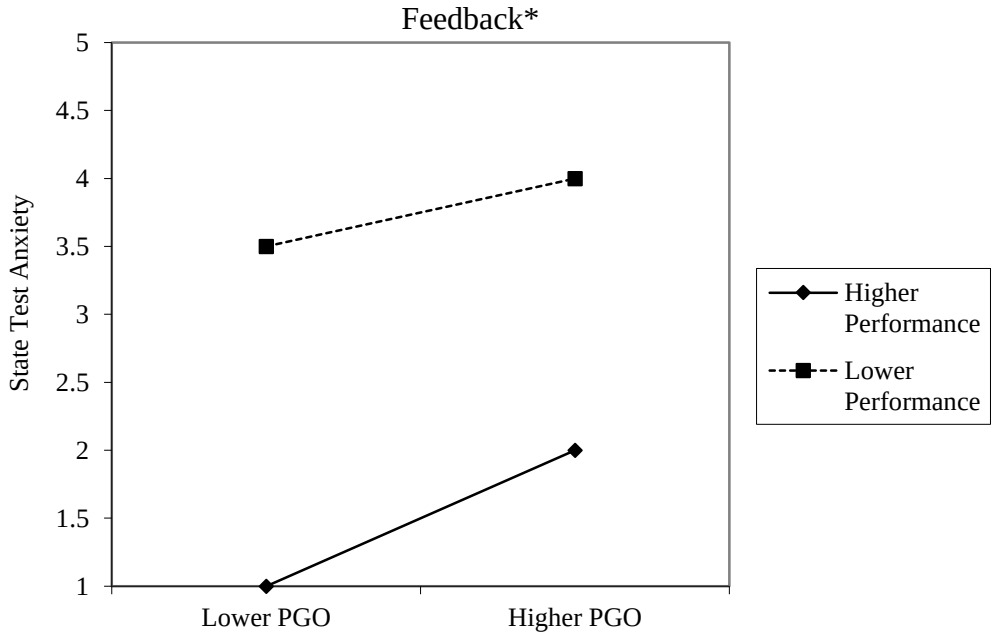
N = XX, **p* < .05, ***p* < .01, coefficient alpha reliabilities on diagonal. PPGO = Prove Performance Goal Orientation, APGO = Avoid Performance Goal Orientation.



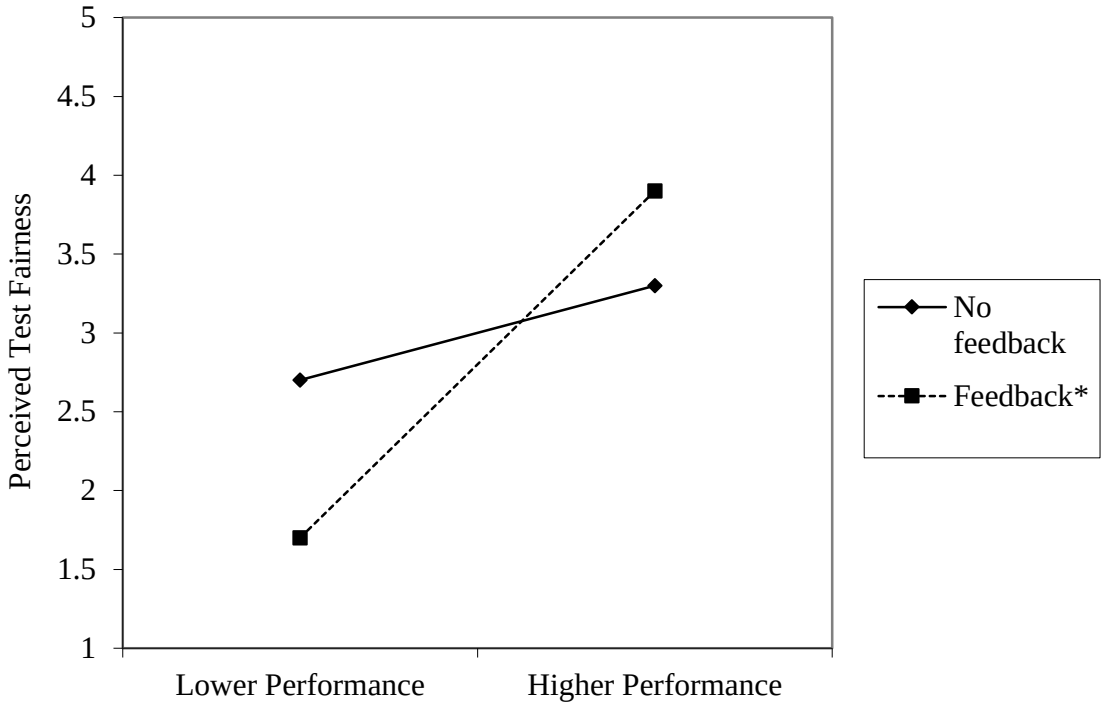
Appendix B, *Figure 1*. Hypothesized interaction between feedback and test performance on state test anxiety.

*Feedback will be tested across three modes: text feedback only, graphical feedback only, and text and graphical feedback together.



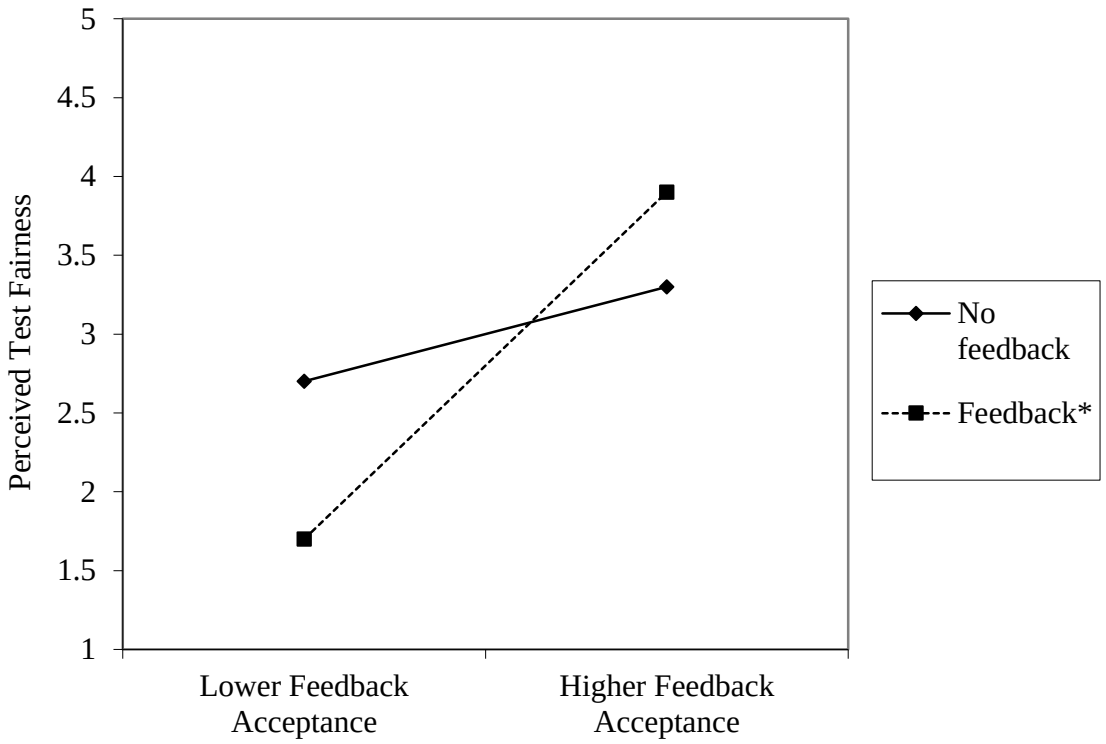


Appendix B, *Figure 2.* Hypothesized three-way interaction between feedback, performance, and performance goal orientation. PGO = Performance Goal Orientation. *Note.* This interaction will be tested for two types of PGO (prove and avoid).



Appendix B, *Figure 3.* Hypothesized interaction between feedback and test performance on perceived test fairness.

*Feedback will be tested across three modes: text feedback only, graphical feedback only, and text and graphical feedback together.



Appendix B, *Figure 4.* Hypothesized interaction between feedback and feedback acceptance on perceived test fairness

*Feedback will be tested across three modes: text feedback only, graphical feedback only, and text and graphical feedback together.

Appendices

Appendix B-1

Measures and Items

Demographics

1. What is your age?
2. What is your gender?
   0 = Female, 1 = Male
3. What is your ethnicity?
   1 =African American, 2 = Asian American, 3 = Caucasian, 4 = Hispanic, 5 = Native American, 6 = Other
4. How many hours per week do you work?
5. What is your country of origin?
6. What is your native language?
7. In what country do you currently reside?
8. What is your level of education?
   1 =Some High School, 2 = High School, 3 = Some college, 4 = College degree, 5 = Masters, 6 = PhD

Reasoning Items

Bank 1
1. (____) : PUCCINI :: SCULPTURE : OPERA
A. Cellini
B. Rembrandt
C. Wagner
D. Petrarch

2. (____) : SPEECH :: COORDINATED : MOVEMENT
A. predictive
B. rapid
C. prophetic
D. articulate

3. INCEPTION : CONCLUSION :: (____) : ARRIVAL
A. upshot
B. culmination
C. departure
D. escapade

4. SCINTILLATING : DULLNESS :: (____) : CALM

A. erudite
B. boisterous
C. cautious
D. exalted

5. ELUCIDATE : CLARITY :: ILLUMINATE : (____)
A. memory
B. problem
C. oblivion
D. light

6. SHARD : POTTERY :: (____) : WOOD
A. acorn
B. smoke
C. chair
D. splinter

7. CONIFER : (____) :: GRASS : PRAIRIE
A. taiga
B. Scandinavia
C. steppe
D. tundra

8. PENURY : MONEY :: STARVATION : (____)
A. sustenance
B. infirmity
C. illness
D. care

9. ARABLE : LAND :: (____) : WATERWAY
A. impenetrable
B. navigable
C. fertile
D. shallow

10. ATTENUATE : SIGNAL :: (____) : ENTHUSIASM
A. exacerbate
B. modify
C. dampen
D. elongate

11. (____) : SLAG :: FIRE : ASHES
A. quarry

B. furnace
C. automobile
D. silo

12. 17 : 19 :: (____) : 37
A. 39
B. 36
C. 34
D. 31

13. SRI LANKA : THAILAND :: (____) : SIAM
A. Ceylon
B. Malaysia
C. Burma
D. Colombo

14. (____) : -LY :: ADJECTIVE : ADVERB
A. ion
B. -ful
C. -ity
D. -ily

15. PITCH : (____) :: WAVELENGTH : AMPLITUDE
A. sound
B. timbre
C. loudness
D. color

Bank 2
1. ELECTED : (____) :: CONDEMNED : EXECUTION
A. graduation
B. inauguration
C. dismissal
D. exhibition

2. DIVIDEND : STOCKHOLDER :: (____) : AUTHOR
A. patent
B. royalty
C. wage
D. interest

3. ARCHIPELAGO : ISLAND :: CONSTELLATION : (____)
A. hamlet

B. zodiac
C. sea
D. star

4. (____) : ELABORATION :: SIMPLIFICATION : DECORATION
A. précis
B. revision
C. emendation
D. prodigality

5. (____) : WALK :: SIMPER : SMILE
A. mince
B. wave
C. waltz
D. hike

6. TWAIN : CLEMENS :: ELIOT : (____)
A. George
B. Mary
C. Bronte
D. Evans

7. FENESTRATION : (____) :: PORTAL : DOOR
A. mural
B. table
C. window
D. atrium

8. (____) : MARBLE :: SEDIMENTARY : METAMORPHIC
A. granite
B. slate
C. limestone
D. basalt

9. (____) : LENIENT :: MISER : CHARITABLE
A. philanthropist
B. virtuoso
C. hedonist
D. authoritarian

10. ALLAY : SUSPICION :: (____) : FEAR
A. plant
B. anger

C. generate
D. calm

11. MENDEL : NEWTON :: GENETICS : (____)
A. history
B. mechanics
C. linguistics
D. astrology

12. DIRECTORY : ADDRESS :: DICTIONARY : (____)
A. spelling
B. solecism
C. thesaurus
D. encyclopedia

13. FOOT : BOOT :: (____) : BOUGH
A. cough
B. bout
C. plough
D. tree

14. (____) : EYE :: COCHLEA : EAR
A. sinus
B. mollusk
C. fovea
D. fontanel

15. (____) : PARALLELOGRAM :: 1 : 2
A. square
B. rhombus
C. trapezium
D. cube

Bank 3
1. BOAST : LANGUAGE :: SWAGGER : (____)
A. anger
B. gait
C. sight
D. wealth

2. BELITTLE : DISPARAGE :: (____) : RIDICULE
A. jeopardize
B. efface

C. assuage
D. deride

3. (____) : INNOCUOUS :: REPREHENSIBLE : PRAISEWORTHY
A. pretentious
B. virulent
C. antiseptic
D. widespread

4. (____) : MINIMALIST :: ORNATE : UNADORNED
A. Rococo
B. Cubist
C. Pastoral
D. Pointillist

5. DESECRATE : (____) :: DESPOIL : BEAUTIFUL
A. holy
B. ugly
C. rich
D. corporeal

6. VIRTUOSO : ACCOMPLISHED :: (____) : GENEROUS
A. aesthete
B. hedonist
C. mogul
D. philanthropist

7. EMIT : MITE :: EVIL : (____)
A. huge
B. vile
C. veil
D. time

8. HOMOGENEOUS : KIND :: (____) : TIME
A. fast
B. disparate
C. diverse
D. contemporary

9. ANTHRACITE : (____) :: HARD : SOFT
A. diamond
B. peat
C. mica

D. uranium

10. AUTHENTIC : APOCRYPHAL :: CONVENTIONAL : (____)
A. dangerous
B. unorthodox
C. religious
D. chaotic

11. (____) : IODINE :: YELLOW : PURPLE
A. water
B. litmus
C. sulfur
D. alcohol

12. STANZA : POEM :: (____) : NOVEL
A. chapter
B. preface
C. sentence
D. language

13. BAIKAL : (____) :: TANGANYIKA : TANZANIA
A. Canada
B. Jutland
C. Sahara
D. Siberia

14. FRAGMENT : MERGE :: (____) : JOIN
A. engage
B. loosen
C. splinter
D. mend

15. PROLIFIC : (____) :: FERTILE : BARREN
A. withered
B. blooming
C. unfruitful
D. stale

Bank 4
1. GLASS : (____) :: METAL : LUSTER
A. portability
B. delicacy
C. porosity

D. translucence

2. (____) : PINE :: SPINE : CACTUS
A. crocus
B. trunk
C. flower
D. needle

3. MACBETH : OTHELLO :: (____) : JEALOUSY
A. royalty
B. solitude
C. ambition
D. foresight

4. OHM : RESISTANCE :: (____) : POTENTIAL
A. ampere
B. joule
C. volt
D. watt

5. (____) : STORY :: LIMERICK : POEM
A. anecdote
B. novel
C. prose
D. overture

6. EXPEDITE : (____) :: FACILITATE : MAKE EASIER
A. conscript
B. make harder
C. stretch
D. hasten

7. DIFFIDENT : CONFIDENCE :: UNSTABLE : (____)
A. peculiarity
B. balance
C. fervor
D. vigor

8. ERUDITE : LEARNING :: (____) : CAPRICE
A. whimsical
B. elegant
C. generous
D. conscientious

9. (____) : SQUARE :: 540 : 360
A. triangle
B. rectangle
C. pentagon
D. hexagon

10. APOCRYPHAL : CERTAINTY :: CHAOTIC : (____)
A. orthodoxy
B. order
C. incongruity
D. piety

11. INDEFATIGABLE : ENERGY :: (____) : WISDOM
A. illustrious
B. sagacious
C. enervated
D. experienced

12. MAUNA LOA : ETNA :: HAWAII : (____)
A. Sicily
B. Alps
C. Crete
D. Bosporus

13. CUBISM : PICASSO :: (____) : DALI
A. Romanticism
B. Classicism
C. Realism
D. Surrealism

14. (____) : ACTOR :: SOLO : SINGER
A. monologue
B. performance
C. repertoire
D. drama

15. STOIC : (____) :: BENEFACTOR : GENEROSITY
A. faith
B. fortitude
C. ineptitude
D. charity

Bank 5

1. EARTH : (____) :: MARS : PHOBOS
A. Venus
B. Sirius
C. moon
D. sun

2. DIPLOMAT : TACT :: (____) : SKILL
A. novice
B. virtuoso
C. sybarite
D. dilettante

3. OUTWIT : RUSE :: FRUSTRATE : (____)
A. desire
B. fiasco
C. irritant
D. statute

4. COTTAGE : (____) :: POND : LAKE
A. hamlet
B. camp
C. tent
D. mansion

5. (____) : FAULKNER : WESSEX : YOKNAPATAWPHA
A. DICKENS
B. HARDY
C. HAWTHORNE
D. SAWYER

6. (____) : VENGEFUL :: CONCILIATE : FLEXIBLE
A. ameliorate
B. retaliate
C. ruminate
D. iterate

7. SPORADIC : CONTINUOUS :: (____) : CACOPHONOUS
A. tranquil
B. cathartic
C. perennial
D. fitful

8. HONE : SHARP :: (____) : SHINY
A. caulk
B. scald
C. burnish
D. rusticate

9. (____) : HYPOTHESIS :: DEDUCE : THEOREM
A. adjudge
B. terminate
C. abrogate
D. formulate

10. DEVANGIRI : CYRILLIC :: HINDI : (____)
A. Russian
B. Swahili
C. French
D. Sanskrit

11. RACONTEUR : NARRATE :: (____) : ENTERTAIN
A. ruffian
B. recluse
C. host
D. spendthrift

12. TAWDRY : SHOWY :: MERETRICIOUS : (____)
A. attractive
B. urbane
C. delicate
D. flexible

13. (____) : MAGNESIUM :: LIQUID : SOLID
A. copper
B. mercury
C. aluminum
D. sulfur

14. APPENDIX : BOOK :: (____) : PLAY
A. . overture
B. . premise
C. . glossary
D. . epilogue

15. HYPOTENUSE : PERIMETER :: ARC : (____)

A. radius
B. diameter
C. circumference
D. diagonal

Bank 6
1. (____) : DISPLEASURE :: FIDGET : RESTLESSNESS
A. hunch
B. frown
C. strut
D. squirm

2. REAM : PAPER :: (____) : HAY
A. hutch
B. bale
C. coop
D. quart

3. PERFORATION : SEAL :: (____) : SKIN
A. contraption
B. adhesion
C. scar
D. laceration

4. ARGON : ELEMENT :: (____) : COMPOUND
A. copper
B. water
C. oxygen
D. lead

5. PROBITY : GUILE :: (____) : LAZINESS
A. industry
B. felicity
C. ingenuity
D. decorum

6. TURNIP : (____) :: POTATO : STEM
A. flower
B. vegetable
C. root
D. gourd

7. 0.125 : (____) :: 0.75 : ¾

A. 1/8
B. 1/9
C. 1/16
D. 1/20

8. (____) : SEPARATE :: JOIN : REND
A. desecrate
B. capitulate
C. promulgate
D. amalgamate

9. CAROUSE : SEDATE :: ENTHUSE : (____)
A. lukewarm
B. voluble
C. hidden
D. controversial

10. ZEALOT : PASSION :: (___) : DECEPTION
A. patriarch
B. quack
C. sybarite
D. libertarian

11. (____) : JUNG :: ECONOMICS : PSYCHIATRY
A. Freud
B. Pasteur
C. Keynes
D. Leavis

12. FRANKENSTEIN : SHELLEY :: DRACULA : (____)
A. Poe
B. Collins
C. Byron
D. Stoker

13. CHOLERIC : ANGER :: (____) : DOUBT
A. narcissistic
B. gullible
C. skeptical
D. sanguine

14. (____) : TREACLE :: SATIN : SMOOTH
A. saccharine

B. volatile
C. viscous
D. brittle

15. AVARICE : (____) :: AMBITION : SUCCESS
A. money
B. success
C. respect
D. trepidation

Bank 7
1. SNEER : CONTEMPT :: (____) : ANGER
A. lament
B. glower
C. confess
D. chide

2. (____) : QUARREL :: SKIRMISH : BATTLE
A. war
B. tiff
C. fiasco
D. rendezvous

3. DAPPLED : SPOTS :: RIDDLED : (____)
A. holes
B. lines
C. squares
D. stains

4. ARENA : CONFLICT :: (____) : DEBATE
A. forum
B. agenda
C. precinct
D. building

5. PETRIFY: STONE :: OSSIFY : (____)
A. disinterest
B. inscription
C. bone
D. saint

6. (____) : CAUSE :: SECTARIAN : SECT
A. skeptic

B. pariah
C. maverick
D. partisan

7. CALORIE : JOULE :: (____) : METER
A. second
B. radian
C. inch
D. kilogram

8. (____) : SURPASSED :: INDISPUTABLE : CHALLENGED
A. grandiose
B. peerless
C. opaque
D. lavish

9. EMBARKATION : VOYAGE :: (____) : PROJECT
A. criterion
B. inception
C. delineation
D. diatribe

10. ACROBAT : (____) :: SURGEON : DEXTROUS
A. nuanced
B. circumspect
C. lithe
D. bombastic

11. STANCH : FLOW :: (____) : OSCILLATION
A. corroborate
B. whet
C. validate
D. damp

12. (____) : NUMBER :: DILATE : SIZE
A. disseminate
B. promulgate
C. atrophy
D. proliferate

13. SOMME : (____) :: 1916 : 1815
A. Waterloo
B. Hastings

C. Iwo Jima
D. Pearl Harbor

14. (____) : BODY :: RATIONAL : MIND
A. sartorial
B. macroscopic
C. corporeal
D. fetal

15. VENUS : (____) :: APHRODITE : ATHENE
A. Minerva
B. Juno
C. Sappho
D. Diana

Bank 8
1. SOLVE : MYSTERY :: (____) : CODE
A. ensure
B. decipher
C. encrypt
D. conquer

2. LIONIZE : CELEBRITY :: (____) : IDOL
A. worship
B. ostracize
C. swindle
D. decorate

3. (____) : PUPIL :: DEPORT : ALIEN
A. inter
B. educate
C. test
D. expel

4. OBOE : (____) :: CELLO : VIOLIN
A. double bass
B. trombone
C. flute
D. piano

5. CONFLAGRATION : (____) :: METROPOLIS : CITY
A. fire
B. building

C. town
D. brook

6. SLUGGARD : ENERGY :: SLATTERN : (____)
A. disorder
B. expression
C. laziness
D. tidiness

7. JUTLAND : (____) :: HOKKAIDO : JAPAN
A. Germany
B. Denmark
C. New Zealand
D. British Isles

8. (____) : LOCAL :: MERCENARY : MONEY
A. parochial
B. miserly
C. capricious
D. germinal

9. INCARCERATION : PRISON :: (____) : ISOLATION
A. moratorium
B. sequestration
C. dejection
D. pedantry

10. INGRATE : GRATITUDE :: PROFLIGATE : (____)
A. amorousness
B. virtue
C. frugality
D. timidity

11. (____) : SINCERE :: FLUSTERED : COMPOSED
A. arrogant
B. disinterested
C. blatant
D. hypocritical

12. MALINGER : WORK :: (____) : OBLIGATION
A. shirk
B. fulfill
C. fight

D. invent

13. UBIQUITOUS : EVERYWHERE :: (____) : FOREVER
A. impervious
B. penetrable
C. endemic
D. eternal

14. pH 5 : pH 7 :: (____) : NEUTRAL
A. alkaline
B. acid
C. basic
D. caustic

15. (____) : EMPHASIS :: ADORN : DECORATION
A. confound
B. underplay
C. underscore
D. undermine

Bank 9
1. WOOD : (____) :: BUTTER : KNIFE
A. string
B. paper
C. saw
D. drill

2. FORD : (____) :: STREAM : MOUNTAIN
A. pass
B. route
C. field
D. scale

3. SUPPORT : PATRON :: (____) : INSPIRATION
A. braggart
B. muse
C. parent
D. soldier

4. (____) : WARDEN :: MUSEUM : PRISON
A. archaeologist
B. servant
C. conductor

D. curator

5. INCUMBENT : OFFICE :: MONARCH : (____)
A. presidency
B. accommodation
C. station
D. throne

6. (____) : EXPERIENCE :: VAGRANT : ABODE
A. protestor
B. magnate
C. neophyte
D. scholar

7. SYMPHONY : MUSIC :: (____) : POEM
A. epic
B. anecdote
C. opera
D. anthem

8. (____) : PALINDROME :: REARRANGED : REVERSED
A. epithet
B. metaphor
C. acronym
D. anagram

9. SYNTHESIS : PARTS :: CONFLUENCE : (____)
A. mergers
B. streams
C. ores
D. parallels

10. SALK : PASTEUR :: POLIO : (____)
A. rabies
B. liquids
C. smallpox
D. antibodies

11. CRITERIA : DATA :: CRITERION : (____)
A. information
B. datum
C. item
D. date

12. BELLICOSE : (____) :: PLIANT : TRACTABILITY
A. compliance
B. aggression
C. vacillation
D. sobriety

13. CHOLERIC : ANGER :: (____) : OPTIMISM
A. sanguine
B. saturnine
C. introverted
D. jaundiced

14. TORTUOUS : (____) :: VENAL : VENIAL
A. trivial
B. arduous
C. torturous
D. corrupt

15. (____) : AMELIORATE :: WORSE : BETTER
A. elucidate
B. engender
C. accredit
D. exacerbate

Bank 10
1. TARSAL : (____) :: CARPAL : WRIST
A. heel
B. toe
C. tunnel
D. hinge

2. PORTION : DOSE :: FOOD : (____)
A. rain
B. drug
C. dessert
D. amount

3. ESSAY : THESIS :: ARTICLE : (____)
A. protagonist
B. book
C. fiction
D. topic

4. MUSICIAN : VIRTUOSO :: BARD : (___)
A. journalist
B. poet
C. politician
D. sophomore

5. FURTIVE : BEHAVIOR :: (_____) : ACTION
A. evanescent
B. cerebral
C. conditioned
D. covert

6. BOVINE : OVINE :: COW : (____)
A. sheep
B. chickens
C. calf
D. lowing

7. DNA : NUCLEOTIDE :: PROTEIN : (____)
A. amino acid
B. ribosome
C. fatty acid
D. carbohydrate

8. BEACON : (____) :: PRECEPT : PRINCIPLE
A. shroud
B. light
C. confusion
D. call

9. EXTRICATE : ENSNARED :: (____) : ENSLAVED
A. entice
B. condone
C. coordinate
D. liberate

10. EDITORIAL : (____) :: BIOGRAPHY : LIFE HISTORY
A. doctrine
B. censorship
C. synopsis
D. opinion

11. FETTER : (____) :: GAG : SPEECH
A. leg
B. movement
C. catacomb
D. law breaking

12. (____) : DEADLOCK :: QUANDARY : DILEMMA
A. moratorium
B. impasse
C. exegesis
D. paradigm

13. CATHARSIS : EMOTION :: ABSOLUTION : (____)
A. malady
B. innocence
C. guilt
D. mourning

14. LYELL : DALTON :: GEOLOGY : (_____)
A. biology
B. chemistry
C. sociology
D. music

15. NASCENT : (____) :: GERMINAL : SENESCENT
A. sophomoric
B. covetous
C. moribund
D. shrewd

State Test Anxiety

Using the scale below as a guide, indicate for each statement how much you agree or disagree with the statement. (*Note.* Response scale will be used for all subsequent items.)

1 = strongly disagree
2 = disagree
3 = somewhat disagree
4 = neither agree nor disagree
5 = somewhat agree
6 = agree
7 = strongly agree

State worry

1. I was concerned about what would happen if I did poorly.
2. Thinking about my performance on the test interfered with my work on the test.
3. During the test I got so nervous that I forgot facts I really knew.
4. I thought about how important the test was for me.

State emotionality

5. While taking the test, I had an uneasy, upset feeling.
6. I felt very panicky when I was taking the test.
7. I felt very jittery when I was taking the test.
8. During the test, I felt very tense.

Goal Orientation

Prove Performance Goal Orientation

1. I'm concerned with showing that I can perform better than my coworkers.
2. I try to figure out what it takes to prove my ability to others at work.
3. I enjoy it when others at work are aware of how well I am doing.
4. I prefer to work on projects where I can prove my ability to others.

Avoid Performance Goal Orientation

5. I would avoid taking on a new task if there was a chance that I would appear rather incompetent to others.
6. Avoiding a show of low ability is more important to me than learning a new skill.
7. I'm concerned about taking on a task at work if my performance would reveal that I had low ability.
8. I prefer to avoid situations at work where I might perform poorly.

Perceived Test Fairness

1. The test used was not a reliable and valid indicator of my capabilities.
2. The test used is an unfair test of a person's true capabilities. R
3. The testing procedure obtained accurate information about each person's capabilities.
4. Performance on the test was influenced by things that should not have been considered. R
5. Under the circumstances, the test was fair.
6. I have strong doubts that the test really measures a person's general reasoning. R
7. I feel other procedures should have been used to assess people. R
8. The test should not have been used to assess people. R

Feedback Acceptance

1. The feedback I received was an accurate evaluation of my performance.
2. I do not agree with the feedback provided. R

3. It is hard to take the feedback seriously. R
4. I do not believe that the feedback is accurate. R

Manipulation Check

1. Each time I answered a test question, I received feedback about how I was doing on the test.
2. I always knew how well I was doing while taking the test because of the feedback display.
3. Feedback was provided during the test to show my score.
4. During the test, there was feedback that gave me updates about my test score.

Attention Check

1. Please answer “strongly agree.”
2. Please answer “disagree.”
3. Please answer “neither agree nor disagree.”

Note. “R” indicates reverse-scored items.

Appendix B-2

Instructions

You are about to complete a series of reasoning items as part of an experiment. The purpose of this experiment is to identify individuals who are best at completing analogy items. After all of the data has been collected, the three examinees who scored the highest will receive a $25 Amazon.com gift card. If you would like to be eligible for this, you will be asked to provide your email address later so that you may be contacted in the event that you have won. Recipients will be notified within three weeks of the experiment's completion.

Thank you for your participation.

Appendix B-3

Examples of Feedback Displays

Text Feedback

Please select the best answer.

Pendulum : clock :: ?

- Asparagus : beet
- Coat : jacket
- Piston : engine
- Shirt : tie
- Cookie : batter

Theta Level:

.3472

Your theta level is an indication of your ability.

Standard Error of Measurement:

.874

The Standard Error of Measurement is the precision with which your score is being calculated. The lower this is, the more precisely your score is estimated.

Item Difficulty

.21

The item difficulty will become higher when you answer questions correctly, and lower when you answer questions incorrectly.

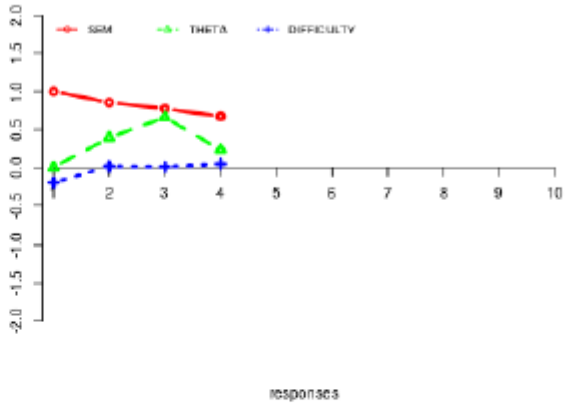
Graphical Feedback

Please select the best answer.

Pendulum : clock :: ?

- Asparagus : beet
- Coat : jacket
- Piston : engine
- Shirt : tie
- Cookie : batter

Theta, Standard Error of Measurement (SEM), and Difficulty Graph



Your theta level is an indication of your ability.

The Standard Error of Measurement is the precision with which your score is being calculated. The lower this is, the more precisely your score is estimated.

The item difficulty will become higher when you answer questions correctly, and lower when you answer questions incorrectly.

Text and Graphical Feedback

Please select the best answer.

Pendulum : clock :: ?

- Asparagus : beet
- Coat : jacket
- Piston : engine
- Shirt : tie
- Cookie : batter

Theta Level:

.3472

Your theta level is an indication of your ability.

Standard Error of Measurement:

.874

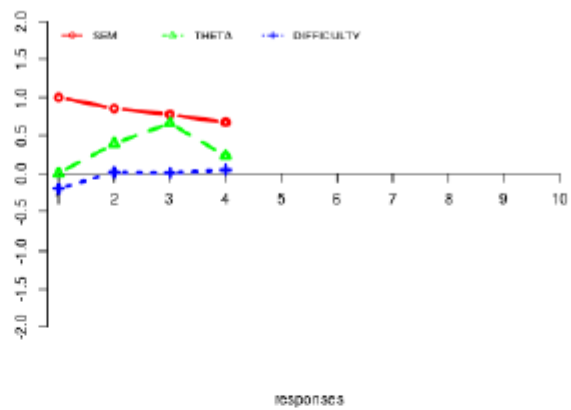
The Standard Error of Measurement is the precision with which your score is being calculated. The lower this is, the more precisely your score is estimated.

Item Difficulty

.21

The item difficulty will become higher when you answer questions correctly, and lower when you answer questions incorrectly.

Theta, Standard Error of Measurement (SEM), and Difficulty Graph



Your theta level is an indication of your ability.

The Standard Error of Measurement is the precision with which your score is being calculated. The lower this is, the more precisely your score is estimated.

The item difficulty will become higher when you answer questions correctly, and lower when you answer questions incorrectly.