MAXIMIZING INFORMATION: APPLICATIONS OF IDEAL POINT MODELING AND INNOVATIVE ITEM DESIGN TO PERSONALITY MEASUREMENT

Heidi Vanessa Leeson

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Education,

The University of Auckland, 2008

ABSTRACT

Recent research has challenged the way in which personality and attitude constructs are measured. Alternatives have been offered as to how non-cognitive responses are modeled, the mode of delivery used when administrating such scales, and the impact of technology in measuring personality. Thus, the major purpose of the studies in this thesis concerns two interrelated issues of personality research, namely the way personality responses are best modeled, and the most optimal mode by which personality items are presented and associated modal issues. Three studies are presented. First, recent developments using an ideal point approach to scale construction are outlined, and an empirical study compares modeling personality items based on an ideal point approach (generalized graded unfolding model; GGUM) and a dominance approach (graded response model: GRM). Second, an extensive review of literature pertaining to the mode effect when transferring paper-and-pencil measures to screen was conducted, in addition to a review of the various types of computerized and innovative items and their associated psychometric information. Finally, nine innovative items were developed using various multimedia features (e.g., video, graphics, and audio) to ascertain the advantages of these methods to present items constructed to elicit response behavior underlying ideal point approaches, namely, typical response behavior.

It was found that the dominance IRT model continued to produce superior model-data fit for most items, more attention needs to be placed on developing principles for constructing ideal point type items, the web-based version supplied 20% more construct information than the paper version, and innovative items seem to

provide more data-model fit for students with lower personality attributes. While the innovative items may require more initial outlay in terms of time and development costs, they have the capacity to provide more information regarding test-takers' personality levels, potentially using fewer items.

ACKNOWLEDGEMENTS

I would like to thank the numerous people, students, and schools who have assisted or participated in the research contained in this thesis. I am especially grateful for the support and guidance of my supervisor Professor John Hattie. Thank you for the invigorating debates on psychometric issues and statistical approaches, and your never-ending enthusiasm for my work. Your feedback, advice, and assistance have made this thesis not only possible, but an exciting adventure. I would also like to thank my second supervisor Associate Professor Mike Townsend. Thanks also to the fabulous team at Project asTTle, you are truly my 'work family' and I have appreciated the incredible support (and humour) that you have all given me throughout this journey. Thank you also Jana and Satomi, you have been great fellow PhD students, and I have appreciated the support from both of you.

I acknowledge the financial support from the University of Auckland, whose scholarship has meant that I could undertake this commitment financially. Special thanks also to the Ministry of Education, who provided the significant funding required for the innovative item development conducted in this thesis. Thank you for making it possible to undertake that project.

At a personal level, a special thanks to my family for providing me with the motivation and desire to achieve this goal, you will never know your impact. Extra special thanks to my Mum for deciding that I should learn to read when I was 3 years old, thank you for introducing to me the worlds created from text in a book.

Lastly, I would like to acknowledge and dedicate this thesis to Andrew, Ben and Lilly. You are everything.

LIST OF TABLES

Table 1	Sample items from the Academic Self-Worth Scale	39
Table 2	Total variance explained in the EFA for factors whose eigenvalues exceed 1.00	50
Table 3	Sub-Scale Characteristics	50
Table 4	GGUM and GRM Average Item Parameter Estimations	52
Table 5	Means, Standard Deviations, and Frequencies of Chi-Square to *df* Ratio in the Calibration Sample	54
Table 6	Number of Fit Plots Providing Best-Fit to Sub-Scales	55
Table 7	The twelve fonts compared in Bernard, Mills, Peterson, et al. (2001) study	77
Table 8	Summary of the main human and technological findings	113
Table 9	Descriptive statistics and cross-mode correlations for the RSPS – Progress items	133
Table 10	Item factor loadings, absolute and incremental fit indices for web-based and paper-and-pencil questionnaires	136
Table 11	Descriptive statistics and difficulty and discrimination item parameters for paper-and-pencil and web-based versions	137
Table 12	Frequencies and Percentages for Reasons for Preferred Mode Choice	144
Table 13	Frequencies and Percentages for the Mode that Best Reflected their Self-Perceptions of Reading Progress	146
Table 14	Frequencies and Percentages for Likes and Dislikes of the Web-based Questionnaire	147
Table 15	Frequencies and Percentages for Participants Attitudes Regarding Different Features of the Web-based Questionnaire	148

LIST OF FIGURES

Figure 1	Fit Plots for the GGUM and GRM (Item 11 of the Reflectivity sub-scale): (a) Strongly Disagree Option (GRM), (b) Strongly Disagree Option (GGUM), (c) Disagree Option (GRM), (d) Disagree Option (GGUM), (e) Neither Disagree Nor Agree Option (GRM), (f) Neither Disagree Nor Agree Option (GGUM), (g) Agree Option (GRM), (h) Agree Option (GGUM), (i) Strongly Agree Option (GRM), (j) Strongly Agree Option (GGUM).	56
Figure 2	Fit Plots for the GGUM and GRM (Item 8 of the Self-Presented sub-scale): (a) Strongly Disagree Option (GRM), (b) Strongly Disagree Option (GGUM), (c) Disagree Option (GRM), (d) Disagree Option (GGUM), (e) Neither Disagree Nor Agree Option (GRM), (f) Neither Disagree Nor Agree Option (GGUM), (g) Agree Option (GRM), (h) Agree Option (GGUM), (i) Strongly Agree Option (GRM), (j) Strongly Agree Option (GGUM)	59
Figure 3	An example of a figural response item (Zenisky & Sireci, 2002).	92
Figure 4	An example of a figural response item (Martinez & Jenkins, 1993).	93
Figure 5	An example of a drag-and-drop item presenting the Periodic Table of the Elements with missing elements (Zenisky & Sireci, 2002).	94
Figure 6	A drag-and-drop item requiring test-takers to select and drag their chart title choices to their correct position on the work area of the item (Microsoft Corporation, 2003).	95
Figure 7	An example of a graphical modeling item format (Bennett, Morley, Quardt, & Rock, 2000).	96
Figure 8	An example of a drop-and-connect item format (Jodoin, 2003).	97
Figure 9	An example of a specifying relationships item type (Zenisky & Sireci, 2002).	98
Figure 10	An example of a create-a-tree item type (Zenisky & Sireci, 2002).	99
Figure 11	An example of an ETS capturing frame item (Bennett, Goodman, Hessinger, Kahn, Ligget, Marshall, & Zack, 1999).	100
Figure 12	An example of an ETS capturing frame item (Bennett, Goodman, Hessinger, Kahn, Ligget, Marshall, & Zack, 1999).	101

Figure 13	An example of a multiple selection item (Shotland, Alliger, & Sales, 1998).	102
Figure 14	An example of an ETS analyzing situation item (Bennett, Goodman, Hessinger, Kahn, Ligget, Marshall, & Zack, 1999).	103
Figure 15	Example of a GRE essay item format whereby test-takers are required to type a response into a text box that has basic editing functions available (cut, paste, undo) (Microsoft Corporation, 2003).	104
Figure 16	Example of an ETS generating examples item (Bennett & Rock, 1998).	105
Figure 17	An example of an ETS document-analysis item that asks the test-taker to analyze and compare the persuasive techniques adopted in wartime propaganda (Bennett, Goodman, Hessinger, Kahn, Ligget, Marshall, & Zack, 1999).	106
Figure 18	An example of a text-editing item (Breland, 1999).	107
Figure 19	Item information for the nine RSPS Progress items in both paper-and-pencil and web-based versions.	139
Figure 20	Average IRT information across theta levels for paper-and-pencil and web-based versions of the RSPS Progress sub-scale	141

TABLE OF CONTENTS

ABSTRACT ..... **I**

ACKNOWLEDGEMENTS ..... **III**

LIST OF TABLES ..... **IV**

LIST OF FIGURES ..... **V**

TABLE OF CONTENTS ..... **VII**

CHAPTER ONE **1**

INTRODUCTION 1

Purpose of the Research *5*

Significance of the Research *6*

Design of the Research *9*

CHAPTER TWO **11**

LITERATURE REVIEW 11

Self-worth Strategies *12*

Self-handicapping 12

Active self-handicapping 13

Self-presented self-handicapping 14

Self-presented affective self-handicapping 15

Defensive Pessimism 15

Active Defensive Expectations 16

Self-Presented Defensive Expectations 16

Reflectivity 16

Response Behaviors *17*

Measurement Theory *19*

Item Response Theory (IRT) 19

Graded Response Model (GRM) 20

Ideal Point Item Response Theory 21

Generalized Graded Unfolding Model (GGUM) 22

Applications of Ideal Point and Item Response Approaches to Personality Items 23

Reading Self-efficacy *26*

Innovative Item Design - Framework *27*

Multimedia Testing 31

Equivalence Testing *32*

Efficiency Testing *35*

CHAPTER THREE **37**

STUDY ONE: COMPARISON MODEL-DATA FIT FROM IRT AND IDEAL POINT IRT MODELS TO AN IDEAL POINT SCALE 37
Method *38*
Participants 38
Scale Development 38
Instrument 38
Self-handicapping items 39
Defensive pessimism items 41
Procedure 42
Statistical Analysis 44
Cross-validation 44
Exploratory Factor Analysis of the Academic Self-Worth Scale 45
IRT Item Parameter Estimation 45
Model-Data Fit 46
Results *49*
Exploratory Factor Analysis of the Academic Self-Worth Scale 49
IRT Item Parameter Estimation 52
Model-Data Fit 54
Discussion *62*

CHAPTER FOUR 66

STUDY TWO: A LITERATURE REVIEW OF HUMAN, TECHNOLOGICAL AND ITEM DESIGN ISSUES IN COMPUTERIZED TESTING 66
Participant Issues *67*
Race, Ethnicity, and Gender 67
Cognitive Processing 68
Ability 69
Familiarity with Computers 71
Computer Anxiety 73
User Interface – Legibility *74*
Screen Size and Resolution 74
Font Characteristics 76
Line Length 79
Number of Lines 82
Interline Spacing 83
White Space 84
User Interface – Interactive *86*
Scrolling 86
Item Review 87
Item Presentation 90
Computerized Item Design Types *91*
Item Format and Response Actions 91
Mouse-based Response Action 91
Figural Response Items 91

Drag-and-drop Item 93
Graphical Modeling Item 95
Drag-and-connect Item 97
Specifying Relationships Item 98
Create-a-tree Item 99
Capturing Frames Item 99
Multiple Selection Item 101
Analyzing Situation Item 102
Text-based Response Action 103
Essay/Short Answer Item 104
Generating Examples Item 105
Document-analysis Item 106
Text-editing Item 107
Innovative Item Response Formats 108
Task Constraints 110
Discussion *112*

CHAPTER FIVE 121

STUDY THREE: WEB-BASED AND PAPER-AND-PENCIL VERSIONS OF THE READER SELF-PERCEPTION SCALE: A COMPARISON OF MEASUREMENT EFFICIENCY AND PARTICIPANTS' PERCEPTIONS 121
Method *123*
Participants 123
Design 123
Measures 124
Reader Self-Perception Scale (RSPS) – Paper-and-Pencil 124
Reader Self-Perception Scale (RSPS) – Web-based 125
RSPS – Progress sub-scale modifications 126
Post-assessment feedback measure 128
Procedure 129
Paper-and-pencil administration 130
Web administration 130
Results *131*
Data Analysis 131
Descriptive Statistics 131
Confirmatory Factor Analysis 134
Item Parameter Statistics 136
Efficiency Analysis 137
Item Level 138
Test Level 139
Mode Preferences 142
Discussion *149*

CHAPTER SIX 157

DISCUSSION 157
Findings *163*
Implications *170*
Future Research *173*
Contribution *178*
REFERENCES **181**

CHAPTER ONE

INTRODUCTION

Whilst personality research has received an increased focus over the last several decades, recently two issues have challenged the way that personality constructs are measured. The first issue relates to the way that responses are modeled given the type of behavior being exhibited by the test-taker, and the second concerns the optimum mode by which personality measures are administered. These issues have arisen from the recent developments of both specialized item response theory (IRT) models and the impact and increased accessibility of the personal computer.

Over the past three decades, considerable research has examined the use of IRT approaches to modeling response data from various personality measures (e.g., Reise, 1999; Reise & Henson, 2000; Rouse, Finger, & Butcher, 1999; Steinberg & Thissen, 1995). Typically, the one- and two-parameter models have been applied to personality scales in an attempt to find a more appropriate treatment of response data than that provided by the more traditional approach where values are simply summed across items. However, while the application of traditional IRT models have provided a significantly more flexible and robust approach to classical test theory's (CTT) handling of personality responses, it cannot be deduced that such models necessarily provide the best *representation* of this data. Instead, recent conjecture has re-emphasized Cronbach's (1949) and Thurstone's (1928, 1931) earlier argument that the response behaviors exhibited by test-takers were different when responding to personality (or attitude) items, than when responding to cognitive ability items. Cronbach and Thurstone proposed that *maximum* behavior occurred when test-takers

were aware that their performance outcomes were to be measured against presubscribed standards. Thus, in many cognitive testing environments, test-taker response behavior is largely constrained to right/wrong responses, often under the pressure of time limits. Conversely, Cronbach referred to the less constrained, but more complex testing environment of non-cognitive ability tests as inducing a *typical* performance behavior from test-takers. Here, because of the lack of pressure (e.g., typically no or reduced pressure of time limits) and requirement for specific knowledge recall and recognition, test-takers exhibited variability in effort and motivation. As a result of these response and behavior distinctions, it was proposed (Coombs, 1964; Roberts, Laughlin, & Wedell, 1999) that the resulting response data would be modeled more appropriately using an "ideal point" approach to scoring (e.g., generalized graded unfolding model), rather than the traditional "dominance-based" models (e.g., graded response model). More recently, studies have explored this possibility and the degree of model-data fit generated from these two modeling approaches (Chernyshenko, 2002; Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Maydeu-Olivares, 2005; Roberts, Laughlin, & Wedell, 1999).

Establishing the most effective method of modeling responses from personality items, however, is only one aspect of optimizing measurement of these constructs. Of similar import is the mode by which these items are presented. Given the technological viability of transferring paper-and-pencil measures to computer screen, there has been active research in the area of the mode differences between these two administration approaches. Previously, most of the research has focused on mode effect issues at the test-taker level. Effects such as, differential responding based on race, ethnicity and gender (e.g., Boodoo, 1998; Parshall & Kromrey, 1993), cognitive

processing differences, such as comprehension (Mead & Drasgow, 1993), memory and comprehension (Mayes, Sims, & Koonce, 2001), reading times (Noyes & Garland, 2003), and familiarity/anxiety with computers (Lynch, 2000; McDonald, 2002; Powers & O'Neill, 1992) have all been investigated in relation to their impact on performance across modes. In contrast, user interface/typology issues, which are inherent design features of computerized scales (e.g., screen size, font characteristics, whitespace), have received surprisingly little attention from measurement researchers. Similarly, psychometric research focusing on the interactive functionality of onscreen measures, such as scrolling and reviewing items is either outdated given current application design and function capabilities, limited in its generalizability, or has not taken sufficient account of the burgeoning literature on the human-computer interface (typically addressed in the engineering and computer science literature).

As well as establishing the potential mode effect differences between paper-and-pencil and computerized tests, research has also focused on utilizing the distinct advantages computerized administration has to offer (e.g., multimedia applications). In addition to providing more effective ways of scoring tasks (e.g., essays, constructed responses), computer administration permits the opportunity for test developers to re-envision how items might be designed and codified in order to allow innovative ways to measure constructs that are assessed, perhaps less effectively via paper administration. The widespread availability and usability of multimedia packages has provided opportunities to create interactive and dynamic test items, which may arguably result both in more authentic and higher order tasks being assessed (Jodoin, 2003). For example, the inclusion of video, graphics, and audio may result in increased capturing of the variance associated with a construct. It is unfortunate that

while there has been substantial growth in transfer of paper-and-pencil scales to computerized versions, the actual measurement advantages gained from innovative items have not yet, to date, been well established. It is thus no surprise that the majority of the literature has focused on the equivalence of paper-and-pencil and computerized versions, and not the potentiality of the computerized or online environment to offer alternatives to item format and design, and subsequent psychometric comparisons with paper versions. Issues such as the establishment of measurement equivalency across modes, while having generated a considerable amount of attention (e.g., Mead & Coussons-Read, 2002; Ployhart, Weekley, Holtz, & Kemp, 2002) have predominantly focused on confirming criterion-related validity. Thus, new types of multimedia designed innovative items that allow for the creation of scenario-based or virtual world environments, have yet to be significantly investigated, let alone psychometrically analyzed.

Therefore, the investigation presented in this thesis developed these two interrelated issues of personality research, namely, the way personality responses are best modeled, and the most optimal mode by which personality items are presented as well as associated modal issues.

Purpose of the Research

Based on the research conducted in the meta-review and the two empirical studies in this thesis, the following three propositions are introduced. The first proposition is that the ideal point model will provide better fit to response patterns derived from an ideal point constructed personality scale than those generated from the dominance-based graded response model (GRM). While comparison between these two modeling approaches have been previously investigated (Chernyshenko et al., 2001; Chernyshenko, 2002) these studies compared IRT models to responses from scales developed using theoretically aligned construction procedures (e.g., ideal point modeling of responses from an ideal point developed scale). In addition, these studies have dichotomized polytomous responses and compared these to dominance-based IRT models, usually the 2PL model. This study is the first to compare the model-data fit performance of these two approaches to a scale constructed specifically using an ideal point approach. Thus, given that both response modeling and scale construction are aligned, it is reasonable to presume that resultant model-data fit produced by the ideal point generalized graded unfolding model (GGUM) will be superior to that provided by the dominance model (GRM).

The second proposition is associated with the web delivered interactive scenario-based innovative items. Here is it proposed that, in comparison to the original paper-and-pencil versions of these items, the innovative items would provide more measurement efficiency across all proficiency levels captured by the scale. Specifically, it is proposed that the degree of information gained via the scenario-based items will be greater at *both* an item and test level, when compared to the paper versions. Jodoin's (2003) comparison of innovative items to paper-and-pencil

versions is the only previous example of an empirical investigation into the information provided by these items. In addition, it is the only study using the IRT approach to assessing measurement efficiency. Although Jodoin's study related to cognitive-ability items, it is proposed that like his study, innovative items will provide more information at all levels across the proficiency continuum than their paper-and-pencil counterparts.

In addition to examining the measurement efficiency of the innovative items administered, the final proposition of this thesis relates to the perceptions of the test-takers to the interactive scenario-based item design. Here it is proposed that the test-taker will prefer the innovative version of the Reader Self-Perception Scale (RSPS) Progress scale. Thus, due to the engaging and interactive nature of the items, it is anticipated that the test-takers' experience and involvement in the scenarios will be the primary focus, with the actual items blending in to the onscreen environment. As such, it is proposed that test-takers will less aware of the testing nature/intent of the innovative items. Previous research has not, to date, determined the attitudes of test-takers to interactive multimedia items, particularly, their preferences to these in comparison with paper-and-pencil versions.

Significance of the Research

Although a reasonable proportion of personality research has been directed at developing and validating measures and taxonomies, there has been less research examining the most optimum approach to scale construction and response modeling for these constructs. As the theoretical underpinnings of dominance-based models does not support the *typical* response behavior proposed to be exhibited when

responding to a non-cognitive ability scale, it is of significant import to establish whether the ideal point assumptions are a valid, if not preferable, approach to both scale construction and modeling of data. At a model level, it is important that the correct approach to item estimation is adopted in order to provide accurate item-person estimates. Similarly, at the scale development level, a compelling reason for adoption of an ideal point approach lies in the construction method where items are included that lie at various points along a trait continuum. As dominance-based (e.g., Likert) approaches to scale development select essentially extreme items (either high or low on the trait continuum), the ideal point approach would result in more effective use, and thus less redundancy, of items. While previous research (e.g., Chernyshenko et al., 2001) has compared estimations produced by models from their correspondingly aligned scales, no previous research has examined the modeling capabilities of a dominance IRT model to response data generated from an ideal point constructed scale.

In addition to understanding response behavior, it is important to appreciate the impact that the mode of test delivery might have on the test-taker. The thesis includes the first literature review of research on the mode effect that focuses on both the human and technological factors associated with changing the test administration from paper to computer. Most research has focused on the human factors associated with the mode effect (such as computer familiarity and anxiety), with almost no focus on the technological user interface (UI) and presentation issues that are an inherent part of a computer-based test. Across all of these mode issues, even less focus has been applied in relation to non-cognitive ability tests, such as personality and attitude measures. Such research is particularly important given the increasing and expected

future movement towards computerized cognitive and non-cognitive assessment. Obviously, the premise of any measurement theory is that through the use of robust scale construction and development procedures, valid and reliable assessments of latent constructs can be achieved. Although there has been a vast array of empirical and psychometric analysis to support many paper-and-pencil personality measures, it is unreasonable to assume that these are applicable for onscreen assessment. Instead, a new history must be established for computerized items, where similarities and differences with paper measures are investigated at both a user and psychometric level.

In addition to these mode issues, a review of many currently applied innovative items is provided in Chapter 3. This review highlights that only a small amount of research has focused on the psychometric analysis of innovative items. In particular, the measurement efficiency derived from the various innovative item designs has only the focus of one study to date (see Jodoin, 2003). If the full design possibilities of innovative items are to develop and evolve, a body of research needs to be established regarding the psychometric properties of these items.

The three studies in this thesis were developed from recent arguments surrounding non-cognitive ability data, reviewing test administration effects, and empirically measuring competing modeling approaches, and innovative item measurement efficiency. Although the first issue related to the way personality responses are modeled, and the second focused on the mode by which personality items are presented, both were concerned with increasing the amount of measurement information derived from personality scales. It is proposed that given the substantial increase in the use of attitude and personality measurement tools across many areas of

industry and academia, the areas focused in this thesis are highly relevant for the development of more sophisticated and arguably more appropriate attitudinal and personality measurement approaches.

Design of the Research

The next chapter presents a critical review of the literature associated with the constructs, approaches, and models used in the two empirical studies in this thesis. First, an overview of the constructs that underpin self-worth strategies was explored, along with the findings of the research that focused on these strategies. In addition, the competing IRT models were presented, together with arguments regarding the type of response behavior that this elicited when test-takers respond to non-cognitive ability scales. The second section of the literature review focused on the construct, design, and analysis relating to the second empirical study of this thesis. The work of Bandura (1977, 1982) was reviewed in relation to the construct of perceived self-efficacy and the approaches that should be adopted when attempting to measure self-efficacy. As the second empirical study involved the design of a web-based scale, previous approaches to innovative item design and the associated use of technology was reviewed in relation to computerized and web-based scales. Finally, literature relating to the psychometric analysis of paper and computerized tests are presented, with focus given to previous studies relating to equivalence and efficiency analysis of computerized and paper-based measures.

In Study One, a personality measure was developed following an ideal point approach to scale construction, and the responses from this scale were used to compare the model-data fit provided by two theoretically diverse approaches to

modeling non-cognitive data. This study was designed to compare the effectiveness of an ideal point model with a traditional dominance-based model.

Study Two aimed to examine issues relating to the transfer of paper-and-pencil measures to screen. The literature review presented research based on its contribution to the human and technological issues related to the mode effect. In addition, a related review of innovative items was presented, examining specific modal differences and design issues (e.g., response actions).

Study Three involved a series of web-based interactive scenario-based items that were developed to empirically assess their psychometric properties in comparison to a paper-and-pencil version. Using the latest multimedia development tools available at the time of this study, nine innovative items were developed that combined various videoed objects in a graphical environment. In order to create a virtual reality experience for the test-taker, the questionnaire's structure was designed around the context of a story involving an upcoming reading test at school. Both the web-based and paper-and-pencil items were analyzed in order to establish the degree of measurement efficiency occurring from each version.

The final chapter in this thesis presents a discussion of the overall findings of the research in relation to the propositions presented here, the significance of the three studies, and the implications for future approaches towards personality testing.

CHAPTER TWO

LITERATURE REVIEW

This chapter presents a review of the various theories, constructs, and measurement approaches that were used across Study One and Study Three in this thesis. In relation to the first study, the self-worth strategy constructs that were adopted for the construction of the ideal point scale are reviewed within the context of academic ability and performance. This is followed by a review of the response behaviors that are proposed to occur when test-takers are responding to non-cognitive ability constructs such as self-worth. Item response theory is then outlined, with focus given to the two different response modeling approaches, namely graded response model (GRM) and generalized graded unfolding model (GGUM) that were applied in Study One. A critical review is then presented outlining previous applications of ideal point and traditional item response models to personality measures.

The final topics in this chapter are related to Study Three in this thesis. Reading self-efficacy is outlined, along with the framework and design approach used in developing the innovative items for this study. Due to the multimedia nature of the innovative items, special attention is given to literature on multimedia testing. The last two sections of this chapter focus on previous evidence regarding the validity and reliability of innovative items.

Self-worth Strategies

Self-worth is a latent construct that refers to an individual's sense of worth; that is, the degree that an individual accepts himself or herself as a worthwhile person (Covington, 1992). By extension, self-worth strategies refer to the lengths that an individual will go to avoid failure that is interpreted as being indicative of low ability, and then equated with low self-worth. Previous research (Garcia & Pintrich, 1994; Martin, 1998; Martin, Marsh, & Debus, 2001) has found that self-handicapping and defensive pessimism are two key strategies adopted by individuals to protect their self-worth. In relation to the academic realm, Martin (1998) validated that the higher order factors of defensive expectations, reflectivity, and self-handicapping effectively represented the construct of academic self-worth. Further, Martin found that active and self-presented dimensions were important aspects of the self-handicapping and defensive expectations factors.

Self-handicapping

First coined by Jones and Berglas (1978), self-handicapping refers to the deliberate obstacles or impediments that an individual will enlist in order to impair their performance on a task or activity. In other words, an individual purposefully partakes in behavior through which they can sabotage their upcoming performance. Although self-sabotage, as a behavior, appears to be counter-intuitive to survival, self-handicapping is by its very nature a survival technique or strategy. By the individual purposely choosing not to partake in a task for which they feel that they may not succeed, individuals protect themselves from the self-deprecating effects of failure on their self-esteem. By extension, self-handicappers typically are also strongly concerned about how a possible performance failure may be perceived by others.

Numerous researchers (e.g., Covington, 1992; Garcia, et al., 1995; Levesque, Lowe, & Mendenhall, 2001; Martin, Marsh, & Debus, 2001; Midgley, Arunkumar, & Urdan, 1996; Midgley & Urdan, 1995) have argued that in addition to self-handicapping being a self-worth preserving strategy, it also acts as a public image management strategy. Within the academic environment, Covington (1992) posits that some students will adopt self-handicapping strategies in order to negate any possibility of being labeled as stupid by their peers or teachers. A self-handicapping student will actively present to others a perception of themselves and their circumstances that accounts for (typically lower) performance outcomes (Baumeister & Scher, 1988; Midgley, Arunkumar, & Urdan, 1996). For example, attributing a poor exam result to an external cause, such as an after school job commitment, allows the individual to "save face" by deflecting any possible internal reasons for poor performance (e.g., lack of ability) to a presented "beyond their control" external source. When self-handicappers internalize their poor performance, excuses are presented as relating to non-ability issues such as stress/test anxiety, fatigue, disinterest, or ill health (Martin, Marsh, & Debus, 2001; Midgley, Arunkumar, & Urdan, 1996, Thompson, 1994). As such, the strategy of self-handicapping can be effectively viewed as comprising of three facets: active, self-presentation, and self-presented affective self-handicapping.

Active self-handicapping

This aspect of self-handicapping occurs when an individual purposely creates or enlists an actual obstacle or impediment to hamper their upcoming performance scenarios. This strategy allows an individual to redirect the cause of academic failure away from their own perceived academic ability, thus protecting their academic self-worth. As Rhodewalt and Davison (1986) posited, self-handicapping allows the

individual to externalize ability attributions because of their adoption of an equally plausible performance-inhibiting cause (e.g., the handicap). Within the academic setting, typical performance-deliberating handicaps adopted include diverting effort into non-related activities (e.g., socializing), letting themselves get physically run-down (e.g., lack of sleep, not eating properly), over-committing time outside school, and putting assignments or exam preparation off until the last minute (e.g., procrastination) (Martin, 1998).

Self-presented self-handicapping

Earlier research (e.g., DeGree & Snyder, 1985; Hirt, Depe, & Gordon, 1991; Smith, Snyder, & Perkins, 1983) has contended that individuals will purposefully present to others obstacles or impediments that have prevented their success or impacted on their ability to perform well. While self-presenting self-handicappers are motivated to present themselves as being at a distinct disadvantage to the potentiality of success, these individuals will typically not partake in active self-handicapper behavior, instead only manifest such scenarios to others. As a result, individuals are able to present themselves as possessing significant levels of academic ability, despite the obstacles that they have endured. When actual obstacles do naturally exist in the individual's environment, the impacts from these are exaggerated by the self-presented self-handicapper. In the possibility of performance failure, the presented hindrances act as legitimate excuses, and as such, distance the reason for poor performance away from the individual's academic ability.

Self-presented affective self-handicapping

As an extension to self-presented self-handicapping, Martin (1998) proposed that an individual's affective state, rather than obstacles or impediments, also represents a form of self-presented self-handicapping. Thus, the individual uses internal states such as fear, anxiety, boredom, frustration, or anger, as an excuse for poor potential performance. As with self-presented self-handicapping, the affective state may or may not be present in the individual at the time. Where affective states are legitimately being experienced by the individual, the state will be greatly embellished beyond its actual true proportions.

Defensive Pessimism

Norem and Cantor (1986) described defensive pessimism as a strategy whereby individuals set themselves unrealistically low expectations in order to circumvent any anxiety that they attribute to their performance. As opposed to self-handicapping, defensive pessimists do not intentionally sabotage their behavior, as the low expectations do not become self-fulfilling. Thus, this strategy allows the individual to protect their academic self-worth. Norem and Cantor argued that this self-protection strategy allows the individual to not only remove or lessen performance-derived anxiety, but also to redirect this anxiety affect to a motivation effect.

To date there is scant research regarding defensive pessimism, with the majority of development in this area originating from Norem, Cantor and their colleagues. Their research has suggested that defensive pessimism consists of three underlying factors: active defensive expectations, self-presented defensive expectations, and reflectivity. In addition, more recent research by Martin has extended these factors to include a self-presented dimension to defensive expectations.

Active Defensive Expectations

Individuals adopting this strategy lack confidence in their academic ability. Regardless of how well they may have done in previous performances, the overwhelming fear of failure experienced by these individuals result in them setting low expectations for their future academic performance scenarios (Martin, 1998). By doing so, the anxiety associated with the fear of failure for these individuals lessens considerably as the minimal academic requirement set is highly achievable, thus lessening significantly the anxiety and pressure that they put on themselves.

Self-Presented Defensive Expectations

Whereas individuals who adopt defensive expectations as a self-worth protection strategy internalize this approach, some individuals work hard to portray to others the low expectations that they have of their ability to perform well. Suggested by Martin as an additional type of defensive expectation, these individuals externalize their defensive expectations regarding upcoming academic performance by presenting themselves as being in a worse academic situation than is actually the case. As these individuals will nearly always in the short-term, out-perform their presented abilities, they partake in this form of impression management to cushion the potentiality of failure, in turn avoiding the disapproval of others.

Reflectivity

This dimension of defensive pessimism reflects the thinking-through behavior adopted by individuals regarding their upcoming performance outcomes. Thus, this dimension is not associated with any active destructive behavior or impression management of their abilities, rather reflectivity involves individuals simply reflecting the potential outcomes of their upcoming performance. These individuals do not

necessarily just focus on the potential negative outcomes, the realities of positive outcomes are also reflected upon. As such, the individual protects the possibility of failure by thinking through how they would feel if they did not perform well, and similarly, how they would feel if they obtained a successful result.

Response Behaviors

Recent research has suggested that individuals may exhibit different behaviors when responding to personality items than when they respond to cognitive ability items (Chernyshenko, 2002; Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Maydeu-Olivares, 2005; Roberts, Laughlin, & Wedell, 1999). Intuitively it seems reasonable to assume that the relatively stress-free, lesser time pressures, no right or wrong answer environment that most individuals experience when completing a personality scale differs markedly from the intense conditions arising from a cognitive-ability test environment. The focus of cognitive testing relates more to whether a test-taker has acquired a particular knowledge or has a capability or skill, whereas the focus for personality testing is on the "internal appraisals" that the individual has of themselves or some attribute, and for which they arguably are often the "expert". The question arises: Do such vastly different test conditions and focus result in equally different responding behavior?

Cronbach (1949) proposed that test-takers did behave differently when responding to the different content requirements of these two types of measures, and referred to this distinction as *maximal* and *typical* behavior. Maximal behavior is exhibited when, knowing the standards being used to evaluate their performance, test-takers are motivated to perform to the best of their abilities in order to maximize their

scores, as in tests of attainment, ability, aptitude, and skills (Sternberg & Kaufman, 1998). In contrast, respondents display typical behavior (e.g., behavior typical of the person at the time of the test), where little intellectual effort is required by respondents to items, as in measures of personality, interests, and attitudes.

Based on these behaviors, Thurstone (1928, 1931) and later Coombs (1964) proposed that when responding to a personality item there is one of two types of response processes occurring, specifically, a *dominance* (cumulative) response or an *ideal point* response. Coombs (1964) described a dominance response process as one where the probability of a correct response increases as a function of the individual's trait position. Applied to non-cognitive ability scales, an "agree" (or a "strongly agree") response increases as a function of the individual's location on the trait continuum. Thus, there is an assumption of a monotonic relationship between an individual's trait level and their scores, where test-takers with greater levels of a trait, will have higher scores than test-takers with lower levels of the trait. This approach is facilitated by the Likert scale development and construction approach, where a cumulative mechanism is also assumed. Conversely, as the name suggests, assumptions underlying an ideal point response process suggests that an individual has an ideal point position on the trait continuum. Therefore, an individual's score is related to their response's proximity to the trait being measured by the item. Specifically, high trait scores are associated with individuals with trait levels closest to that being measured by the item.

While there have been numerous studies that have applied various IRT dominance models (e.g., one-parameter logistic (1PL), two-parameter logistic (2PL), three-parameter logistic models (3PL), and graded response model (GRM) to

personality scales (e.g., Ellis, Becker, & Kimmel, 1993; Fraley, Waller, & Brennan, 2000; Harvey & Murry, 1994; Reise & Waller, 1990), little research has investigated the appropriateness of these cumulative models. The subsequent section gives a brief introduction of IRT, and a description of the graded response and ideal point models employed in this study.

Measurement Theory

Item Response Theory (IRT)

IRT is a modern measurement approach that relates the characteristics of items (item parameters) and characteristics of individuals (latent traits) to the probability of providing a particular response (Hambleton, Swaminathan & Rogers, 1991).

One of the advantages of IRT is that both item (b = difficulty, a = discrimination, c = pseudo-guessing) and person parameters (θ = proficiency) are invariant. This independence means that both item and person parameters are neither reliant on the subset of items, nor on the distribution of the latent trait in the population of respondents respectively (Stark, Chernyshenko, Chuah, Lee, & Wadlington, 2001). In addition, due to the relationship between item and person estimations, both parameters are on the same metric. This permits direct comparisons between the trait or ability level required by the item, and the level of that characteristic in the individual (Scherbaum, Finlinson, Barden, & Tamanini, 2006). While IRTs potential for solving problems in testing and measurement is well recognized, the advantages of its use can only be realized when a fit exists between the IRT model of choice and the test data of interest (Hambleton, 1989).

Graded Response Model (GRM)

Widely applied, the GRM has been become one of the best known and widely applied IRT models for polytomous (e.g., Likert) response models (Hambleton et al., 1991). Samejima (1969) developed the GRM as an extension of the 2PL model to allow for the analysis of items with polytomous ordered response categories, where parameters are estimated for *m* ordered response options. Under this model, analysis identifies the relationships between the item or option parameters, the person parameters, and the particular option that has been selected (Scherbaum, Finlinson, Barden, & Tamanini, 2006). The category response function $P\left(v_i = k \middle| \theta = t\right)$ is the probability of the response *k* to item *i* defined as

$$P\left(v_i = k \middle| \theta = t\right) = \frac{1}{1+\exp[-1.7a_i\left(t-b_{i,k}\right)]} - \frac{1}{1-\exp\left[-1.7a_i\left(t-b_{i,k+1}\right)\right]}, \qquad (1)$$

where v_i = an individual's response to the polytomously scored item *i*; *k* is the particular option selected by the respondent ($k = 1, \ldots, s_i$, where s_i is the number of options for item i); a_i is the item discrimination parameter, which in this model is assumed to be constant for each option within an item; *b* is the difficulty parameter, which varies across each option given the constraints $b_{k-1} < b_k < b_{k+1}$, and b_{Si+1} is taken as $+\infty$ (Chernyshenko et al., 2001). Thus, it is assumed that given the ordered response set under the graded response model the latent trait value is smaller for test-takers that respond "strongly disagree" than it is for test-takers that respond "disagree". This assumption highlights the dominance response process that underlies this model (Scherbaum et al., 2006).

GRM item parameters can be estimated using MULTILOG 7 (Thissen, Chen, & Bock, 2003) computer program, which employs optimal full information marginal maximum likelihood (MML) item estimations procedures by the use of the EM algorithm (Bock & Aitkin, 1981) (see Thissen et al., 2003 for further details).

Ideal Point Item Response Theory

Thurstone (1928, 1931) posited that an ideal point approach assumes that respondents endorse items based on how closely they believe that the item reflects their own position (e.g., their ideal point). Based on the premise of proximity, an "agree" (or a "strongly agree") response is determined by the extent that their own position (e.g., attitude, viewpoint) is reflected by the content of the item. In other words, if an item is positioned below (more negative than the individual's attitude) or above (more positive than the individual's attitude) the individual's position on the trait continuum then the probability increases that the individual will disagree with the item (Roberts, Donoghue, & Laughlin, 2000). As such, the *typical* performance modeled under ideal point assumptions produce nonmonotonic response functions that have a single peak and are symmetric about the origin $\left(\theta_j - \delta_i\right) = 0$, where θ_j denotes the location of the *j*th individual on the continuum, and δ_i denotes the position of the *i*th item on the continuum. (Roberts et al., 2000). From Thurstone's (1928, 1931) initial ideal point procedures, Coombs (1964) coined the term *unfolding* to represent the process of locating items and respondents' positions on the trait continuum. While Roberts et al. noted that numerous parametric (Andrich, 1996; Andrich & Luo, 1993; Desarbo & Hoffman, 1986; Hoijtink, 1990, 1991; Roberts et al., 2000; Verhelst & Verstralen, 1993) and nonparametric (Cliff, Collins, Zatkin, Gallipeau, & McCormick, 1988; van Schuur, 1984) unfolding models have been devised, the

Generalized Graded Unfolding Model (GGUM) is the only parametric model for graded responses that allows the discrimination parameter (α_i) of the item to vary, and thus permit response category threshold parameters to vary across items (Roberts et al., 2000). As such, the GGUM leads itself to be compared to other IRT models where varying α_i parameters are a feature.

Generalized Graded Unfolding Model (GGUM)

The estimation approach of the GGUM represents an ideal point approach to analyzing response behavior. The GGUM has been chosen for analysis in this study as it represents the most generalized and flexible model from the numerous parametric and nonparametric options available. In particular, item estimations generated under the GGUM permits both the subjective response category threshold parameters (τ_{ik}) and discrimination parameters (a_i) to vary across items. In addition, this model permits analysis of both dichotomous and polytomous item responses. Item parameters under this model are estimated using a marginal maximum likelihood (MML) approach (Bock & Aitkin, 1981; Bock & Lieberman, 1970), and person parameters estimated using an expected a posteriori (EAP) procedure (see Roberts et al., 2000 for further details). In combination, the *typical* performance modeled under ideal point assumptions produce nonmonotonic response functions that have a single peak and are symmetric about the origin $\left(\theta_j - \delta_i\right) = 0$, with the shape of the item's function being a product of its a_i and the relative location of the subject response category (SRC) for that item (τ_{ik}) (Roberts et al., 2000). Thus, this model is defined as

$$P[Z_i = z \mid \theta_j] = \frac{\exp\left(\alpha_i[z(\theta_j - \delta_i) - \sum_{k=0}^{z}\tau_{ik}]\right) + \exp\left(\alpha_i[M - z](\theta_j - \delta_i) - \sum_{k=0}^{z}\tau_{ik}\right)}{\sum_{w=0}^{C_i}\left[\exp\left(a_i[w(\theta_j - \delta_i) - \sum_{k=0}^{w}\tau_{ik}]\right) + \exp\left(\alpha_i\left[(M_i - w)(\theta_j - \delta_i) - \sum_{k=0}^{w}\tau_{ik}\right]\right)\right]}, \quad (2)$$

where α_i = the discrimination index of the *i*th item, δ_i = the location of the *i*th item on the latent continuum, θ_j = the location of the *j*th individual on the latent continuum, τ_{ik} = the relative location of the *k*th subjective response category threshold of the *i*th item, z_i is an observed response to the *i*th item with $z = 0, 1, \ldots, C_i$, $z = 0$ representing the strongest level of disagreement and $z = C_i$ representing that strongest level of agreement for the *i*th item, and $M_i = 2C_i + 1$.

Item parameters were estimated using the windows-based GGUM2004 computer program (Roberts, Donoghue, & Laughlin, 2000; Roberts, Fang, Cui, & Wang, 2006). In applicable simulations using GGUM2004's predecessor GGUM2000, Roberts et al. found that accurate MML item parameter estimations and EAP estimations occurred when sample sizes consisted of a minimum of 750 respondents, and 15 to 20 items were present. However, Roberts et al. also found that 10 to 15 items resulted in slightly higher error levels, and suggested that unless absolute θ_j accuracy was necessary, a minimum of 10 items would suffice where relative accuracy was required, or when test efficiency is of greater concern.

Applications of Ideal Point and Item Response Approaches to Personality Items

Previous studies have indicated that ideal point models provided good fit to attitude items (even when the items were not intentionally developed to fit this model). For example, using an ideal point approach towards a scale measuring

attitudes towards abortion, Roberts et al. (1999) found that although item characteristic curves (ICC) showed monotonically increasing (or decreasing) response functions for extreme positive (or negative) statements, more moderate and neutral items displayed a typical single peaked nonmonotonic response function. Interestingly, items whose content reflected the most extreme positive or negative positions on the trait continuum displayed nonmonotonically shaped response functions (Roberts et al., 1999). Given the overall reasonable representation of monotonically increasing functions across the majority of attitude items, Roberts et al. suggested that the decision to follow either a Likert or Thurstonian approach might be in part dictated by the type of individual (e.g., those with extreme attitudes) that the researcher may want to identify.

Chernyshenko et al. investigated how well a selection of parametric models (2PL and 3PL models, and GRM) typically used for analysis of personality scales fit the data when compared to the fit provided by Levine's (1984) nonparametric MFS model. Analysis of data from the Sixteen Personality Factor Questionnaire (16PF) and the Big Five Personality scale revealed that overall none of the parametric dominance models provided adequate fit. In comparison, the more complex nonparametric MFS model provided fewer misfits across both personality scales. Although Chernyshenko et al. highlighted some issues that may have impacted on the poor performance shown by the parametric models; they argued that the results probably reflected the inappropriateness of applying dominance-based models to non-cognitive ability based items. Thus Chernyshenko et al. concluded that the good fit provided by the MFS nonmonotonic function indicates that an ideal point response process may underlie non-cognitive responses processes.

Maydeu-Olivares (2005) challenged this recommendation, and suggested instead that a distinction between the measurement of *personality* and *attitude* may exist. While an ideal point approach may be appropriate for attitudinal scales, Maydeu-Olivares suggested that the traditional dominance-based process might be more appropriate for personality-based scales. Specifically, when respondents are required to ascertain the degree to which personality-based description applies to them, then an individual endorsement of an item will represent their standing at that theta level or higher, as with cognitive-based responses (Maydeu-Olivares, 2005). In addition to using the models and methodology used in the Chernyshenko et al. study, Maydeu-Olivares also included Bock's nominal model, Masters' partial credit model, an extension of Masters' partial credit model (Thissen & Steinberg, 1986), and a normal ogive model version of Samejima's (1969) graded response model (GRM) using the limited information estimation. Results showed that the GRM outperformed all the other models, with the full information version providing the best fit. Although not as successful as the limited and full versions of the GRM, the large number of parameters estimated by the MFS model provided reasonable fit to both scales, which concurs with the findings from the Chernyshenko et al. study.

The good fit provided by the GRM, however, may be a reflection of the scale's construction procedure, rather than a reflection on the appropriateness of dominance IRT models to personality response patterns. Where the Likert procedure to scale construction seeks to avoid statements reflecting either neutral or extreme positions, the ideal point method requires that all locations on the trait continuum are represented. Given that previous applications of an ideal point model has only ever been applied to scales developed based on a dominance approach (e.g., Likert) it

seems more appropriate that the effectiveness (or otherwise) of the ideal point approach should be applied to a personality scale constructed given the same theoretical assumptions.

Chernyshenko (2002) provided the only investigation whereby an ideal point model was applied to a personality scale constructed using ideal point assumptions. Using three different scale construction procedures (e.g., dominance classical test theory, dominance IRT, and ideal point IRT), Chernyshenko constructed three six-facet measures of conscientiousness. Comparing the fit of each conscientiousness scale to its respective model, he found that the ideal point approach showed greater fit to the ideal point constructed scale than did the other two approaches to their respective scales. Across the three scales, items from the ideal point scale provided more information, hence, greater measurement precision.

Reading Self-efficacy

The construct of interest in the second empirical study is grounded in Bandura's (1977, 1982) theory of perceived self-efficacy, and applied in relation to reading progress. Bandura (1993) defined self-efficacy as beliefs that "influence how people feel, think, motivate themselves and behave" (p. 118). It is proposed that an individual's actions are predetermined by the beliefs that they have in their capability to exercise control over their functioning (Bandura, 1993; Lynch, 2000). Thus, these beliefs will either inhibit or motivate the individual, as "unless people believe that they can produce desired effects by their actions, they have little incentive to act" (Bandura, Barbaranelli, Caprara & Pastorelli, 1996 p. 1206). In relation to reading behavior, Henk and Melnick (1995) argued similarly that the perception an individual

has regarding their reading ability would influence the degree to which they are motivated to read, and the effort and persistence given when processing and comprehending text. However, it is important to note that self-efficacy cannot be viewed as a universal concept, instead perceptions, even with the same subject area, can be highly contextual and context specific (Bandura, 1993). Within reading itself, individuals might have a high sense of efficacy regarding their ability to comprehend text, but perceive a lack of ability when recognizing words. Given the potential variability of an individual's self-efficacy at levels within a task, Bandura (1986) recommended that self-efficacy measures should not be aimed at capturing a general efficacy towards an area, rather, be targeted at specific behaviors or tasks (Mathewson, 1994).

Innovative Item Design - Framework

This section discusses the structural definition of innovative items, and the classification framework adopted in relation to the scenario-based innovative item design developed for the second empirical study in this thesis (see Chapter 5).

To date, the term *innovative item* has encompassed a broad and diverse range of item designs that often, only share as their commonality, the fact that they are computerized. As the literature review in Study 2 highlighted, the term innovative item has been applied to such items ranging from essentially static passage/text editing (e.g., Breland, 1999; Parshall, Davey, & Pashley, 2000) to interactive video (IAV) computerized assessment (Dyer, Desmarais, Midkiff, Colihan, & Olson, 1992; Drasgow, Olson-Buchanan, & Moberg, 1999). This diversity indicates that the term is highly generic, encompassing designs that differ vastly in terms of presentation and

test-taker interactivity. Drasgow and Mattern (2006) argued that establishing whether an item's design is innovative or not is too simplistic, suggesting instead that Parshall, Davey, & Pashley's (2000) framework for innovative items is used to classify computerized items. This framework classifies items based on five specific design aspects: item format, response action, media inclusion, level of interactivity, and scoring algorithm. Using this framework, the key features of the scenario-based test design developed in this thesis are presented.

Item format refers to the type of response that is required by the participant, for example, the test-taker might have available a selection of responses to select from, or be required to construct a response, e.g., type a short answer (Parshall et al., 2000). The interactive scenario-based items presented a selected response design, whereby test-takers chose from the answer options presented to them. At the end of every dialogue, one of the screen actors delivered the item (in the form of a question) directly to the test-taker. At that point, a flash animated response form appeared on the screen, displaying the question that had just been asked at the top of the form, with the response options ("Yes"/"No") positioned underneath. This response form required the test-taker to perform a physical action to respond to an item. Parshall et al. defines the *response action* feature of an innovative item as relating to the input devices that test-takers must use (e.g., keyboard, mouse, joystick, and trackballs) to record their response. Thus, for the scenario-based items, the test-taker was required to use their mouse to select one of the radio buttons associated with either the "Yes" or "No" response option. Once the response option was selected, the test-taker then needed to click the "submit" button to proceed to the next scene. In addition, as part of the interactive component of these items, the test-taker used the mouse to move the

pointer to different locations on the screen (and beyond the initially visible screen). Further, in order to interact with the various groups of students or individuals in view, the test-taker was required to move their mouse pointer over the onscreen characters, and click to initiate dialogue.

As the name suggests, Parshall et al. refer to *media inclusion* as the various material that is included in an item that is in addition to text-based material. The scenario-based items in this thesis included audio, graphics, and video. Using a range of background designs (e.g., outdoor school environment, school lobby, school corridors, classrooms), graphical objects (e.g., desks, chairs, whiteboards), and associated audio, each scene was designed to create a virtual environment that supported the storyline context. In addition, actors (e.g., students, teachers) were videoed, keyed and placed within the virtual graphical environment.

Parshall et al. refer to *interactivity* as the degree that an item is designed to adapt and/or engage a test-taker. Low interactivity is typified when test-takers simply select a response to complete an item. In contrast, high interactivity occurs when, as in the United States Medical Licensing Examination, a test comprises a simulation (e.g., a clinical scenario), which requires the test-taker to use the information to construct a solution/answer. In the area of personality testing, a high degree of cognitive complexity is not, and should not, be required to respond to an item. Here the degree of interactivity might be better referred to as the degree to which the test-taker interacts with the environment presented. Thus, interactivity is related to the test-taker's experience and responding actions, rather than complex ability-based activities. The scenario-based items were designed to present a highly engaging and interactive experience for the test-taker. Within each scene, there exist numerous

dialogues (e.g., groups of students, single students) that the test-taker listens and responds to. A conversational style discourse was written, typically consisting of either short conversations (one or two dialogues), or longer conversations (four to five dialogues), delivered from two to four screen actors. During this discourse, these screen actors interacted with other screen actors and directed the conversation towards the test-taker. Although the test-taker is a passive listener to these dialogues (i.e., does not respond during the conversation), the inclusive style of discourse was adopted to give the test-taker as much a feeling of involvement in each scene as possible. In addition to responding to the questions posed by the screen characters, several graphical objects and environments could be interacted with by the test-taker. For example, at the beginning of each scene, the test-taker was able to select which group of students (or student) they wished to interact with by clicking on that area. In addition, some of the background objects moved or changed state when clicked (e.g., fluorescent light turns on and off). Non-intrusive cue points were used to indicate that these objects could be interacted with. This extra interactivity was included to provide some fun and engagement through the progression of the questionnaire, thus replicating a feature of many of the virtual world gaming applications commercially available.

Since progression through each of the scenes was fixed, in other words, test-takers' responses to each of the items did not impact on the subsequent progression to the next scene, the *scoring algorithm* devised for the web-based questionnaire simply recorded the test-taker's demographic information and responses through the test. However, given that test-takers could elect to interact with students in any order

within each scene, the scoring algorithm was written to capture the responses specific to an item identity, rather than sequential order.

Multimedia Testing

As the previous literature review indicates, the capabilities of computerized testing present numerous opportunities for test developers to include various graphical aspects into items (Drasgow & Mattern, 2006). To date, however, only a few published studies present tests that have been created using a combination of multimedia features such as integrated audio, computer graphical technology and video. The most closely aligned innovative approach to the interactive scenario-based items in this study is the interactive video (IAV) computerized assessment tool (e.g., Drasgow, Olson-Buchanan, & Moberg, 1999; Olson-Buchanan et al., 1998). IAV items consist of video and sound, whereby the test-taker watches a video clip and at certain points during the clip is asked questions regarding the construct of interest. Certain items throughout the test are adaptive in that the test-taker's response dictates which video clip will be next presented (Drasgow et al., 1999).The most early example of an IAV assessment is the Workplace Solution test (Desmarais et al., 1992; Desmarais, Masi, Olson, Barbera, & Dyer, 1994; Dyer et al., 1992; Midkiff, Dyer, Desmarais, Rogg, & McCusker, 1992) where situational judgment skills within the workplace were assessed using a series of 30 scenes set in the context of a fictional organization. Another organizational tool, Allstate Multimedia In-Basket (Ashworth & McHenry, 1992) used video clips representing the context of an airline customer service department to assess potential applicants' ability to conduct various tasks and display key knowledge and skills identified for the position. In addition to videoed instructions from the fictional "boss", test-takers interact with various graphics

onscreen (e.g., filing cabinets), in order to complete the required tasks (McHenry & Schmitt, 1994). Unfortunately, no psychometric information is available for either the Workplace Solution test or the Allstate Multimedia In-Basket assessment applications.

One of the most recent and sophisticated uses of the IAV approach was in the development of the Conflict Resolution Skills Assessment (CRSA; Olson-Buchanan et al., 1998). Based on the KO (Keenan & Olson, 1991) model of conflict resolution, the stimulus for this measure consisted of the presentation of a scene where a conflict was taking place (typically 1-3 minutes in duration). At significant junctions during the scene, the test-taker was presented with several options for resolving the conflict, and based on the response given, the conflict scene continued to unfold. Thus, the IAV is adaptive in that certain scenes progress based on the test-taker's response to the previous video clip. While the internal consistency of the test was high (r = .85), and no adverse impact was found amongst test-takers, only a modest correlation (r = .14) was found between the criterion (derived from a composite of on-the-job performance ratings from supervisors) and the CRSA (Olson-Buchanan et al., 1998). Thus, although Olson-Buchanan et al. acknowledged that further work was required to understand the results found in this study, especially in relation to validity, they did argue that the increased authenticity created with the interactive video aspect was an important design concept to pursue and develop.

Equivalence Testing

Although the studies above have specifically set out to design assessment tools that are not equivalent to traditional paper measures, typically computer- and web-

based measures have been designed to replicate in format their paper-and-pencil versions. However faithfully a measure might be replicated onscreen, various mode effect issues, at both a human and technological level, that might result in different cross-mode performance. Understanding the reality of the naturally non-equivalent physical dimensions between the two modes (e.g., physical interface and difference in interactivity/flexibility), together with the various psychological (e.g., computer familiarity, anxiety, and social desirability), and cognitive (e.g., memory and comprehension) participant characteristics that result, researchers have been motivated to establish equivalence between computerized and paper-and-pencil versions. As Buchanan (2003) argued, there must be evidence of equivalence between paper-and-pencil and web versions of measures, rather than the *assumption* that it exists. Lievens and Harris (2003) asserted that equivalence research has essentially fallen into three categories: differences in data collection of psychosocial data, equivalence of different approaches to web-based testing, and equivalence between web-based and paper-and-pencil measures. It is the latter category of research that focuses on the psychometric issues pertaining to cross-mode administration. Unfortunately, this category of equivalence research has received the least attention (Buchanan, 2003; Lievens & Harris, 2003). Typically focus has been on the measurement equivalence derived from paper-and-pencil measures that have been reproduced within a computerized environment. To date, evidence from these studies has generally showed measurement equivalence exist between paper and computerized versions. For example, comparing web-based and paper-and-pencil versions of the Big Five personality scale, Salgado and Moscoso (2003) found equivalence across observed scores, factor analysis structures, and reliability coefficients. Similar psychometric equivalences were found by Bartram and Brown

(2004) when investigating proctored and unproctored comparisons of the Occupational Personality Questionnaire (OPQ-32: SHL, 2000) across administration modes. In particular, both the unproctored web-based OPQ-32 and its proctored paper version displayed similar observed scores, factor intercorrelations, and reliability coefficients.

However, not all studies have found evidence of measurement equivalence. Instead, results from some studies have indicated that even when observed scores from both modes are comparable, the psychometric properties (e.g., factor structure) of each version are essentially different (Meade, Michels, & Lautenschlager, 2007). For example, using both paper-and-pencil and web-based measures of conscientiousness, agreeableness, and emotional stability for job applicants, Ployhart, Weekley, Holtz, & Kemp (2003) demonstrated that all three web-based scales produced better psychometric properties such as distributions, lower means, more variance, and greater internal consistency than the paper-and-pencil versions. A small collection of studies has also found that online versions have produced loadings on dimensions that have been unexpected and different from paper versions (see Buchanan, Johnson, & Goldberg, 2005; Johnson, 2000; Woolhouse & Meyers, 1999). For example, Johnson (2000) found that the latent structure of a web-mediated version of an International Personality Item Pool (IPIP: Goldberg, 1999) had changed slightly, where a small number of items (at a sub-scale level) loaded most highly onto factors that were incorrect based on the latent structure of the paper-and-pencil version.

Efficiency Testing

Beyond the small collection of studies investigating the validity of web-based measures through assessments of equivalence, the efficiency (e.g., reliability) of these measures has received even less attention. Of these that has focused on the reliability of computerized/web-based items, typically classical test theory (CTT) approaches have been employed (e.g., Buchanan & Smith, 1999; Davis, 1999; Pasveer and Ellard, 1998; Pettit, 2002; Ployhart, Weekley, Holtz, & Kemp, 2002). In these studies, the primary focus was replication, rather than measurement efficiency. Specifically, CTT reliability refers to the consistency of an obtained score if it were possible to replicate the measurement procedure over several separate occasions (Barlow & Proschan, 1996). Unfortunately, under this theory, the derived reliability and standard error statistics relate to test scores, and not the actual measurement instrument (Doran, 2005). Thus, this approach, while useful in establishing the consistency of scores, has less utility when wanting to examine and compare the measurement efficiency of two versions of a measure. In contrast, IRT approach to reliability is aimed at ascertaining the efficiency of the measurement procedure itself, at the point along the theta (θ) continuum where the measure is providing the most precision (Hambleton, 1989). In other words, the amount of measurement information that is provided by item i at θ, and collectively by a test at θ. Unfortunately, regardless of the distinct utility of applying an IRT approach to assessing the efficiency of innovative items, Jodoin (2003) has provided the only study using this approach. Using the innovative versions (e.g., drag-and-drop, create-a-tree) of Microsoft Certified Systems Engineer (MCSE) items and their paper-and-pencil equivalents (multiple-choice), Jodoin compared the amount of measurement efficiency provided by both versions and then against the unit

of time to complete each item. Jodoin found that information for the innovative items exceeded corresponding multiple-choice items across all theta levels (-4.0 to 4.0). However, when examining the amount of information produced by each item as a function of the time taken to complete, multiple-choice items provided more information per minute than their innovative counterparts. Thus, Jodoin concluded that while innovative items provided more information, thus more measurement efficiency, this was at the expense of increased completion time compared to multiple-choice versions. Obviously, within the context of cognitive-ability testing, where time is a factor that impacts on performance, the information being supplied by an item per unit of time is of particular import when assessing the efficacy being provided by items.

CHAPTER THREE

STUDY ONE: COMPARISON MODEL-DATA FIT FROM IRT AND IDEAL POINT IRT MODELS TO AN IDEAL POINT SCALE

Study One sought to compare the effectiveness of an ideal point model to a personality scale developed under the principles of an ideal point scale. In addition, the goodness-of-fit of a traditional polytomous dominance IRT model (GRM) to the ideal point developed scale is examined. All previous comparative fit analysis (e.g., Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Chernyshenko, 2002) have been applied to response data obtained from dominance constructed scales, thus, no attempt has been made to examine the degree of fit provided by a dominance IRT model to an ideal point constructed personality scale. As such, the present investigation extends previous research by examining the fit provided by an ideal point IRT model and a dominance IRT model to an ideal point constructed personality scale.

Model-data fit will be compared using the category response functions generated under two IRT models representing each approach, namely the ideal point-based generalized graded unfolding model (GGUM) and the dominance-based GRM. The two parametric models were chosen for comparison based on the similar characterization of responses, namely, models estimate the same number of parameters, and allow for both discrimination and response thresholds to vary across items. In addition, the GRM and GGUM are both appropriate for polytomous responses.

Given the argument offered by Chernyshenko et al. (2001) that an ideal point response process may be a more appropriate approach to modeling personality data, and further findings confirming this conjecture (Chernyshenko, 2002), it is expected that the ideal point model (GGUM) will provide better fit to the ideal point response patterns than the dominance-based GRM. The research question for this study was: Which measurement model provides the best fit to the ideal point constructed Academic Self-Worth Scale?

Method

Participants

Participants consisted of 1020 secondary school students (527 females, 493 males) from four secondary schools in the Auckland area. Students ranged in age from 16-20 years, with a mean age of 16.70 years (*SD* = .77) covering secondary school Years 11, 12, and 13.

Scale Development

Instrument

Participants completed the 74-item Academic Self-Worth Scale (ASWS) consisting of three sub-scales measuring Defensive Expectations (30 items), Reflectivity (15 items), and Self-handicapping (29 items). Items for each sub-scale were developed and adapted using ideal point scale construction guidelines (see Procedures Section), from pre-existing scales for each construct namely the Academic Self-Handicapping Scale (Midgley, Arunkumar, & Urdan, 1996), Life Orientation Test (Scheier, Carver, & Bridges, 1994), Academic Process Questionnaire (Martin,

1998), and Defensive Pessimism Questionnaire (Norem & Cantor, 1986). Within the Self-handicapping and Defensive Pessimism sub-scales, six distinct strategy approaches (see Table 1) are represented.

Table 1
Sample items from the Academic Self-Worth Scale

Sub-scale	Example Item
Self-handicapping:	
Active self-handicapping	"I intentionally over-commit my time outside school, which keeps me from studying, *so I have an excuse if I don't do as well as I hoped*"
Self-presented self-handicapping	"I tell others that I fool around the night before an exam, even if I actually don't, *so I can use this as the reason, if I don't do as well as I hoped*"
Self-presented affective self-handicapping	"When an assignment is due, I let people think that I'm more tense and uptight than I really am, *so I can use this as the excuse, if I don't do as well as I hoped*"
Defensive Pessimism:	
Defensive expectations	"*No matter how well I've done in the past*, I have serious doubts about my overall academic ability"
Self-presented defensive expectations	"*No matter how well I've done in the past*, I tell people that I'm more pessimistic about future academic performances, than I really am"
Reflectivity	"Fear of failure is always on my mind as I prepare for exams"

Note. Predicates (defensive pessimism) and trailers (self-handicapping) are indicated by italics.

Self-handicapping items

Self-handicapping items were derived and adapted from Midgley et al., (1996), Strube (1986), and Martin (1998). Active self-handicapping items reflect the active impediments that an individual will employ in order to purposefully impede upcoming measurements of their ability. Items include excuses that typify the deliberate and destructive actions adopted by students, such as, "I let myself get run-down when

assignments or exams are due", and "Instead of doing study after school, I tend to hang out with friends during the term".

Similarly, self-presented self-handicapping items consist of content that portray the individual as having obstacles to their performance, but unlike the active self-handicapping content, these impediments either do not actually exist or are excessively exaggerated (Martin, 1998). Here individuals are not motivated to actively ruin their chances of performing well, instead they focus on presenting themselves as being disinterested or at a disadvantage, thus publicly managing any potentiality of academic failure. Examples of such image management items are, "I tell people that I put assignments and study off until the last moment more than I actually do", and "I let people know that I am involved in a lot more activities than I really am when exams or assignments are due".

In addition, self-presenting affective items were also incorporated to reflect the less tangible emotional states presented by individuals. Like the self-presented items, the individuals are either not actually experiencing the portrayed affective states or the extent to which they are occurring is significantly embellished. For example, "When an exam or assignment is due, I let people think that I'm more tense and uptight than I really am", and "When I have an exam coming up, I sometimes tell others that I'm more frustrated with my prep than I really am".

Based on the item design from Midgley et al. (active and self-presented) and Martin (self-presented affective self-handicapping), items finished with a similar trailer relating to the motivation for presenting an image regarding their upcoming performance scenario (see Table 1). The use of this trailer helps maintain the

respondents' focus on the underlying reasons for presenting their alibi for potential poor performance, a priori (Martin, 1998).

Defensive pessimism items

Defensive expectation and reflectivity items were based from the strategy prototypes devised by Norem and Cantor (1986) and the Life Orientation Test (Scheier, Carver, & Bridges, 1994), and drawn from the further adaptations made by Martin (1998).

In this scale, active defensive expectation items reflected the strategies whereby individuals adopt unrealistically low academic expectations in order to protect themselves from the negative expectation of failure (Norem & Cantor, 1986). Norem and Cantor's original measure of defensive pessimism was conducted after scores from the Optimism-Pessimism Pre-screening Questionnaire had been scored. This approach was based on the assumption that students who were pessimistic regarding their performance on previous assessments were more likely to have defensively pessimistic orientation. Martin's adaptation of Norem and Cantor's prototypes does not make such an assumption, and suggests instead the use of the predicate "No matter how well I have done in the past…" as a way of integrating a screener into each of the items (see Table 1).

As with self-handicapping strategies research Martin (1998) adapted active defensive expectation scenarios into self-presented versions. Also, as with the self-handicapping counterparts, the content in the self-presented defensive expectations items reflect students presenting themselves as being *more* pessimistic than they actually are towards upcoming performance scenarios on their ability (e.g., "I tell everyone that I question my academic ability, more than I actually do").

The reflectivity items for this scale were derived from work of Norem and Illingworth (1993) and Martin (1998), and represent a dimension of defensive pessimism that sees students thinking through (reflective) the potentiality of failure *and* success. Therefore, where students with active defensive expectations *expect* negative performance outcomes, students who adopt reflectivity as a self-worth strategy simply *think* about the possibilities of negative performance outcomes (Martin, Marsh, & Debus, 1999). As such, items in the reflectivity sub-scale reflect both positive and negative reflections regarding assessed performance (e.g., "I think about how I will feel if I do very poorly in tests and assignments", "Imagining how I will feel after a successful result motivates me to work hard").

Procedure

The Academic Self-Worth Scale developed for this study follows the guidelines recommended by Chernyshenko (2002) for scale construction given an ideal point approach. The first step, item pool generation, involved creating items within each sub-scale to cover the entire trait continuum. With the assistance of two experienced item writers, both of whom are knowledgeable in the subject of academic self-worth, additional items were written for the three sub-scales. Particular attention was given to writing what was considered neutral and extreme reflections of the trait being measured. Guidelines on how to write non-cognitive ability items to reflect areas on the trait continuum provided by Michell (1994) were also used during this process. In total, there were 150 items in the initial pool: 40 from the Academic Process Questionnaire (APQ; Martin, 1998) and 110 developed items. The structure and design of the items from the APQ provided the template from which the additional 110 items were developed.

The second step involved rating each item according to what aspect of the trait continuum the item reflected. After being thoroughly briefed regarding the definition of each construct two judges were asked to rate items based on three categories: weak, moderate, and high representations of the construct. For example, *weak* representations of self-handicapping reflect individuals allowing external obstacles to be imposed upon themselves (e.g., "I let my friends distract me from paying attention in class or from doing my study or assignments"), whereas, an example of *high* self-handicapping reflects a deliberate destructive internal action performed by the individual (e.g., "I purposely don't get enough sleep or rest before upcoming exams and assignments"). Between these extremes, an exemplar of moderate self-handicapping highlights the occasional employment of obstacles (e.g., "I often spend time doing non-urgent things just before an exam or assignment is due").

The third step involved selecting the items that showed the greatest inter-rater agreement, ensuring that a good representation across the three continuum categories was achieved for each sub-scale.

As the items taken from the Martin's (1998) study were originally devised for an adult student population, the final step involved ascertaining the reading level of the items developed for this questionnaire. Two experts (secondary school teachers) rated the reading level of the scale, assessing the items clarity and wording.

The final version of the ASWS contained three sub-scales: Defensive Expectations (29 items), Reflectivity (15 items), and Self-Handicapping (30 items). Items on all sub-scales were rated on a 5-point Likert scale (1 = "strongly disagree", 2 = "disagree", 3 = "neither disagree nor agree", 4 = "agree", and 5 = "strongly agree"). Given the scale construction procedure outlined above, the final theta values of the

ASWS items were expected to span evenly across the -3 to 3 of the trait continuum. In particular, given the focus of including items to strongly represent the middle of the trait range (e.g., -1 to 1), it was anticipated that average of items estimates would be close or at 0 theta.

The ASWS was circulated to participating schools, and administered to students using approved ethics procedures.

Statistical Analysis

Cross-validation

The dimensionality and model-data fit were assessed using a cross-validation procedure as recommended by Drasgow, Levine, Tsien, Williams, and Mead (1995). This involves the arbitrary splitting of the total sample into two data sets, a calibration sample used for the estimation of item parameters, and a validation sample used as the "empirical sample" of response data. The use of separate samples for estimating parameters and examining fit is preferable as it avoids artifacts (e.g., sampling fluctuations) that result from the overfitting from models with greater numbers of parameters (Drasgow et al., 1995). Thus, any fit advantage gained from complex models estimating (many parameters) is negated, allowing for a more equitable comparison to simpler models (fewer parameters). Further, cross-validation avoids unrealistically high fit results provided when item estimates and responses are from the same sample. Thus, the cross-validation technique provides a more stringent and robust approach to analyzing model-data fit. The calibration sample (n = 510) was generated by selecting every second respondent from the initial data set, and used for item parameter estimates under GRM and GGUM. The remaining sample formed the

validation sample (n = 510) was used as the empirical proportions across the models examined.

Exploratory Factor Analysis of the Academic Self-Worth Scale

To assess the structure of the ASWS an exploratory factor analysis (EFA) was conducted with the aim of establishing the number of common factors influencing this measure. As the assumption of multivariate normality was satisfied, factor analysis utilized the maximum likelihood method for extraction with an oblique rotation. This rotation method was selected based on the desire to extract factor patterns regardless of their degree of correlation, and based on previous evidence (e.g., Martin, 1998) suggesting that the constructs in the ASWS are not orthogonal. Therefore, an oblique rotation approach will present a more precise and realistic depiction of how the three constructs in the ASWS are likely to be related to one another (Fabrigar, Wegener, MacCallum & Strahan, 1999).

IRT Item Parameter Estimation

Prior to item parameter estimation, data from the positively worded items (reflectivity sub-scale) were reversed scored**.** The GRM item parameters for each of the four sub-scale's calibration samples were estimated using MULTILOG 7 (Thissen, Chen, & Bock, 2003). This program employs optimal full information marginal maximum likelihood (MML) item estimation procedures by the use of the EM algorithm (Bock & Aitkin, 1981). Under the ideal point model item parameter estimations for each of the four sub-scales' calibration samples were estimated using GGUM2004 (Roberts, Fang, Cui, & Wang, 2006). Because of the assumptions of an ideal point approach, data from positively worded reflectivity items was not reversed scored. Like MULTILOG, item parameters are estimated using the MML approach,

with the exception that person parameters are estimated via the expected a posteriori (EAP) procedure.

Model-Data Fit

An essential precursor to examining model-data fit is that data must conform to model assumptions regarding the dimensionality of the models to be applied. However, an issue that arises when confronted with the assessment of dimensionality under an ideal point approach lies in that typical procedures, such as principle components or maximum-likelihood methods, assume that a dominance process underlies the responses to the items (Chernyshenko, 2002). Specifically, an individual's high observed score reflects a high standing on a latent trait. As the response data in this study was assumed to reflect an ideal point process, based on the construction methods of the ASWS, it would have been appropriate to assess the scale's dimensionality with an unfolding procedure. However, Habing, Finch, and Roberts (2005) provide the only study that attempted to develop a test of dimensionality, Yen's Q_3 statistics, specifically for the ideal point approach. Although Habing et al. found promise for this approach, further investigations are required before the performance of the modified $Q_{3_{i,j}}^{unf}$ statistic can be a used as a reliable assessment of unidimensionality.

An approach that avoids issues relating to the use of dominance-based procedures is the chi-square fit statistic. Previous studies have shown that chi-square fit statistics derived from single items (singlets), bundles of two (doublets) (Glas, 1988; Van den Wollenberg, 1982), and three (triplets) (Chernyshenko et al., 2001; Drasgow, Levine, Tsien, Williams, & Mead, 1995) item combinations provide a useful indication of the degree of dimensionality and local independence existing in the response data.

Although chi-square statistics for single items have been found to be largely insensitive to violations of unidimensionality and by extension local independence, chi-squares computed for bundles or two (doubles) and three (triples) items have been found to be sensitive to such issues. Where single chi-square statistics are computed based on the expected frequency that respondents would select option *k*, chi-statistics for doubles and triples are established based on the expected and observed probabilities that respondents would endorse specific options on two (or three) items under a unidimensional model (Drasgow et al., 1995). Thus, to compute a chi-square doubles statistic, the expected frequency of the $(k, k')^{th}$ cell in the two-way table is calculated, with the observed frequencies for the two-way contingency table counted. For example, Item 1, Option 1, and Item 2, Option 2 would be computed. Cells with expected frequencies of < 5 are aggregated, and if the sum of the expected frequency is still <5, the cells are then combined with the cell with the expected frequency that *least* exceeded 5. From these calculations the chi-square two-way is produced. By extension, the same procedure is conducted for the triples statistic where a three-way contingency table comparison of the IRT model's observed and expected probabilities of endorsement (e.g., Item 1, Option 1, Item 2, Option 2, and Item 3, Option 3) is computed.

As such, chi-square statistics based on pairs and triples provide an indication for how well the IRT model is able to predict the interaction between the items, in other words the ability of the IRT model to reconstruct the patterns of responses. Where chi-square fit statistics for doubles and triples are small (e.g., chi-square to degrees of freedom ratio = <3.0) then the assumption of unidimensionality is adequately met (Drasgow et al., 1995). Adopting interval criterion of Chernyshenko et al., the chi-

square to *df* ratios were arranged into six units: very small (<1), small (≥1 to <2), medium (≥2 to <3), moderately large (≥3 to <4), large (≥4 to <5), and very large (≥5).

Drasgow et al. (1995) showed that if the chi-square fit statistic is adjusted to the magnitude expected in a sample of 3,000, then effective cross comparison of models consisting of a different number of parameters and/or where sample sizes differ can be conducted. As the models examined in this study consisted of the same number of parameters and the same sample size, an adjusted chi-square was not employed.

In addition to statistical fit analysis, graphical goodness-of-fit procedures were also applied to the data. Adopting Levine and Williams' (1991, 1993) graphical fit plot method and EMPOCC program (Williams,1999), the MODFIT computer program provides a measure of fit between models through the comparison of theoretically derived item/option response function (estimated using the calibration sample item estimates) and empirical option response function (estimated using the validation sample of responses). MODFIT divides the theta continuum into 25 equally spaced points, where theta is then estimated for each of the respondents from the validation sample, with the total number of respondents totaled at each point (Chernyshenko et al., 2001). In addition, where respondent numbers permit, each of the theta points are estimated with their 95% confidence interval.

The following criteria was used to evaluate which IRT model provided the greatest degree of congruence between the model's theoretical assumptions and the empirical pattern of the actual response data: the distance between the theoretical and empirical option response functions, representation of the higher and lower tails of the response function, and the size of confidence intervals. The reader is referred to

Drasgow et al. (1995) for a detailed discussion about these statistical and graphical goodness-of-fit analyzes.

Results

Exploratory Factor Analysis of the Academic Self-Worth Scale

Two rules, namely, Cattell's (1966) scree plot and Kaiser's (1960) eigenvalues rule (eigenvalues greater than 1) were used to determine the latent structure of the ASWS. The scree plot of eigenvalues showed a clear break in the data, with four factors positioned above this break. Similarly, although fourteen factors had eigenvalues larger than 1.00, accounting for 53.40% of the variability, a four-factor solution was considered the most parsimonious explanation of the data, as these factors accounted for more than 4% of the variability (see Table 2). Thus, while the structure of the ASWS suggested three factors, the self-presenting aspects represented in both self-handicapping and defensive expectations loaded together on a single additional factor. Specifically, the first factor accounted for 19.10% of the variability and focused, with the exception of one item, on the self-presented dimension from both the self-handicapping and defensive expectation constructs. The second factor accounted for 6.64% of the variance and appears, with the exception of three items, to focus on defensive expectations. The third factor accounted for 5.81% of the variance and with the exception of one item, focuses on self-handicapping. The fourth factor accounted for 4.62% of the variance and consisted of all the reflectivity items.

In sum, twenty-eight items (18, 19, 22, 25, 26, 29, 30, 33, 34, 37, 38, 40, 41, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 72) loaded solely on the first factor (self-presenting), eighteen items (14, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69,

70, 71, 73, 74) loaded solely on the second factor (defensive expectations), fourteen items (16, 17, 20, 21, 23, 24, 27, 28, 31, 32, 35, 36, 39, 43) loaded solely on the third factor (self-handicapping), and fourteen items (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15) loaded solely on the fourth factor (reflectivity). Across the four factors, 5 items (14, 18, 20, 70, 74) cross-loaded or loaded onto factors which did not relate to the construct that they sought to measure and were removed from further analysis.

Given these results, further analysis was conducted using the four factors (self-presenting, defensive expectations, self-handicapping, and reflectivity) identified in this EFA.

Table 2
Total variance explained in the EFA for factors whose eigenvalues exceed 1.00

Factor	Eigenvalue	% of Variance	Cumulative % of Variance
1	14.14	19.10	19.11
2	4.91	6.64	25.75
3	4.30	5.81	31.56
4	3.42	4.62	36.18
5	1.99	2.70	38.88
6	1.60	2.17	41.05
7	1.39	1.87	42.92
8	1.26	1.71	44.63
9	1.17	1.58	46.21
10	1.13	1.52	47.73
11	1.09	1.48	49.21
12	1.08	1.45	50.66
13	1.03	1.39	52.05
14	1.00	1.35	53.40

Table 3 presents a summary of the scale and data characteristics across the four sub-scales. Caution needs to be taken regarding the interpretability of these descriptive statistics, however, as traditional item analysis such as estimates of

reliability, and total score and associated statistics, are not applicable to an ideal point method. Alpha coefficients ranged from .75 to .91 with an average across scales of .85.

Table 3
Sub-Scale Characteristics

Sub-Scale (# Items)	Number of Valid Cases			Mean	SD	Alpha
	Total	Calibration	Validation			
Defensive Expectations (15)	1020	510	510	28.32	12.21	.90
Reflectivity (14)	1020	510	510	33.85	7.98	.75
Self-Handicapping (13)	1020	510	510	23.69	9.59	.83
Self-Presenting (27)	1020	510	510	43.11	17.18	.91

IRT Item Parameter Estimation

Table 4 lists the average of item parameter estimates derived from the GGUM and the GRM. Item locations parameter averages across both models showed distinctly different item-theta locations for each respective response category. Across the four sub-scales, GGUM parameters showed items representing average to extreme negative theta levels. In comparison, the GRM item estimations showed a representation of items spanning across the entire trait continuum. Given the predominance of negatively worded items across all four sub-scales, the location of the GGUM parameters may be a more accurate reflection of the actual trait location of the items. Alternatively, the GGUM estimations might reflect a failure of the scale construction methodology, where a representation of items across both negative and positive areas of the trait continuum was sought. Therefore, although a balanced theta representation of items were selected for this scale, only the GRM reflected the distribution of expected item estimations.

The GRM and GGUM estimated similar discrimination parameters across the four sub-scales. Interestingly, the amount of information, thus measurement precision, provided by an item under both models is indicated by the relationship between discrimination value and range of location parameters under both models. High item

information is achieved under the GRM when the discrimination parameter is large and the range of location parameters is narrow. Similarly, the GGUM stipulates that maximum information is achieved when discrimination parameters are large and the inter-threshold distances between the SRC locations are small, thus resulting in a smaller location range. Although covering different areas of the trait continuum, the ranges between response categories for self-handicapping, reflectivity, and defensive expectations across both models were extremely similar. Parameter averages for the self-presenting sub-scale showed greater measurement precision provided by the GGUM, with the range of SRC locations significantly smaller than those estimated by the GRM.

Conversely, both models generated similar moderate discrimination parameters. This finding may relate to the negatively worded nature of the sub-scales, with only the Reflectivity sub-scale containing both positively and negatively worded items. Concerns regarding negatively worded items have been posited in numerous research focusing on issues such as potential method artifact (e.g., Goldsmith, 1986; Horan, DiStefano, & Motl, 2003; Marsh, 1996; Zumbo, Gelin, & Hubley, 2001), and response bias (e.g., Hensley, 1988; Rowe & Rowe, 1997; Sandoval, 1977, 1981) associated with these items. In relation specifically to item discrimination, Fletcher and Hattie (2004) found that negatively worded items from the Physical Self-Description Questionnaire (PSDQ: Marsh, Richards, Johnson, Roche, & Tremayne, 1994) showed lower discrimination parameters and a greater range of location parameters, thus, providing low information and poor measurement precision in comparison to the positively worded items.

Table 4
GGUM and GRM Average Item Parameter Estimations

Sub-Scales (# Items)	GGUM						GRM				
	$\hat{\alpha}_i$	$\hat{\delta}_i$	$\hat{\tau}_i 1$	$\hat{\tau}_i 2$	$\hat{\tau}_i 3$	$\hat{\tau}_i 4$	a	$b1$	$b2$	$b3$	$b4$
Defensive Expectations (15)	1.00	-3.07	-4.50	-3.16	-2.76	-0.85	0.99	-1.63	-0.34	0.61	2.04
Reflectivity (14)	0.70	-2.59	-5.22	-3.48	-3.11	-0.21	0.64	-3.22	-1.35	0.03	2.26
Self-Handicapping (13)	0.77	-3.31	-4.55	-3.22	-3.25	-1.04	0.85	-1.54	-0.28	0.58	2.09
Self-Presenting (27)	1.47	0.15	-1.84	-0.65	-0.18	0.85	0.79	-1.77	-0.03	1.20	2.87

Model-Data Fit

Table 5 shows the chi-squares to degrees of freedom ratios, means and standard deviations for the singlets, doublets, and triplets for each of the four sub-scales under both models. The model that produced the smallest mean (i.e., best model-data fit) for that item combination, under the respective sub-scale is indicated in boldface. Although the frequency of chi-square to degrees of freedom ratios, and thus magnitude of the means were similar across both models, Table 5 shows that the GRM produced the best model-data fit in 75% of possible item combinations. There are not large differences between the estimations from each model, although overall the GRM appears to be more effective at predicting patterns of responses for combinations of two or three items. The chi-square to *df* ratios across the four sub-scales indicated that the data under both models adequately met the assumption of unidimensionality, with all but six of the 24 ratio means less than the 3.0 criterion. Interestingly, the six indications of a moderately large violation of unidimensionality were exhibited by both models for the same item combinations under the same sub-

scales (see Table 5, underlined values). Under both models, doublet and triplet results indicated that the Self-Presented sub-scale may present a degree of multidimensionality in the calibration sample, or both models fit the data inadequately. Although the EFA clearly clustered the self-presentation items, it should be noted that this factor was originally developed for both defensive expectation and self-handicapping sub-scales. Thus, the impact of these two distinct constructs underlying the dominant Self-Presented factor may be reflected in the slightly elevated chi-square ratios. However, the chi-square ratios were not considered to greatly violate the less than 3.0 criterion. Item interaction was modeled the most effectively in the reflectivity sub-scale, with the exception of GGUM doublets, both models producing small average chi-square distribution frequencies ratios.

Table 5
Means, Standard Deviations, and Frequencies of Chi-Square to df Ratio in the Calibration Sample

			Frequency Distribution of χ^2 to *df* Ratio							
Model	Sub-Scale (# Items)	Item Combination	<1	1-<2	2-<3	3-<4	4-<5	5>	Mean	SD
GGUM	Defensive Expectations (15)									
		Singlets	3	4	5	1	1	1	2.035	1.331
		Doublets	0	2	10	3	0	0	2.625	0.625
		Triplets	0	1	2	2	0	0	2.666	0.696
	Reflectivity (14)									
		Singlets	5	5	2	2	0	0	**1.608**	1.200
		Doublets	0	8	7	3	0	0	2.239	0.679
		Triplets	0	6	3	1	0	0	1.958	0.604
	Self-Handicapping (13)									
		Singlets	0	1	5	1	3	3	**3.547**	1.530
		Doublets	0	1	7	3	4	0	3.034	0.957
		Triplets	0	0	6	1	0	0	2.563	0.318
	Self-Presented (27)									
		Singlets	5	12	1	4	2	3	2.367	1.678
		Doublets	0	2	10	8	5	2	**3.256**	1.222
		Triplets	0	0	3	3	2	1	3.503	0.984
GRM	Defensive Expectations (15)									

	Singlets	4	7	2	1	1	0	**1.737**	1.138
	Doublets	0	4	10	1	0	0	**2.251**	0.574
	Triplets	0	1	3	1	0	0	**2.329**	0.660
Reflectivity (14)									
	Singlets	5	4	3	2	0	0	1.636	1.183
	Doublets	1	9	7	1	0	0	**1.936**	0.651
	Triplets	0	9	1	0	0	0	**1.675**	0.434
Self-Handicapping (13)									
	Singlets	0	1	4	2	3	3	3.638	1.353
	Doublets	0	1	8	5	0	1	**2.939**	0.896
	Triplets	0	1	4	2	0	0	**2.542**	0.457
Self-Presented (27)									
	Singlets	7	10	2	5	1	2	**2.148**	1.462
	Doublets	0	3	10	9	2	3	3.403	1.230
	Triplets	0	1	3	2	3	0	**3.307**	1.048

Note. The lowest mean in each row between the two models is indicated in bold. Moderately large violations of unidimensionality are underlined.

Fit plots representing item estimates from both models were constructed by MODFIT for each of the five response options. As it was not feasible to display all of the 345 fit plots generated, plots will be presented that illustrate the overall fit plots, and results summarized. Table 6 shows which model provided the best fit to the items in each sub-scale. With the exception of the Defensive Expectation sub-scale, fit plots showed that a large percentage of items were represented equally well by both the GGUM and GRM.

Table 6
Number of Fit Plots Providing Best-Fit to Sub-Scales

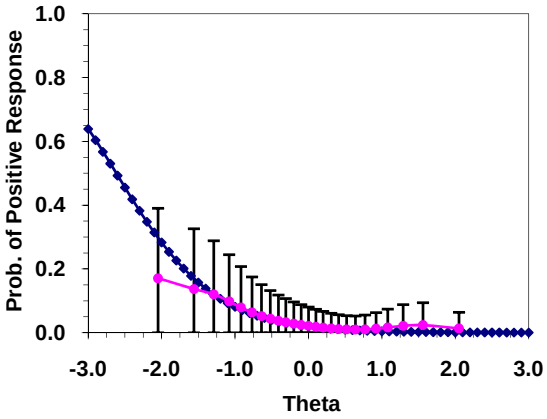
	Frequency of Plots Judged Best-Fit		
Sub-Scale (# Items)	GGUM	GRM	Both
Defensive Expectations (15)	2 (13%)	8 (53%)	5 (33%)
Reflectivity (14)	3 (21%)	4 (29%)	7 (50%)
Self-Handicapping (13)	3 (23%)	1 (8%)	9 (69%)
Self-Presented (27)	5 (20%)	10 (37%)	12 (44%)
All Items (69)	13 (18%)	23 (32%)	33 (48%)

Note. The percentages of frequencies in parentheses.

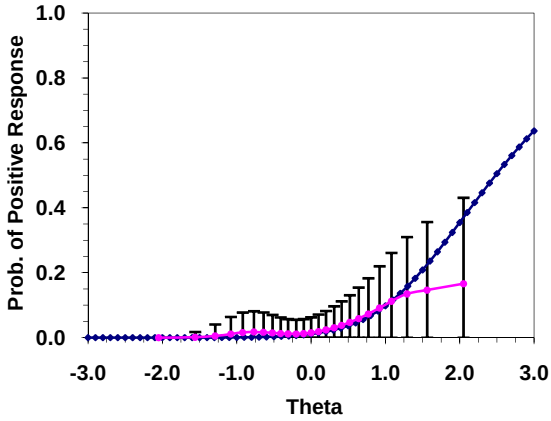
Figure 1 presents Item 11 from the Reflectivity sub-scale as a typical example of the close level of agreement in fit between both models. Item 11 fit plots generated under each model show an almost perfect reflection (symmetrical from the point of origin, where theta equals zero) of theoretical option response and empirical option response functions. In addition, discrepancies between the empirical proportions and the estimated option response functions largely occur at the same inverse trait levels on the continuum. For example, note that the oscillation in the empirical "neither disagree nor agree" option response function (see figures e and f), is located at 1.00 under the GGUM and -1.00 under the GRM. Although highly similar, the GRM produces slightly less better fit between the two functions, apparent most for the "agree" and "strongly agree" response options.

Thus, Item 11 is representative of the fit provided by both models across the Self-Handicapping, Defensive Expectations, and Reflectivity sub-scales. The shape of the item response functions under these sub-scales was largely monotonic, allowing explanation of response patterns to be modeling successfully by the s-shaped functions generated by the GRM. Of interest was the good fit provided by the nonmonotonic GGUM functions, highlighting the ability of this model to provide fit to responses that are more reflective of a dominance process.

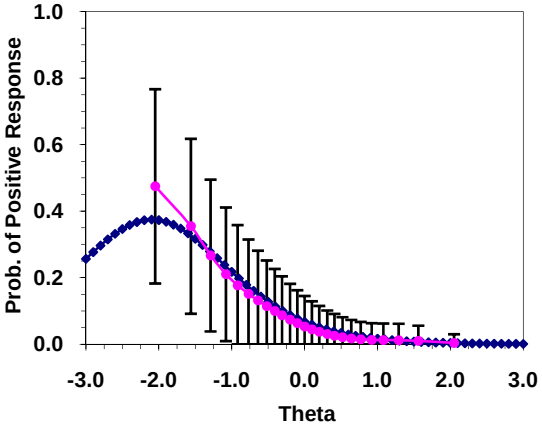
(a)


Prob. of Positive Response
1.0
0.8
0.6
0.4
0.2
0.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
Theta

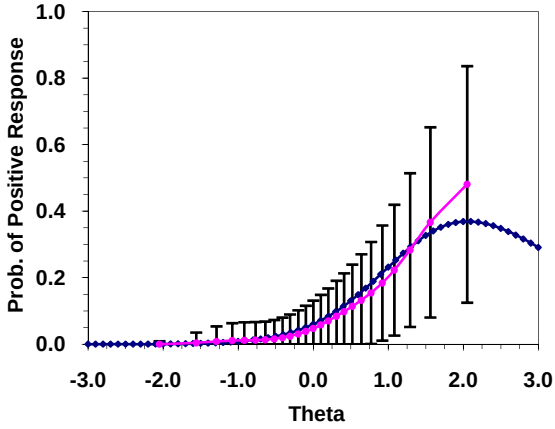
(b)


Prob. of Positive Response
1.0
0.8
0.6
0.4
0.2
0.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
Theta

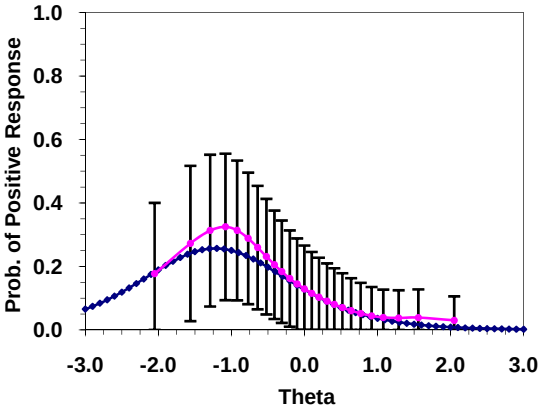
(c)


Prob. of Positive Response
1.0
0.8
0.6
0.4
0.2
0.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
Theta

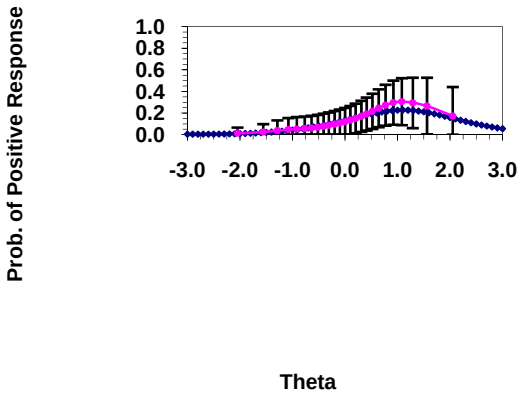
(d)


Prob. of Positive Response
1.0
0.8
0.6
0.4
0.2
0.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
Theta

(e)


Prob. of Positive Response
1.0
0.8
0.6
0.4
0.2
0.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
Theta

(f)


Prob. of Positive Response
1.0
0.8
0.6
0.4
0.2
0.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
Theta

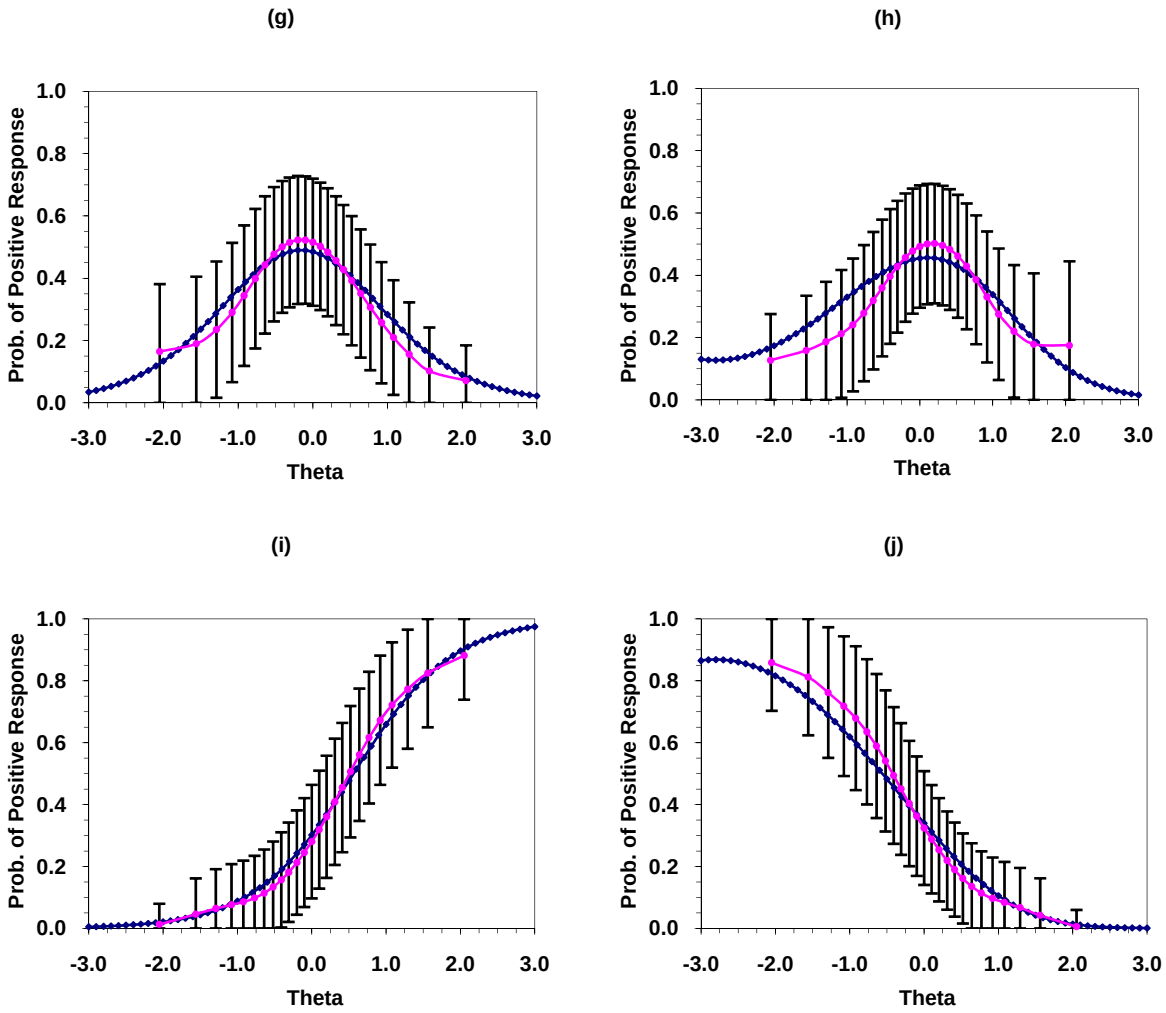
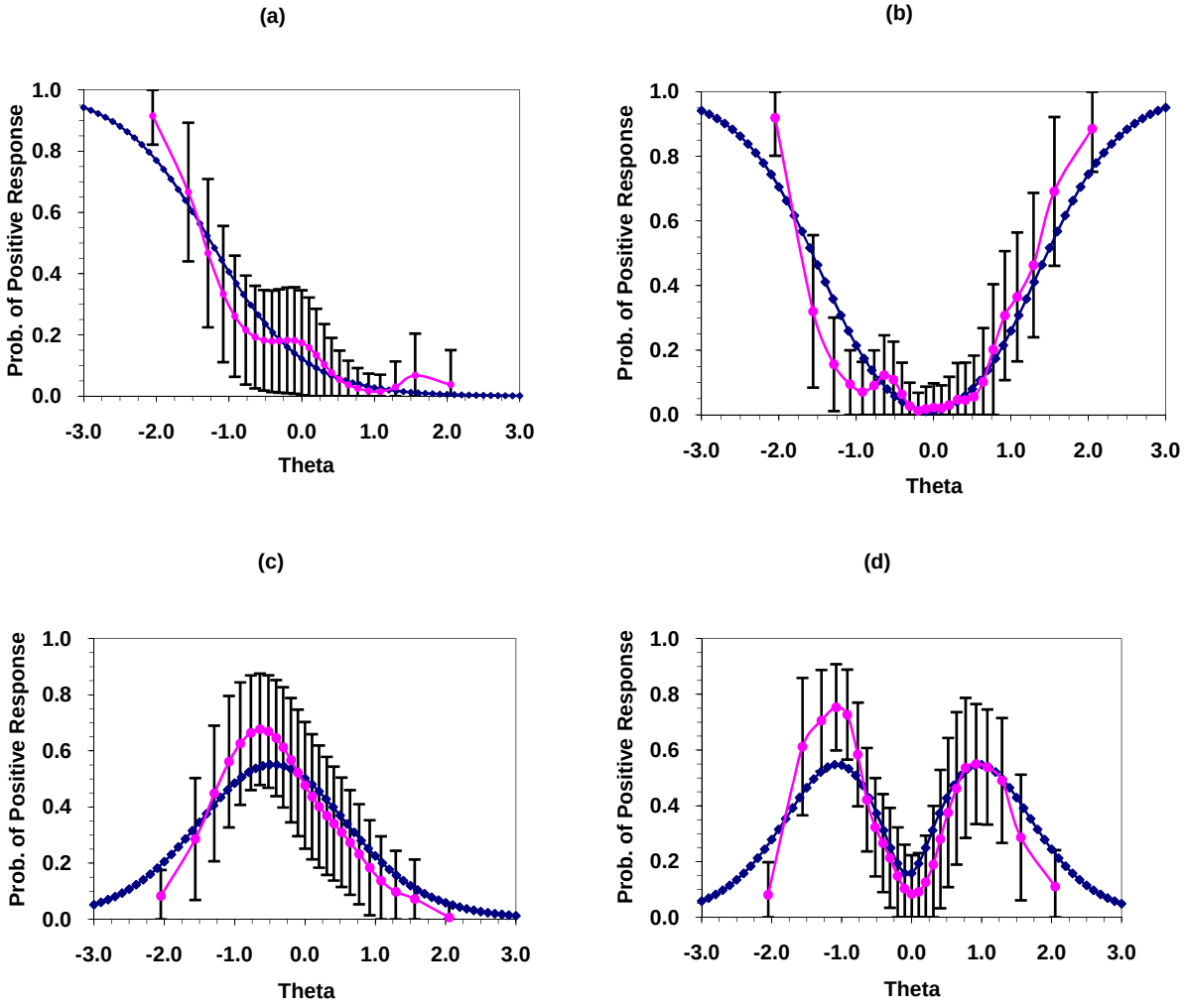


Figure 1. Fit Plots for the GGUM and GRM (Item 11 of the Reflectivity sub-scale): (a) Strongly Disagree Option (GRM), (b) Strongly Disagree Option (GGUM), (c) Disagree Option (GRM), (d) Disagree Option (GGUM), (e) Neither Disagree Nor Agree Option (GRM), (f) Neither Disagree Nor Agree Option (GGUM), (g) Agree Option (GRM), (h) Agree Option (GGUM), (i) Strongly Agree Option (GRM), (j) Strongly Agree Option (GGUM).

Only the Self-Presented GGUM theoretical and empirical option response functions displayed the typical ideal point unfolding shape across the bounds of theta -3.0 to 3.0. Figure 2 presents an item (Item 8) from the Self-Presented sub-scale that is representative of the typical options functions generated by both models in this sub-scale. As expected, the moderate degree of misfit shown by chi-square to degrees of freedom ratios for this sub-scale is similarly reflected in both models estimated plot functions. GGUM plots indicate *typical* performance, as responses are clearly

nonmonotonic across both theoretical and empirical response categories with peaks being symmetric about the point of origin $\left(\theta_j - \delta_i\right) = 0$. Interestingly, the GGUM empirical functions are not well estimated by the theoretical option response function, especially for empirical responses lying around the functions upper bound (e.g., functions highest peak) on the trait continuum. In comparison, the GRM appears to provide a similar degree of fit when the data is modeled under *maximal* performance assumptions, with Figures 2a and 2c actually showing less oscillations under the dominance-based approach.



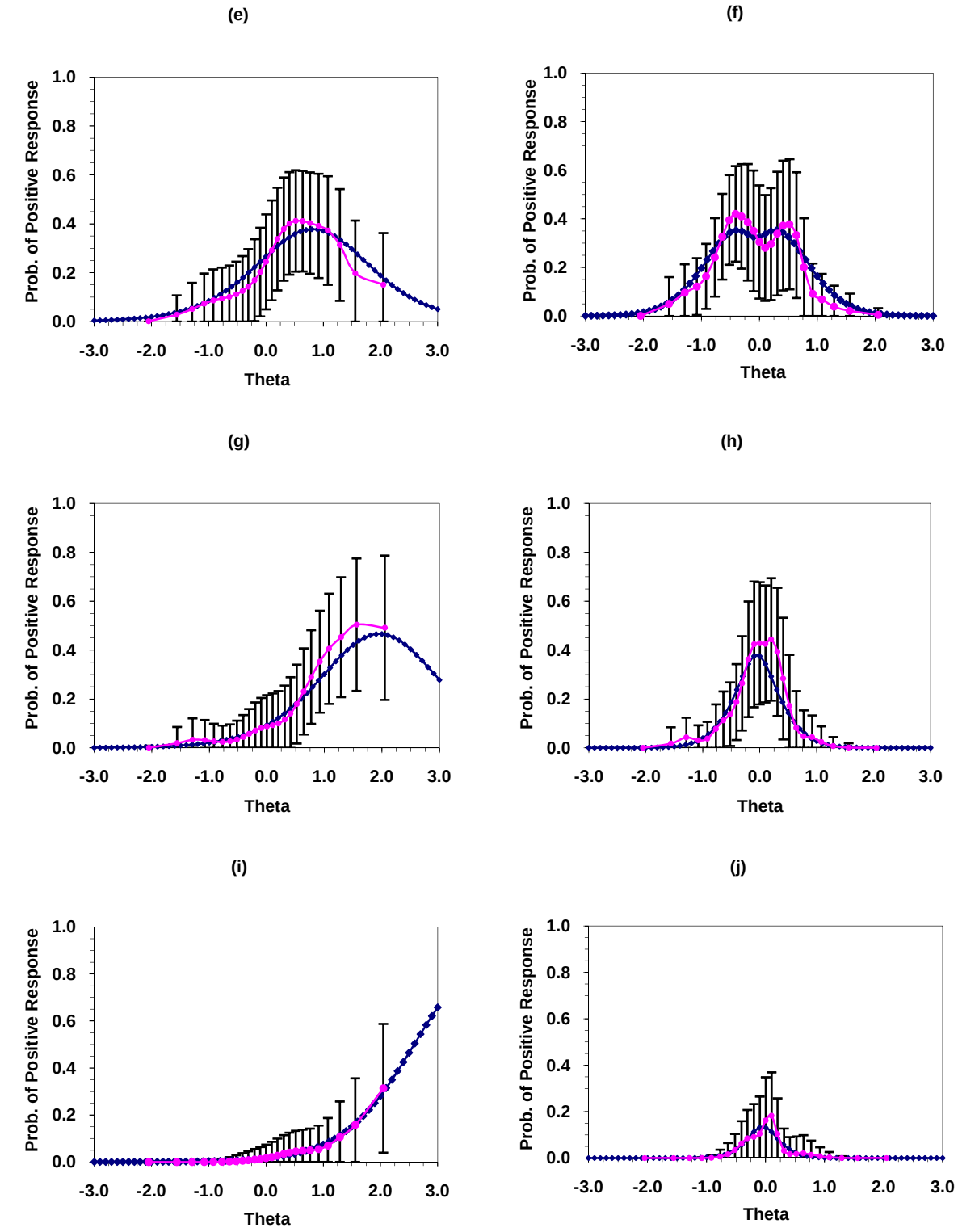


Figure 2. Fit Plots for the GGUM and GRM (Item 8 of the Self-Presented sub-scale): (a) Strongly Disagree Option (GRM), (b) Strongly Disagree Option (GGUM), (c) Disagree Option (GRM), (d) Disagree Option (GGUM), (e) Neither Disagree Nor Agree Option (GRM), (f) Neither Disagree Nor Agree Option (GGUM), (g) Agree Option (GRM), (h) Agree Option (GGUM), (i) Strongly Agree Option (GRM), (j) Strongly Agree Option (GGUM)

Discussion

This study sought to compare the application of two theoretically diverse approaches (dominance and ideal point) to analyzing personality data derived from an ideal point constructed scale. Previous research has suggested that the traditional IRT approach to analyzing personality responses from dominance constructed scales may not adequately reflect the response behavior (nature of responding) of test-takers. This study compared the application of two theoretically diverse approaches (dominance and ideal point) to analyzing personality data based on a scale constructed from the principles of ideal point. The study is only the second to construct a scale under the ideal point assumptions, and the first to examine the fit of a dominance-based IRT model to an ideal point scale. While the scale was developed using the ideal point methodology, the ideal point model (via GGUM) did not provide the expected superior fit to the response patterns. Instead, the graphical results from this study showed that for nearly half of the items, the fit provided by the GGUM estimations showed little if any difference to that provided by the GRM. It should be noted however, that while the GGUM did not outperform the GRM in regards to model-data fit, the GGUM estimates did fit the data well.

Statistical fit analysis supported the fit plot findings, with chi-square statistics showing comparable magnitudes of fit across all four sub-scales. So similar were the two models that the same statistical misfit was found for the same item combinations in the Self-Handicapping (singlets) and Self-Presented (doublets and triplets) sub-scales. Chi-square fit analysis revealed that GGUM produced outright superior model-data fit in only a quarter of the item combinations across all scales.

The question that arises is why does the GRM seemingly perform against expectations by providing comparable fit to the GGUM? The fit performance by the GRM challenges previous conjecture and findings (e.g., Chernyshenko et al., 2001; Roberts et al., 2000) suggesting that dominance models may not provide best fit for non-cognitive ability data. A possible explanation may lie at the scale construction level, with the prevalent fit from the GRM evidence of a scale that did not accurately reflect a true ideal point structure. Even given adherence to construction of an ideal point scale, Chernyshenko (2002) found that the majority of GGUM estimated items displayed monotonic item response functions. The similar finding in this study suggests that, despite the best of intentions, the resultant sub-scales did not contain items representing the middle location on the trait continuum. The GGUM location parameters in this study showed that in three sub-scales (Defensive Expectations, Reflectivity, and Self-Handicapping), few of the initially developed and rated "neutral" items had actually transpired into items that reflected a neutral position on the trait continuum. The GGUM location parameters for the Self-Presented sub-scale provided the only exception, with a large proportion of parameters representing the neutral locations that they had been initially judged. According to Chernyshenko, a good representation of neutrally located items, results in more typical nonmonotonic ideal point shaped response functions. Given then that the Self-Presented sub-scale was potentially more representative of its ideal point construction than the other sub-scales, it may have been expected that the GGUM item parameters would have produced the best fit to the response data. However, chi-square to degrees of freedom ratio showed that the GGUM item estimates accounted for only 5% of outright best fit, with both models performing similarly across the sub-scale. Interestingly, previous research (Roberts et al., 1999) has found that Likert developed items, representing a

relatively extreme position on a trait continuum, can be adequately fit by an ideal point model. The results of this study suggests that the when a Thurstonian approach to scale construction has been applied, monotonic functions generated by a polytomous dominance-based model, such as the GRM, may provide effective estimates of response patterns across the entire trait continuum.

This study adopted the preferable practice of cross-validation to evaluate model fit. This approach has not been used in previous studies given concerns regarding the size of the calibration sample. Regarding the fit methodology adopted, as Maydeu-Olivares (2005) posits previous cross-validation methods have used calibration samples in excess of 1,000 observations. This study provides an example of the successful modeling of personality data using only 510 respondents as a calibration sample. Given that the real-world application of personality questionnaires typically will not result in large sample sizes, this study indicates that the robust procedure of cross-validation can be successfully applied to smaller sample sizes, and result in adequate model-data fit outcomes.

The findings in this study add to what has previously been investigated regarding the application of IRT models to personality data. Generally, the conjecture has fallen into two camps, one arguing that findings suggest that traditional polytomous IRT models, typically the GRM, is unsuited to the specific nature of non-cognitive ability scales (e.g., Chernyshenko et al., 2001; Roberts et al., 2000), and the other finding that given the establishment of data unidimensionality, the GRM is the preferable option for modeling personality data (e.g., Maydeu-Olivares, 2005).

Further research is clearly needed in regards to the utility of an ideal point approach to personality scale construction. To date, only a few studies have attempted

to construct personality measures given the ideal point framework. As findings in this study alluded to, current ideal point scale construction practices may not be producing a truly valid ideal point scale.

These findings suggest that the GRM may provide adequate fit to scales reflecting an ideal point approach, providing support to Maydeu-Olivares' argument that dominance models, predominant in cognitive ability modeling, are equally appropriate with personality data.

CHAPTER FOUR

STUDY TWO: A LITERATURE REVIEW OF HUMAN, TECHNOLOGICAL AND ITEM DESIGN ISSUES IN COMPUTERIZED TESTING

By reviewing literature that has focused on both the human and technological issues, Study Two investigated the potential mode effect of computerized tests on performance. Empirical evidence has found that identical computerized and paper-and-pencil tests have not produced equivalent test-taker performance (e.g., Parshall & Kromrey, 1993; Gallagher, Bridgeman, & Cahalan, 2000). Further, while a computerized environment permits greater scope regarding innovative item design, such items requires a minimal task competency from the respondent beyond that being measured by the item. An overview is given of studies that have directly compared computerized and paper-based tests and which have sought to establish test mode effects such as reading rates, comprehension, and memory. Given the potential for innovation that the computerized environment offers, this review also examined the design of innovative item types and response formats, and the associated task constraints and minimal computer skills needed by the test-taker in order to respond to an item.

Thus, the research question for this study was:

What are the factors that produce a mode effect between paper-based and computerized tests?

Participant Issues

Race, Ethnicity, and Gender

An important area of interest is whether differences in performance for members of specific groups exist, however surprisingly little research has been directed at the fundamentally important issues of test fairness and adverse impact (Boodoo, 1998). In an earlier study, Parshall and Kromrey (1993) investigated the impact of demographic variables (e.g., gender, race, and age) to scores from both paper and computerized versions of the Graduate Record Examination (GRE). All three demographics were found to be associated with test mode, with computer-based administration favoring white males, although all other males performed better on the PPT version. Interestingly, female test-takers showed no difference in performance across either test mode. A potential confound of this study, however, was self-selection bias, as test-takers choose which test mode they wish to complete. An extensive study of differences in subgroup performance consolidated data from prior Educational Testing Service (ETS) research. Gallagher, Bridgeman, and Cahalan (2000) examined whether the differences between paper-and-pencil test (PPT) and computer-based test (CBT) performance were similar across subgroups, defined by race/ethnicity, gender, and native language. Although overall differences between the subgroups were small, some consistent patterns were found. For example, Gallagher et al. found that performance of African-American and Hispanic test-takers across computerized versions was either better than, or equal to, their performance across PPTs. Whereas across all tests, female test-takers' performance appeared to be slightly negatively affected by computerized tests. When gendered performance was compared within racial/ethnic groups, however, only differences for white female test-takers were

statistically significant, although the actual differences were small (Gallagher et al., 2000).

Cognitive Processing

Mead and Drasgow (1993) provided a meta-analysis of all the major research undertaken in the 1980s and early 1990s by comparing the computerized and paper-and-pencil versions of 123 timed power tests and 36 speeded tests. Their review of these 159 correlations obtained from PPTs of cognitive ability found that after correcting for measurement error, the estimated cross-mode correlations were .97 for timed power tests, and .72 for speeded tests. Mead and Drasgow concluded that test mode affected speeded tests, probably because of the longer time that test-takers typically took to read text from the screen. This was also found in previous studies (Dillon, 1992; Gould & Grischkowsky, 1984; Kak, 1981; Muter, Latremouille, Treurniet, & Beam, 1982; Smedshammar, Frenckner, Nordquist, & Romberger, 1989; Smith & Savory, 1989; Wright & Lickorish, 1983).

A more recent study by Mayes, Sims, and Koonce (2001) examined whether participants' memory and comprehension differed when reading an article from a Vision Display Terminal (VDT) compared to paper text. In contrast to earlier research, this study found that participants reading a paper copy of the article took longer to finish than those reading from the VDT (Mayes et al., 2001). Further, participants reading from the VDT were found to remember information just as well as those reading from a paper-based format, however, comprehension scores were lower for participants reading from the VDT.

Noyes and Garland (2003) examined comparable text across both test modes in terms of reading times, number of correct answers, and memory awareness ratings.

Adopting the Remember-Know learning paradigm, devised by Conway, Gardiner, Perfect, Anderson, and Cohen (1997), they found that memory awareness ratings were derived from participants' reflections of their initial recall of the answer (e.g., whether they received more than one type of memory in order to formulate their answer). This approach was adopted to avoid the confound that occurred in the Mayes et al. study due to the association between academic performance and memory abilities (Noyes & Garland, 2003). Congruent with the findings of Mayes et al., Noyes and Garland found no significant differences between matched computer and paper-based text in terms of time taken to read the material. However, in contrast to Mayes et al., Noyes and Garland found no significant differences in the level of comprehension achieved between test modes. Data showed statistically significant differences in memory awareness patterns, with the Remember frequencies from the computer group almost twice that of the Know frequencies, which suggested different cognitive processing took place when learning from VDTs and paper. Noyes and Garland suggested that these differences emanated from the physical characteristics of each condition, and suggested that further research is needed to establish how the various attributes of VDTs influence cognitive processing.

Ability

The interaction between an individual's ability levels and test mode has produced variable results. In examining perceived cognitive workload as presenting a potential impact on performance between test modes, Noyes, Garland and Robins (2004) asked participants to rate their perceived effort associated on a comprehension task, using the NASA Task Load Index (NASA-TLX; Hart & Straveland, 1988). Findings indicated that in addition to there being a greater perception of cognitive workload

required to complete the CBT, participants with low comprehension experienced greater perceived cognitive workload than participants with higher comprehension ability.

Another study examined participants' overall ability, based on academic placement categories: regular education, gifted education, or special education-learning disabled, in relation to performance on a PPT and equivalent CBT. When comparing scores obtained on both test modes, Poggio, Glasnapp, Yang, and Poggio (2005) found that participant's performance did not produce any unexpected main effect separations across academic classifications. Specifically, the participants' academic placement did not differentially influence their performance on either test mode.

Students' academic prior attainment, as measured by their total "A" level scores, and grade averages in Biology, Chemistry, Physics, and Mathematics, was identified by Watson (2001) as the key biological factor impacting on conceptual gains acquired from a computer assisted learning (CAL) generated simulation of an Introductory Genetics experiment. Findings showed that students that were both high academic achievers *and* frequent users of computers obtained the greatest conceptual gains from the computerized experiment simulations. Conversely, the lowest conceptual gains occurred amongst participants with lower than average "A" levels and subject grade averages, and who adjudged themselves to be less confident and infrequent computer users. While Watson (2001) proposed several factors that affected the conceptual gains acquired from the CAL experiments, two key biographical factors were previous academic attainment and familiarity with computers. A similar investigation by Clariana and Wallace (2002) examined the interaction of test mode and content attainment of freshman business undergraduates. Using equivalent paper-based and

computer-based tests, results from test data and test-takers' self-report information showed that high-attaining students performed significantly better on the CBT version, than on the paper-based version. Interestingly, neither test mode impacted on the test scores of the low-attaining students.

Familiarity with Computers

A variable that has been frequently cited as a major contributor to differences in test-taker performance is that of computer familiarity. Although Taylor, Kirsch, Eignor, and Jamieson (1999) claimed that this variable might be one of the most critical reasons behind differences in test mode performance, other studies have produced inconclusive evidence regarding the impact of computer familiarity on performance. In one of the first studies examining computer familiarity, Lee (1986) administered an arithmetic reasoning test via paper and computer, as well as a computer experience questionnaire to test-takers. Against expectations, Lee found that low- and high-use groups showed no significant differences in performance. Powers and O'Neill (1992) assessed students on computerized reading and mathematics questions. Split into two groups, one condition required test-takers to complete a pre-exam computer familiarization tutorial (e.g., how to use a mouse, how to navigate through the test), and allowed access to an online help function throughout the test, whereas the other condition gave test-takers the pre-exam tutorial only. Neither self-reported previous computer experience, nor did additional online help account for significant differences in computerized reading or mathematics scores. Powers and O'Neill concluded that the brief pre-exam training received by test-takers might have been sufficient to negate a pre-existing lack of familiarization with computers.

The primary rationale behind changing the Test of English as a Foreign Language (TOEFL) to a computerized version, was that the paper version of multiple-choice items only measured lower order processing skills, and not the higher order processing skills typically utilized in constructing and communicating meaning (Lynch, 2000). However, the Educational Testing Service (ETS) stated that the greatest danger to the proposed improved validity of the CBT version was the effect of computer familiarity (Kirsch, Jamieson, Taylor, & Eignor, 1998). After administering an on-line familiarization tutorial and controlling for ability, Kirsch et al. examined the relationship between levels of computer familiarity and performance on the computerized TOEFL. Results showed no relationship between test-takers' level of computer familiarity and performance on the CBT. No meaningful differences were found between male and female test-takers, or between test-takers tested at domestic and foreign sites (Kirsch et al., 1998). Due to practical constraints that restricted the use of a control group, it was unknown to what extent the tutorial eliminated or minimized performance differences due to prior levels of computer familiarity (Kirsch et al., 1998). However it is likely, as with the Powers and O'Neill (1992) study, that either pretest computer training negated the low pre-familiarity levels of test-takers, or computer familiarity may have played only a small part in performance, as it does not appear to have the significant impact once assumed. Regarding the former point, a number of authors have suggested that if computer familiarity is a key factor associated with the test mode effect, it may be rapidly diminishing with increased access to computers in schools and the home (Clariana & Wallace, 2002; Kirsch et al., 1998; Lynch, 2000; McDonald, 2002).

Computer Anxiety

According to McDonald (2002), "computer anxiety refers to the fear experienced when interacting with a computer or anticipating an interaction" (p. 305). Anxiety towards computers overlaps with the previously discussed construct of familiarity. Indeed, a large amount of literature on computer anxiety has reinforced the exposure hypothesis, whereby computer anxiety is a consequence of a general lack of experience and familiarity with computers (Levine & Donitsa-Schmidt, 1998). A comprehensive meta-analytic study by Chua, Chen, and Wong (1999) summarized the results of computer anxiety studies published between 1990 and 1996. Chua et al. concluded that results generally showed that computer anxiety was inversely related to computer experience.

In contrast, research has also produced findings that conflict with anxiety's inverse relationship to computer experience. Durndell and Lightbody (1994) examined anxiety levels against the increased use of computers over an eight-year period. A sample of 16-18 year olds was compared to similar groups in 1986 and 1989 on reported use of computers, and levels of anxiety experienced. Although reported use of computers in both the home and school environments had increased dramatically since 1986, results showed no evidence of a decrease in computer anxiety for either males or females (Durndell & Lightbody, 1994). In another study, Todman and Lawrenson (1992) failed to establish a relationship between computer experience and anxiety in their sample of first-year university students and 9-year-old school children. Using findings from self-report inventories measuring computer anxiety, mathematics anxiety, trait anxiety, and computer experience, results from both groups failed to show a relationship between anxiety about computers and levels of computer experience (Todman & Lawrenson, 1992).

As these studies show, the actual effect that computer anxiety has on performance on computer-based tests is generally unclear (McDonald, 2002). It is nevertheless well understood that anxiety reduces the capacity of the working memory; therefore, tasks that require only moderate levels of working memory may be successfully completed regardless of the anxiety level of the test-taker (McDonald, 2002).

User Interface – Legibility

Screen Size and Resolution

There have been rapid improvements in computer display technology, with higher resolution capabilities being one of the most prominent features of current monitors (Dillon, 1994). It has been suggested that the higher the screen resolution, the higher the readability of the text. Using data from fifteen previous experiments, Gould, Alfaro, Finn, Haupt, and Minuto (1987) investigated the relationship between display resolution and reading performance by plotting the within-subject ratio of reading rates on Cathode Ray Tubes (CRTs) versus paper copies. CRT reading speeds were equivalent to those on paper only in high resolution conditions (1000 x 800) (Gould et al., 1987). Harpster, Freivalds, Shulman, and Leibowitz (1989) came to a similar conclusion in their study of the effects of display resolution on accommodative accuracy and visual search performance. In the paper and high resolution (640 x 200) conditions, they found that both accommodative accuracy and visual search performance were significantly better than under the low resolution (320 x 200) condition.

Ziefle (1998) extended previous research on screen resolution by investigating the effects of visual information processing and display resolution on fatigue. Using a

paper resolution (255dpi) and two screen resolution conditions (1664 x 1200 and 832 x 600), results showed that reading performance was significantly better in the paper condition than in the two CRT conditions (Ziefle, 1998). Examining the two CRT resolutions on eye movement parameters (fixation, duration, and number of fixations) in a visual search task, Ziefle found that the higher the resolution the more optimal visual performance was, and the less visual fatigue was evident.

Combining screen size and resolution, Bridgeman, Lennon, and Jackenthal (2002) investigated the impact that variations in the amount and legibility of information displayed on a screen might have on test performance. Using verbal and math SAT I items, Bridgeman et al. adopted three display conditions: 17-inch monitor (1024 x 768), 17-inch monitor (640 x 480), and a 15-inch monitor (640 x 480). Findings showed that low screen resolution impacted negatively on the verbal scores for both male and female test-takers, however, no main effects nor interactions were found on mathematic performance. Test-takers in the high resolution (1024 x 768) condition performed better than either of the 17-inch or 15-inch screens low resolution (640 x 480) conditions. As the amount of scrolling required changed given each condition, so too did the test-takers' ratings on the assumed effect of scrolling on their performance. Under the 15-inch (640 x 480) condition 18% of test-takers reported that they felt that scrolling had interfered with their performance, compared to only 8% under the 17-inch (640 x 480) condition, and 6% under the 17-inch (1024 x 768) condition. Bridgeman et al. concluded that the major differences between the high and low resolution conditions were due, potentially, to the "substantially greater need to scroll through reading passages in the low resolution condition" (p. 23).

Font Characteristics

Another dimension of interest is the font used in presenting information – both on paper and onscreen. Fonts are typically divided into three general visual categories: serif, sans serif, and ornate fonts. Serif fonts are distinguishable by the small curls or appendixes at the end of each letter, whereas sans serif fonts are characterized by straight lines with no curls or appendixes. Ornate fonts are fonts that do not fit the previous two categories.

Within traditional print, there is a commonly held view that serif fonts are easier to read (Tullis, Boynton, & Hersh, 1995). This is based on the idea that the serifs (appendixes) lead the eye through the text, providing quicker recognition to the brain due to their more distinctive shape (Byrne, 2004). This ease in processing typically translates to faster reading speed and comprehension (Dillon, 1994). It is for this reason that most printed documents use serif fonts, especially when large chucks of text are presented (Byrne, 2004; Wakeman, 2000). By comparison, the cleaner, bolder, and less busy character shapes of the sans serif fonts have typically been adopted for display type text, such as titles, headlines, and advertising slogans (Byrne, 2004). Under high resolution settings, the same font choice rules apply to electronic text (Boyarski, Neuwirth, Forlizzi, & Regli, 1998; de Rossi, 2002).

The two most commonly used fonts on the web and for electronic documents are the serif font Times New Roman (TNR) and sans serif font Arial (Bernard & Mills, 2000). The popularity of these fonts is interesting given that they were designed for the high resolution possible on paper, and tend to break up at smaller sizes (< 10pt) on the lower resolutions typically found on screen. It has been suggested that the continued prominence of these two fonts has more to do with their historically widespread use as paper fonts, than their actual suitability for the computer screen

(Lynch & Horton, 2002). Two other popular electronic fonts, that are highly legible on lower screen resolutions, are Verdana and Georgia. This popularity is not surprising as these fonts were specifically designed for legibility on the computer screen. Based on exaggerated x-heights (height of the lower-case "x" letter) their design means that they are a larger font compared to the traditional paper designed fonts of the same point (pt) size. In addition, the spaces between the Verdana and Georgia characters are expanded, making the character shapes more distinctive regardless of whether serifs are present (Tullis, et al., 1995).

A study by Bernard, Mills, Peterson, and Storrer (2001) compared the twelve most used fonts for electronic text in regards to reading speed, legibility (reading time/accuracy), and perception of font legibility. In keeping with the majority of web sites, font size was kept constant at 12pt across all fonts. Table 7 shows the twelve fonts studied.

Table 7
The twelve fonts compared in Bernard, Mills, Peterson, et al. (2001) study

Sans Serif Fonts	**Serif Fonts**	**Ornate Fonts**
Agency FB (Agency)	Century Schoolbook (Schoolbook)	Bradley Hand ITC (Bradley)
Arial	Courier New (Courier)	*Monotype Corsiva (Corsiva)*
Comic Sans MS (Comic)	Georgia	
Tahoma	Goudy Old Style (Goudy)	
Verdana	Times New Roman (TNR)	

Examination of mean reading time for each font type revealed that Tahoma and TNR provided the most efficient reading rates, with Courier, Bradley, and Corsiva producing the most inefficient reading rates amongst participants (Bernard, Mills, Peterson, et al., 2001). Interestingly, neither Verdana nor Georgia fonts facilitated

reading speed, despite being specifically designed for the computer screen. Reading legibility showed no significant font type effects, which suggests that fonts at 12pt are fairly robust in their ability to accurately convey meaning to the reader (Bernard, Mills, Peterson, et al., 2001). However, as a high resolution setting (1024 x 768) was used throughout the study, Bernard, Mills, Peterson, et al. suggested that both traditional and computer specific, serif and sans serif fonts would have presented well.

The size of text characters may also impact on reading speed and legibility. Tullis et al. (1995) examined the readability of four different fonts (Small Fonts, Arial, MS Sans Serif, MS Serif) in sizes from 6.00 to 9.75pts (sizes within the Microsoft Windows environment). All of the fonts, with the exception of 8.25pt MS Sans Serif, performed well in terms of reading speed and accuracy across the size range of 8.25pt to 9.75pt. Due to their poor performance, Tullis et al. advised that 7.5pt Arial and Small Fonts (6.0 – 6.75pts) should generally be avoided.

Examining larger fonts, Bernard, Lida, Riley, Hackler, and Janzen (2002) compared the top eight most commonly used electronic fonts at 10pt, 12pt, and 14pt sizes in terms of reading effectiveness (reading time/accuracy), reading time, perceptions of font legibility, and font attractiveness. No significant font size or type effects were found for reading efficiency. Conversely, reading time revealed significant font type and size differences, with TNR and Arial being read significantly faster then Courier, Schoolbook, and Georgia, and fonts at 10pt size read significantly more slowly than fonts at 12pt size. Participants' perceptions of legibility showed a significant font type x size interaction, with 10pt Tahoma perceived as the most legible font, closely followed by 10 pt Georgia and 12pt Courier. The poorest perceived legibility was 14pt and 10pt Comic, and 10pt Schoolbook. Findings also revealed significant differences in perceived font attractiveness. Georgia was

perceived as being more significantly attractive (irrespective of size) than Arial, Courier, and Comic, while TNR was perceived as significantly more attractive than Courier.

Another factor that should be taken into account when choosing a font relates to the characteristics of the audience. Only one study has specifically focused on children's preferences for different types and sizes of fonts. Bernard, Mills, Frank, and McKown (2001) examined four types of fonts at 12pt and 14pt sizes in order to determine the font combination most readable and most preferred by children (aged 9 to 11 years). Fonts examined were those that are commonly seen in children's text, namely the serif fonts – TNR and Courier, and sans serif fonts – Arial and Comic at 12pt and 14pt sizes. Although no interactions were significant, analysis of perceptions of reading ease indicated that TNR was perceived as being significantly less easy to read than Arial and Comic fonts (Bernard, Mills, Frank, et al., 2001). Perceptions of reading faster with a particular size of font revealed a significant main effect for font size favoring fonts at 14pt. Bernard, Mills, Frank, et al. found perceptions of font attractiveness again revealed a significant main effect for the 14pt font size, with TNR perceived as significantly less attractive than the Comic font. Overall, the children in this study most preferred the 14pt Arial and the 12pt Comic font types (Bernard, Mills, Frank, et al., 2001).

Line Length

Line length is also another important element in the presentation of text. An ideal line length for text is based on the physiology of the human eye. Because of the restrictions imposed by a small area on the retina called the macula, which is used for high visual acuity, the visual field arc covers only a width of approximately eight

centimeters at normal reading distances (Lynch & Horton, 2002). This coverage equates to around 12 words per line, or about 70 characters per line (cpl). Findings from research examining line length and the relative legibility of text on paper have produced mixed results. For example, Spencer (1968), in a study of reading rates and comprehension, concluded that line length is best at 70cpl, whereas other studies have found varying optimal lengths, with a range of 52cpl to 70cpl typical (Rayner & Pollatsek, 1989; Tinker, 1963). Findings from these studies suggest that line lengths exceeding 70cpl require readers to move their heads slightly or strain their eye muscles to track over the long lines of text. As a result, reading speed slows and retention rates fall as the long visual trip back to the left margin often causes the reader to lose track of the next line (Horton, 1989).

However, most web pages violate the suggested line lengths proposed by book typography research, with lines of text on the majority of web pages far exceeding the optimum for reading. Research on electronic line length has produced conflicting findings as to what constitutes optimal text layout. Some studies examining onscreen reading rate found that longer line lengths (75cpl and 100cpl) were read faster than very short lines, probably due to chunking (Duchnicky & Kolers, 1983; Dyson & Kipping, 1998). Duchnicky and Kolers (1983) studied the effect of line length on reading times from scrolled text on VDTs. The lines of text were of three different lengths: full-screen width, two-thirds-screen width, and one-third-screen width. Each of the three lengths was used with an 80- (average of 74.8 cpl) and a 40- (average of 12.5 cpl) character line set. Results found that for both character line sets the longer line lengths resulted in faster total reading times for the passage. Reading speed increased 28% from the one-third-screen width to the full-screen width. In sum,

Duchnicky and Kolers (1983) found that longer lines of text are read more efficiently from VDTs than shorter lines.

Dyson and Haselgrove (2001) suggest that these findings may be due to two differences between these modes of reading. First, there are the differences in visual angles between paper and screen. In particular, the reader of text onscreen tends to sit further back from the text than does the reader reading text from paper. Therefore, "a longer line length onscreen may subtend a similar visual angle to a moderate line length in print" (Dyson & Haselgrove, 2001, p. 588). No studies, however, appeared to have measured this phenomenon. Second, the impact of scrolling increased when participants were required to read narrower passages of text. Dyson and Haselgrove (2001) reported that reading rates may be faster at longer line lengths onscreen as readers are able to spend less time in scrolling movements and more time in uninterrupted reading. However, both Dyson and Kipping (1998) and Dyson and Haselgrove (2001) found that participants did appear to adjust their scrolling patterns according to the line length. Specifically, participants continued to read while scrolling at shorter line lengths, and stopped reading to scroll at longer line lengths. Interestingly, Lynch and Horton (2002) and Piolat, Roussey, and Thunin (1997) hypothesized that the time spent pausing to scroll in longer line lengths may in fact help in the consolidation of reading material. However, no research has investigated this phenomenon.

The effect of line length on comprehension has provided inconclusive results. Neither Duchnicky and Kolers (1983) nor Dyson and Kipping (1998) found differences in comprehension based on line length, however the same comprehension test was used in both studies, therefore these findings may be a function of the test used. Using a more elaborate test of comprehension requiring participants to recall

details and make inferences, Dyson and Haselgrove (2001) found that line length did influence readers' comprehension, with text at medium line length (55cpl) producing higher comprehension scores than at the short (25cpl) or long (100cpl) lengths.

Bernard, Fernandez, and Hull (2002) provided the only investigation of the potentially differing reading speeds and preferences of line length of children (9 to 12 years old) and adults (18 to 61 years old). Presenting three line-length conditions – 132cpl, 76cpl, and 45cpl – Bernard, Fernandez, et al. found no significant differences in mean reading time for either children or adults. Adult participants preferred the medium (76cpl) line length, with the short line (45cpl) length most preferred by children. Interestingly, both adults and children liked the full-length (132cpl) condition the least, even though this condition required only minimal amounts of scrolling.

Number of Lines

Unfortunately, little research has focused on the effect of the number of lines of text presented on reading performance. Duchnicky and Kolers (1983) investigated the effect of the number of lines displayed onscreen on reading times. Displaying conditions of 1, 2, 3, 4, or 20 lines of text, they found that reading speed increased by 9% when the number of lines displayed was increased from 1 to 20 lines. Further, results showed that the conditions 4 and 20 lines of text were read faster by participants, than the conditions of 1 and 2 lines of text. Duchnicky and Kolers concluded that 4 lines of text on a screen at a time were read as efficiently as a full screen of text.

Interline Spacing

The amount of whitespace that occurs *between* lines of text is referred to as leading, or interline spacing. Kolers, Duchnicky, and Ferguson (1981) sought to measure the eye movements of participants as they read texts with two different interline spacing conditions (single versus double). They found that single interline spacing resulted in significantly more fixations per line, therefore resulting in fewer words read per fixation. Further, total reading time was slightly longer under the single interline spacing condition, reducing total reading time by 2% (Kolers et al., 1981). This finding supports the argument by Morrison and Inhoff (1981) that an increase in the blank area between lines decreases lateral masking (the interference of surrounding letters on word perception), resulting in more accurate return sweeps to the next line. In part of their study, Kruk and Muter (1984) examined the effects of varying interline spacing on speed of reading from a video monitor. As found in Kolers et al.'s 1981 study, results showed a significant difference between conditions with mean reading speed being 10.9% slower in the single interline spacing condition than in the double interline spacing condition.

Only a few studies appear to have focused on the potential impacts of interline spacing, and they were conducted over two decades ago. Therefore, it is unknown how applicable the previous findings are, given today's screen technology and font design. It appears, however, that current information follows the logic of these studies' findings. For example, Lynch and Horton (2002) argue that interline spacing strongly affects the legibility of text blocks, with too much space making it difficult for the eye to locate the start of the next line, and too little space confusing the lines of type, as the ascenders of one line get jumbled with the descenders of the line above. According to Lynch and Horton (2002) the typology rules for interline spacing in

print are directly transferable to screen. Specifically, interline spacing of text blocks should be set at approximately 2 points above the size of the font. For example, a 12pt font should be set with 14pts of interline spacing (Lynch & Horton, 2002). Further, Lynch and Horton suggested that adopting generous interline spacing should compensate for screen situations where there are long line lengths and lower resolution settings, for example, 12-point type with 14 to 16 points of interline spacing.

White Space

Another issue connected to line length relates to the amount of screen visible given the width of the text passages, namely white space (Lynch & Horton, 2002). For traditional printed text there is a rule that the use of white space adds not only to the attractiveness of the text, but aids in directing the viewer's attention to the regions where important information is present (Mullet & Sano, 1995). As a result, white space helps in preventing the influence of distracting, unimportant information by spatially organizing information onscreen (Lynch & Horton, 2002). Early research on the effect of white space on computer displays (van Nes, 1986; de Bruijn, de Mul, & van Oostendorp, 1992) tended to support the layout logic from the printed media. It is important, therefore, to establish how much white space should be used. Spool, Scanlon, Schroeder, Snyder and DeAngelo (1999) conducted a usability study of eight major commercial web sites by creating a textual "scavenger hunt", whereby participants had to answer four types of questions (simple facts, comparison of facts, judgment, comparison of judgments) on each site. Spool et al. found that the more white space a site had, the poorer participants performed in terms of the success in finding information. However, this study has some weaknesses in the methods and

measures employed. For example, differing display formats and information were presented at each web site, with only a subjective estimation given of how much white space was present on each web page. The methodology adopted by Bernard, Chaparro and Thomasson (2000) solved both of these issues by creating three content identical web sites that varied only in the amount of measurable white space displayed. Participants were required to find errors and hyperlinks. The Low condition presented little white space; with one character space (3mm) separating one column of text from another, with no additional space between paragraphs. The Medium condition had four character spaces (9mm) between each column and a blank line (9mm) between each paragraph. The third condition, High, had eleven character spaces (19mm) between each column, with four blank lines (19mm) between each paragraph. Although no significant differences were found in the time taken to find errors or hyperlinks amongst the three conditions, participants were significantly more satisfied with the amount of white space found in the Medium condition.

McMullin, Varnhagen, Heng and Apedoe (2002) investigated the effects of line length and white space on the comprehension of text presented on the web. The four conditions used combined line length, either narrow (55cpl) or wide (115cpl), and white space variations, either one- or two-columns. A main effect for white space was found, with participants obtaining higher comprehension scores in the one-column presentations across both line lengths. However, the effects of line length and visual fixations might have confounded this result.

User Interface – Interactive

Scrolling

Perhaps the most notable difference between reading text from paper versus a computer screen lies in the ease with which paper can be manipulated. Manual dexterity skills such as using fingers to turn pages, keeping one finger on a section as a location aid, and flicking through a few pages while browsing the contents of a test, are acquired early in a reader's life and are transferable to all document types (Dillon, 1994).

Obviously, such second-nature tasks are not possible or are difficult to replicate within the CBT environment, with manipulation of electronic text typically involving the use of the mouse to move a scroll bar. Lovelace and Southall (1983) suggested that readers establish a visual memory for where an item is spatially located on a page. It is proposed that scrolling weakens the relationship between an item and its location, offering the reader only relative positional cues that the item may have with its immediate neighbors (Dillon, 1994). Choi and Tinkler (2002) evaluated the ability of third- and tenth-graders to read and comprehend long passages onscreen and on paper. Here, the amount of reading material in a passage was more than what could be presented in a single screen, forcing test-takers to use the vertical scroll bars to read through the entire passage. Differential patterns in item difficulty showed that reading items were more difficult in their computerized form for both grades, especially impacting negatively on third-grade test-taker performance. Items that required test-takers to scan text by interrelating a key word or phrase with a key word or phrase in a test question, were more difficult for third-grade test-takers on the computerized test version (Choi & Tinkler, 2002). Choi and Tinkler suggested that the need to scroll

through reading passages on the computer screen presented the most obvious difference in item presentation. Furthermore, Choi and Tinkler suggested that the computerized reading items might have had less of a mode effect if an electronic marker had been made available, allowing students to highlight a selection of text in a passage, thereby providing a closer emulation of natural PPT-taking practice.

Item Review

The effects of item review in PPTs have been studied for some 70 years, and results have consistently indicated that the majority of test-takers will change only a few of their responses during the review, with test scores usually improving as a result (Revuelta, Ximénez, & Olea, 2003). Item review refers to the ability to review, skip, and/or change items. Although the availability of item review is a taken for granted feature of PPTs, this option is often not available, or even feasible, for many CBT designs. If the CBT is based on the format where items are tailored to a test-taker's estimated ability after each response, such as computer adaptive tests (CATs), then the ability to allow reviewing is highly problematic. When review capabilities have been allowed on CATs this has led to increased testing time, reductions in measurement precision, complicated item administration algorithms, and lower test score validity due to answer response strategies that can yield inflated ability estimates (Vispoel, 1998). In contrast, although computerized fixed-item tests (CFITs) typically prohibit item review, structural reviewing can be allowed with the help of sophisticated algorithms. Unfortunately, there are few studies to date that address the effects of allowing item review on the results obtained from CFITs (Vispoel, 1998), and as the following studies demonstrate, findings have not been definitive.

Using two 50-item tests consisting of only selected-response items Eaves and Smith (1986) allowed their PPT group the usual flexibility of being able to move around freely from item to item, to change the order of items answered, and to review and change answers. Conversely, the CBT group was presented only one item stimulus at a time, and once an item had been answered, test-takers were unable to review or change that response (Eaves & Smith, 1986). Results showed that test flexibility did not affect test-taker performance, with no differences found between the two item flexibility conditions (Eaves & Smith, 1986).

Spray, Ackerman, Reckase and Carlson (1989) hypothesized that differences in paper and computerized performances on a test may be due to the differing flexibilities each offers test-takers. Using tests devised for the Marine Corps Communication-Electronics School (MCCES) Spray et al. developed test software that mimics, as closely as possible, the inherent flexibility of the PPT format. Test-takers were permitted to navigate through the test freely, and to review and change previous responses. Mean and cumulative score distributions for the CBT version were not significantly different to the PPT version, with no item bias due to mode effects found. Spray et al. concluded that score equivalence between the item presentation media could occur when test flexibility equivalence is achieved for both computer and paper versions. In a similar study, Luecht, Hadadi, Swanson, and Case (1998) investigated the effects of test flexibility using three content and statistically parallel versions of the National Board of Medical Examiners' (NBME) Comprehensive Basic Sciences Examination (CBSE). One version of the CBSE was a PPT design, with the other two CBSEs being designed as operationally distinct; one allowing test-takers to navigate to any item, go back to review, and change previously presented items; and the other not allowing any test flexibility (Luecht et al., 1998).

As with the Eaves and Smith (1986) study, findings showed that lack of test flexibility did not negatively impact on test-taker performance. Further, test-taker performance on the PPT CBSE did not differ substantially from either of the two computerized versions. Interestingly, a follow-up survey showed that 20% of test-takers felt that the "no review" feature was what they liked *least* about that version of the CBT (Luecht et al., 1998). Vispoel (2000) argued that there were several reasons why allowing reviewing/changing/skipping of items may be advantageous: reviewing and changing items may increase test score validity as test-takers are able to change incorrect answers resulting from typing errors, misreading of items, temporary lapses in memory, and reconceptualisations of answers to previous items.

Among several areas of investigation, Vispoel (2000) sought to explore the possible interaction between test anxiety and test flexibility (permitting or not permitting item review). In addition, Vispoel compared the answer changes before and after review in the review condition, and compared the answer-changing behavior of test-takers within the low- and high-ability groups. Data consisting of vocabulary items from the Iowa Tests of Education Development (ITED) revealed that 45% of test-takers took advantage of the opportunity to review answers. However, on average less than 4% of test-takers changed their answers, with test-takers changing more answers from incorrect to correct than from correct to incorrect by a ratio of 2.25 to 1 (Vispoel, 2000). Results of test-taker ability levels showed that as ability level increased, testing time and the number of answer changes decreased, with an increase in the incorrect to correct ratio (4 to 1). The study did not find a main effect or Test Anxiety x Review interaction.

Item Presentation

There have been some post hoc suggestions that presenting items individually, as is typically done in CBT formats, may have a detrimental effect on performance (Dimock & Cormier, 1991). A study by Hofer and Green (1985) showed that groups of items presented onscreen might lead to more hurried responses resulting in less attention to detail than on a PPT. Conversely, both Greaud and Green (1986), on a test of clerical skills, and Lee (1986), on a test of arithmetic reasoning, found a facilitative effect when CBT items were presented in groups. In one of their experiments, Dimock and Cormier (1991) sought to determine whether presenting items in groups on a PPT versus presenting items individually on a CBT would impact on performance. In order to avoid the potential confounds of computer familiarity or anxiety, Dimock and Cormier presented Verbal Reasoning items via index cards (7.5cm x 12.75cm) to mimic the presentation of individual computerized items. Results showed that test-takers scored significantly higher on the PPT format than on the card format, suggesting there are format differences between presenting groups of items and individual items (Dimock & Cormier, 1991). However, a potential confound in this study lies in the novelty of being presented items on cards. For example, in a second experiment, Dimock and Cormier did not find performance on items presented via index cards to be equivalent to items presented via computer. Unfortunately, no subsequent studies have investigated the impact of item presentation differences between modes.

Computerized Item Design Types

Item Format and Response Actions

Using categories based on the mechanism by which items are answered – mouse or keyboard, the following sections will review some of the item format types available. Where available in literature, empirical evidence will also be offered.

Mouse-based Response Action

Many designers of innovative items have taken advantage of the flexibility that the mouse and onscreen cursor can offer item design. Items can be designed to include pull-down menus, arrow keys, and scroll bars. Further, mouse and cursor availability allows text to be highlighted and text/graphics and figures to be moved. Using some of these actions, the following represents some of the most commonly used mouse-based response items.

Figural Response Items

Test-takers respond to these items by selecting part of a figure or graphic, which is then dragged to a position on a grid or area, given the parameters or constraints in the item stem (Zenisky & Sireci, 2002). The advantage of this item type lies in its ability to assess knowledge that is difficult to tap by either verbal or quantitative representations, or by more static means of testing (Martinez & Jenkins, 1993). O'Neill and Folk (1996) refer to math item types where test-takers select a figure or graphic and place it on a histogram, scale, or dial. An example is shown in Figure 3, where a figure is selected in order to represent the given mathematical function.

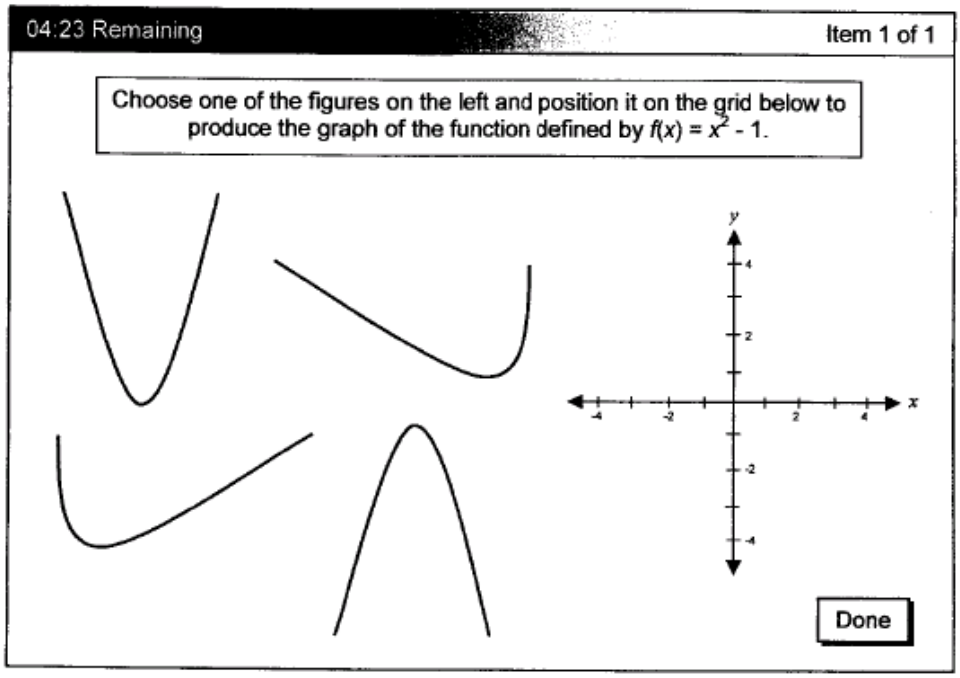


Figure 3
An example of a figural response item (Zenisky & Sireci, 2002).

Martinez (1993) compared figural response (FR) type items in an architecture exam with multiple-choice items, for their ability to predict architectural problem-solving proficiency. In a problem requiring test-takers to design a preschool in the virtual three-dimensional space represented by a contour map, only FR items were significant predictors of performance. Subjective judgments from test-takers found that FR items were believed to be more like what an architect does (77.8%) than the multiple-choice items, with 67.5% of test-takers judging FR items to be better indicators of the subject's knowledge of architecture. Therefore, face validity appears high for figural response items.

In a pilot study investigating items designed for the domain of cell and molecular biology, Martinez and Jenkins (1993) examined the relationships between FR and open-ended verbal questions, as measures of figural and verbal ability. Figure 4 shows an example of the biology FR items designed for the pilot study.

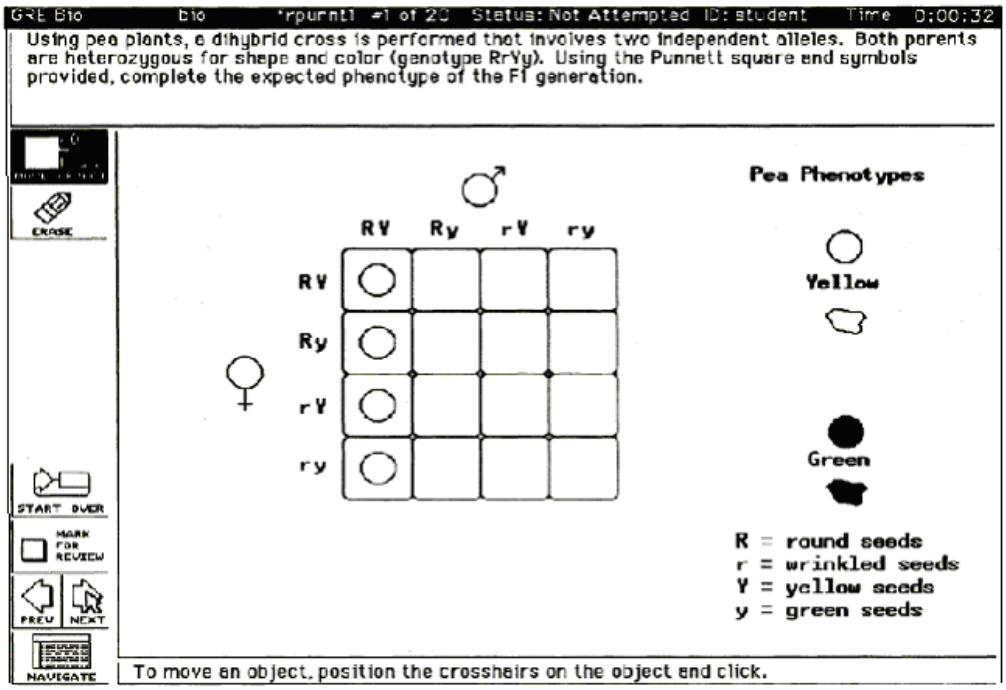


Figure 4.
An example of a figural response item (Martinez & Jenkins, 1993).

Findings revealed that FR items were better able to distinguish between experts and novices than the open-ended verbal response items. Further, FR scores were related to both figural and verbal aptitudes, whereas verbal response scores related only to verbal aptitude. Martinez and Jenkins (1993) propose that FR items may, due to their verbal stems, require a cognitive interplay between figural and verbal symbol systems.

Drag-and-drop Item

This item type presents a scenario or problem, which requires test-takers to use a mouse to click and drag the correct object/s to the appropriate position on the screen. The drag-and-drop (DD) item shown in Figure 5, asks test-takers to drag the five elements given to their correct position on the Periodic Table.

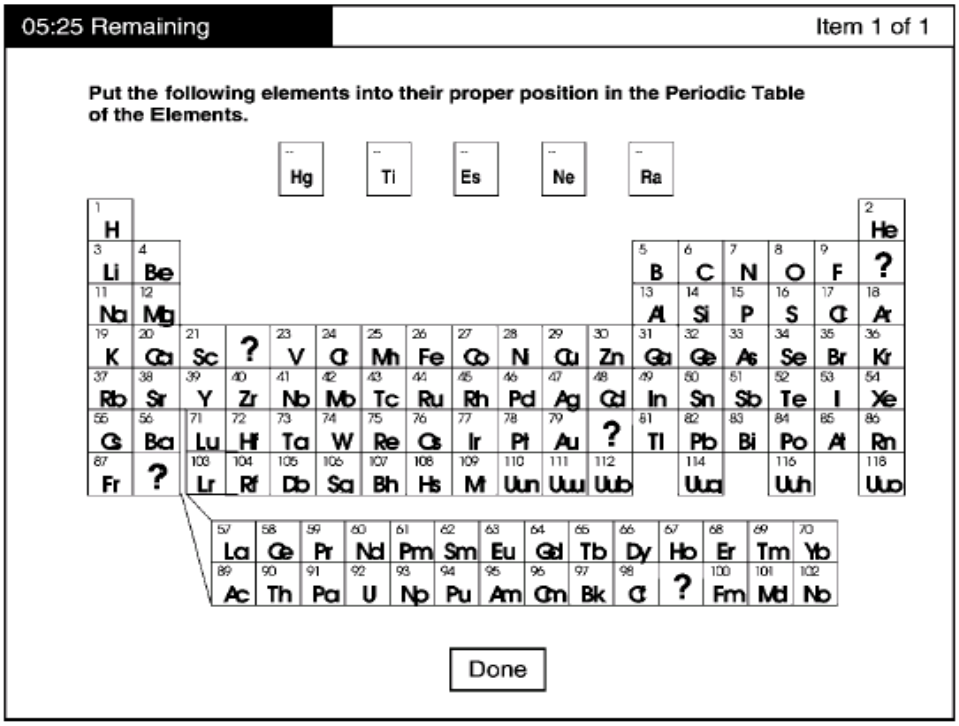


Figure 5.
An example of a drag-and-drop item presenting the Periodic Table of the Elements with missing elements (Zenisky & Sireci, 2002).

The DD item is an example of a largely non-text item, where graphics are used to represent both the problem or scenario, and response options. These items can be presented in a more elaborate fashion as shown in Figure 6, where this item is part of a longer case study question in a Microsoft certification examination.

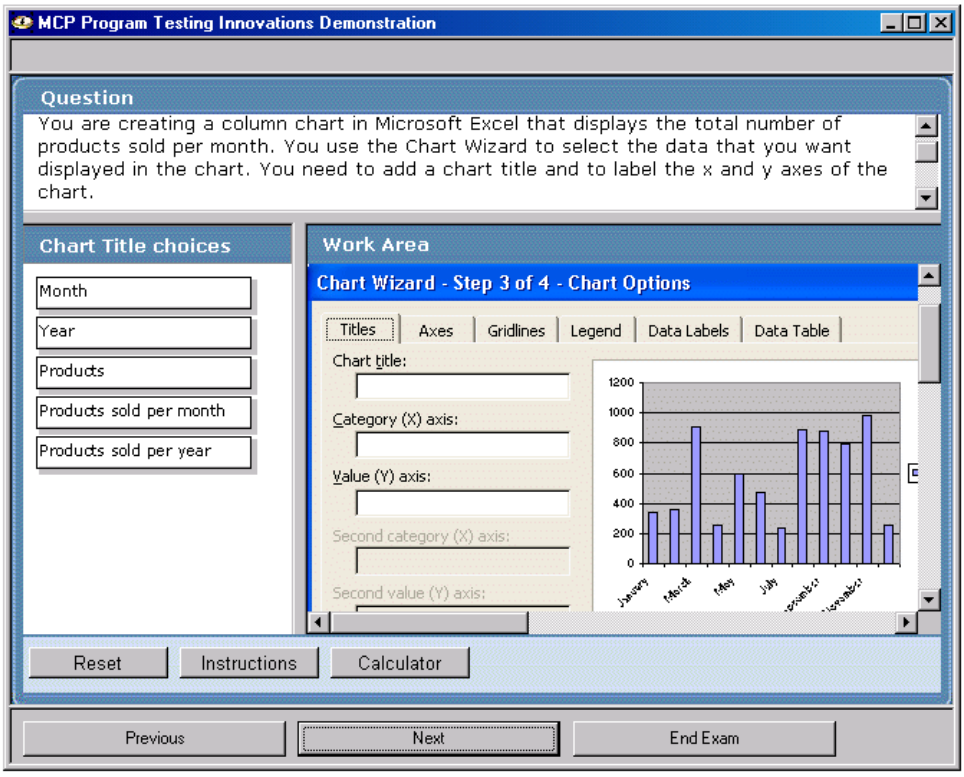


Figure 6.
A drag-and-drop item requiring test-takers to select and drag their chart title choices to their correct position on the work area of the item (Microsoft Corporation, 2003).

Graphical Modeling Item

This item type requires test-takers to respond by plotting points on a grid. This constructed response item was devised by Bennett, Morley, and Quardt (2000) in order to broaden the range of mathematical problem solving possibilities for the computerized Graduate Record Examination (GRE) General Test. An example of this item type is shown in Figure 7.

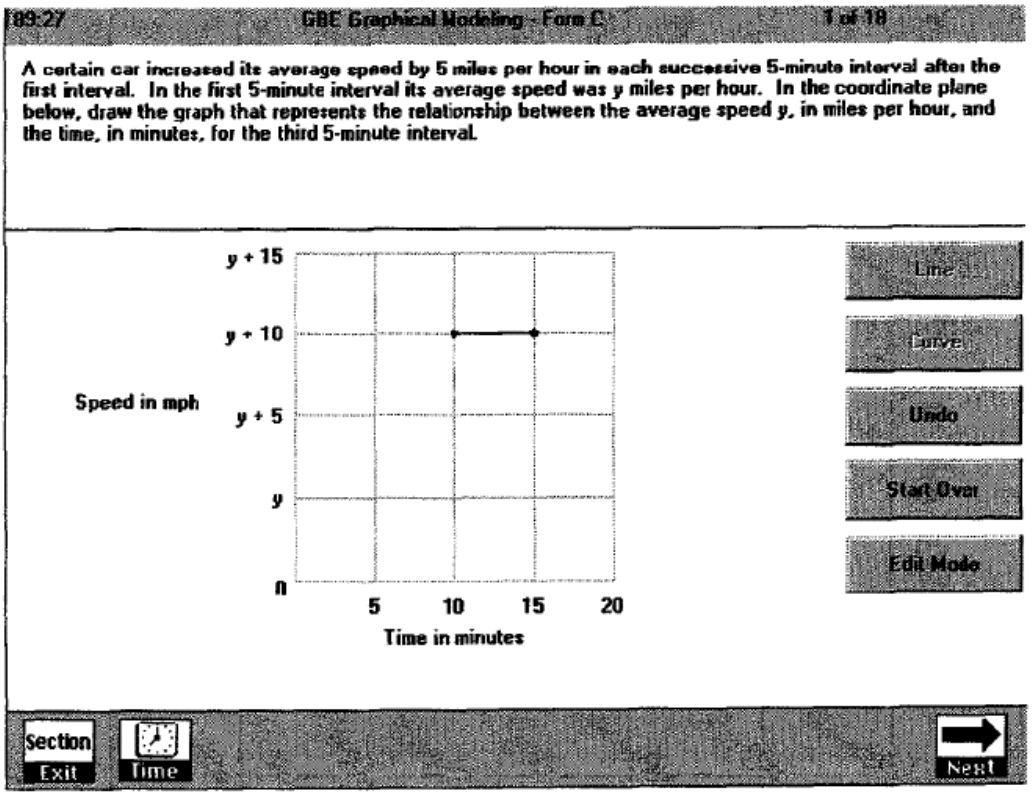


Figure 7. An example of a graphical modeling item format (Bennett, Morley, Quardt, & Rock, 2000).

The Bennett, Morley, Quardt, and Rock (2000) study of the graphical modeling (GM) format, focused on: psychometric characteristics, features that might affect item difficulty, effect of gender, and test-taker perceptions of these items. Psychometric analysis revealed that the GM test showed high reliability (low .90s). Scores were moderately related to GRE quantitative performance; however, GM's disattenuated correlation with the GRE quantitative test did not approximate unity. Bennett et al. suggested that the GM test might require skills that are independent of those required by the GRE quantitative test. Regarding the exploratory difficulty analysis, the manipulated feature of problem structure showed a dependable effect. No significant gender differences, independent of those already known to be associated with the GRE quantitative section, were detected (Bennett et al., 2000). Interestingly, participants preferred the regular multiple-choice graphical reasoning items to the GM

items; however, test-takers thought the GM items were a fairer indicator of their ability to undertake graduate study (Bennett et al., 2000).

Drag-and-connect Item

The threefold design of this item format consists of a question or problem statement, a pool of possible entities that may be connected, and a pool of possible connection links between the entities (Jodoin, 2003). Test-takers are required to identify and connect the relationships or associations between entities. Several movable objects can be arranged in different locations on the answer area. Figure 8 shows one of the drag-and-connect (DC) items analyzed by Jodoin (2003).

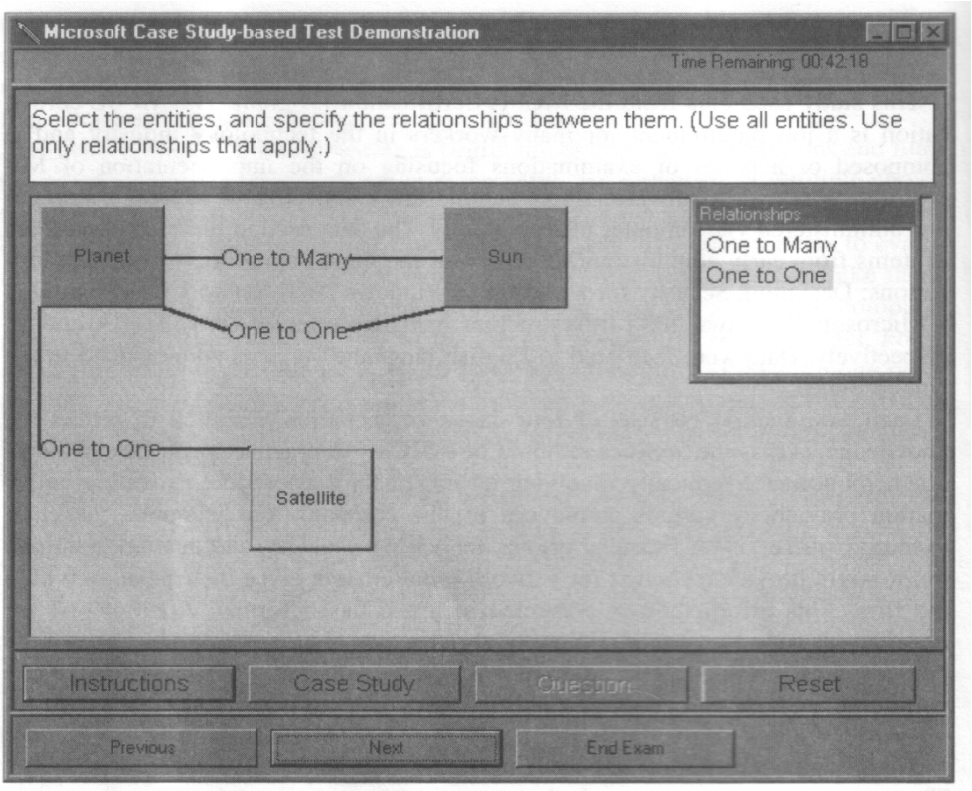


Figure 8.
An example of a drop-and-connect item format (Jodoin, 2003).

Using items from the Microsoft Certified Systems Engineers (MCSE) Certification Program (see Figure 8), Jodoin compared the reliability of DC items against the test's multiple-choice items. Using IRT, DC items were found to provide

considerably more information, therefore greater reliability, than multiple-choice items across all ability levels. However, the DC items took longer to complete than the multiple-choice items, thus resulting in less information per minute of testing time than their multiple-choice counterparts (Jodoin, 2003).

Specifying Relationships Item

An extension of the DC items described above is the specifying relationships (SR) item type. This item format requires the test-taker to move objects so as to link them in a flowchart through selecting their correct relationships with other objects (Zenisky & Sireci, 2002). Figure 9 shows an SR item, where the relationships of the three states of water need to be connected to the appropriate scientific process.

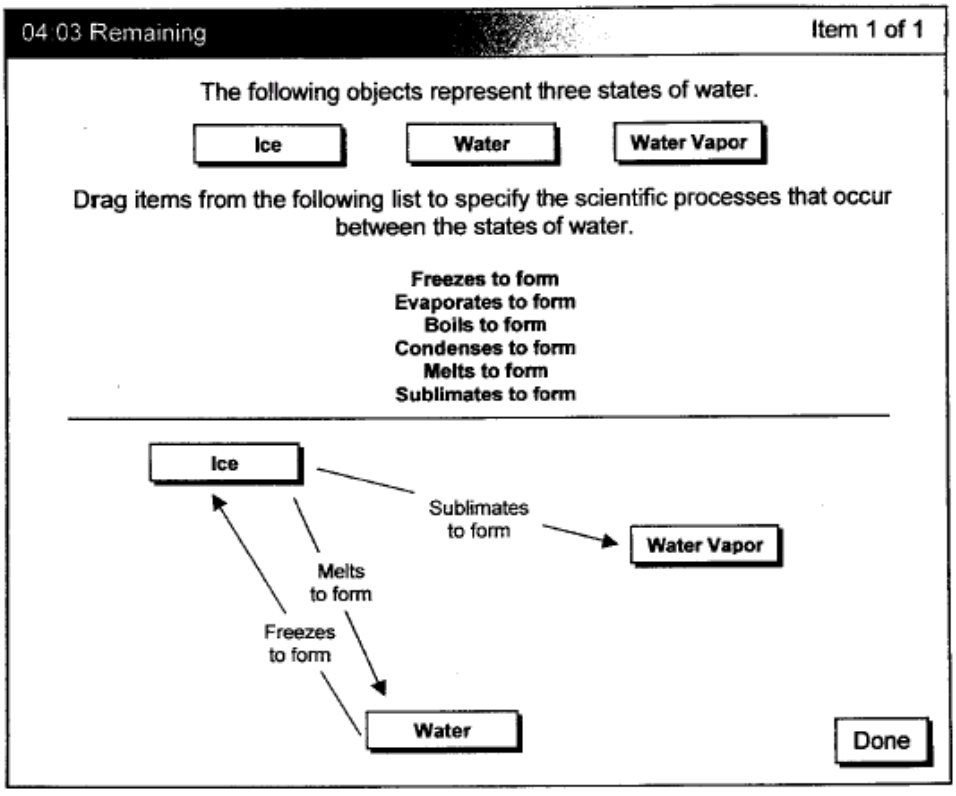


Figure 9.
An example of a specifying relationships item type (Zenisky & Sireci, 2002).

Create-a-tree Item

The create-a-tree (CT) item format presents the test-taker with a question or problem statement, and specifies the way in which the test-takers should order the elements in the process (Zenisky & Sireci, 2002). The correct solution steps are moved to the appropriate place in the diagrammatic outline or tree structure (see Figure 10) (Jodoin, 2003).

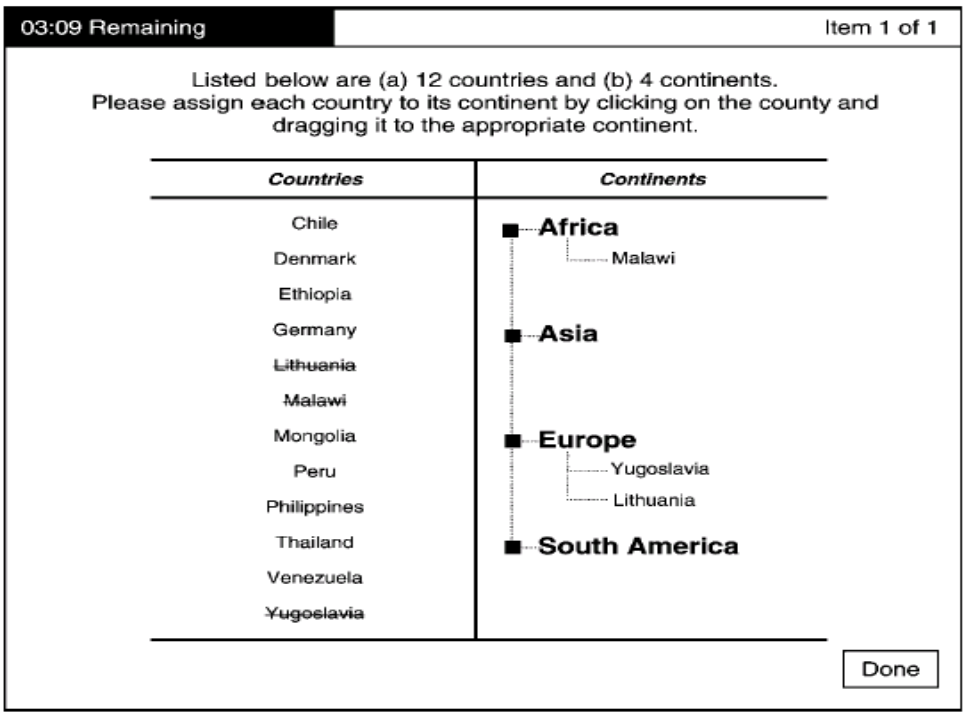


Figure 10. An example of a create-a-tree item type (Zenisky & Sireci, 2002).

The CT item format was also analyzed by Jodoin (2003), and as with the DC items examined, CT items provided more information than multiple-choice. However, CT items also took longer to complete than the multiple-choice items.

Capturing Frames Item

This item format requires the test-taker to click on portions of a graphic or video frame in order to capture the best representation of what is required. An example of a

capturing frames (CF) item is shown in Figure 11. This item requires the test-taker, after playing an animated map of the Middle East's borders since World War II, to capture the frame that illustrates the specifics of the item stem.

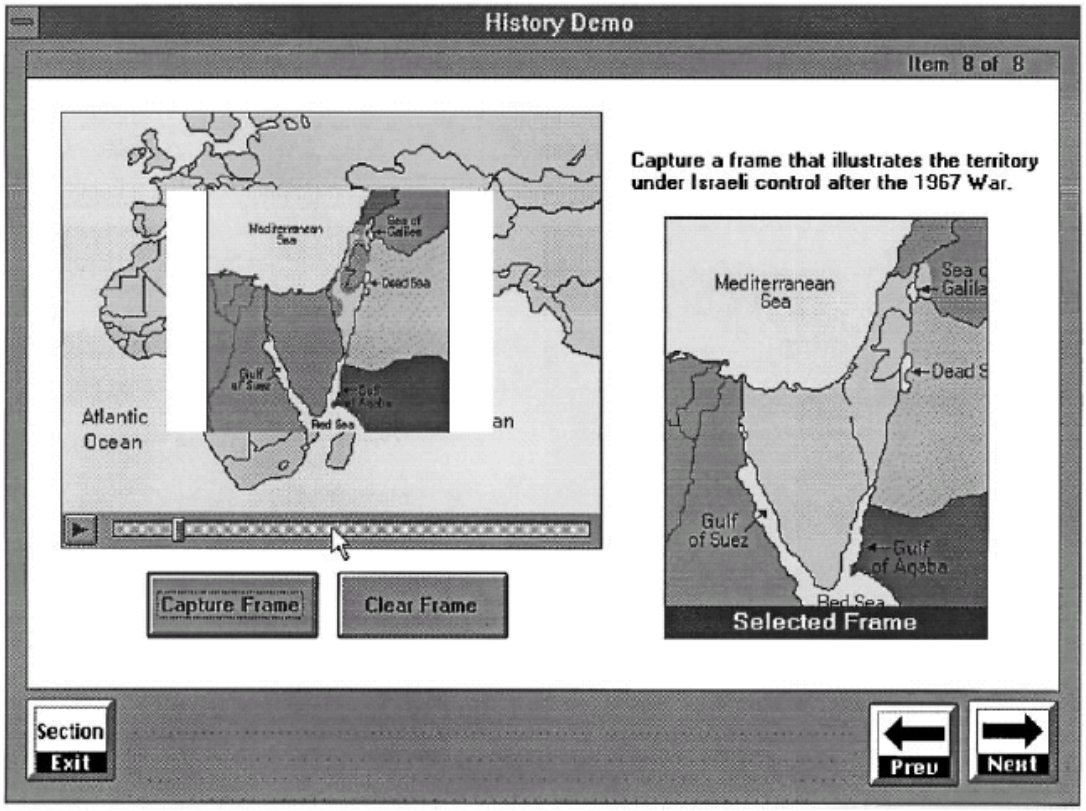


Figure 11.
An example of an ETS capturing frame item (Bennett, Goodman, Hessinger, Kahn, Ligget, Marshall, & Zack, 1999).

Another example of a CF item is shown in Figure 12. Here the test-taker is required to play a video of a child kicking a ball, then select the frame that best exemplifies the child's positioning of their non-kicking foot.

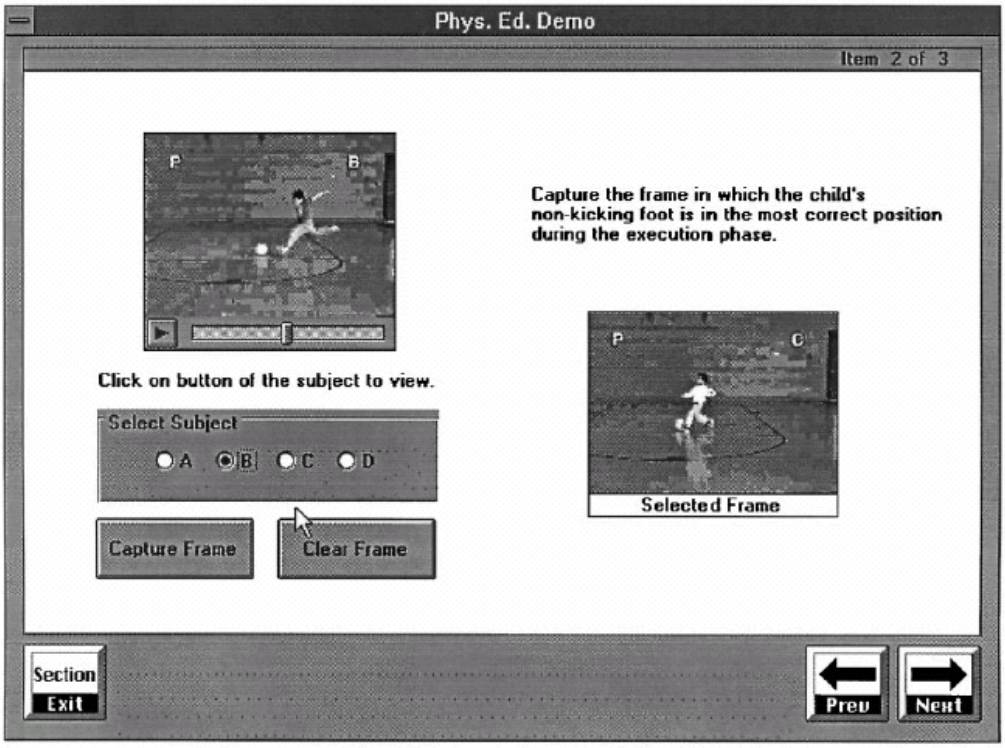


Figure 12.
An example of an ETS capturing frame item (Bennett, et al., 1999).

Multiple Selection Item

A multiple selection (MS) item has been developed as an extension of the multiple-choice format. These items differ in that the test-taker is expected to select as many of the onscreen images or answers that apply. Shotland, Alliger, and Sales (1998) refer to a MS item (see Figure 13), which is part of a battery of tests used to assess applicants for an Insurance Claim Representative position.

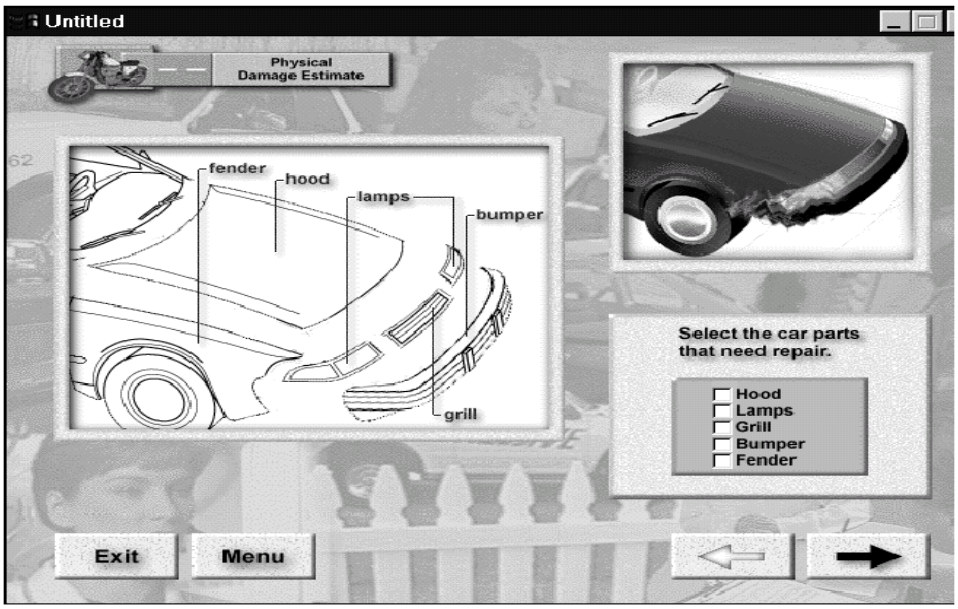


Figure 13.
An example of a multiple selection item (Shotland, Alliger, & Sales, 1998).

This MS item (see Figure 13) consists of three elements: pictorial referencing, response inputting, and branching. The test-taker is required to check the automobile parts that reflect the areas damaged in the top right portion of the display (Shotland et al., 1998). The advantage of the MS format over multiple-choice lies in the fact that needing multiple responses from the test-taker reduces the chance of answering the item correctly by guessing (Zenisky & Sireci, 2002).

Analyzing Situation Item

In these items, test-takers are required to watch or listen to various pieces of information from which they make a diagnosis/decision. An example of this item format is shown in Figure 14.

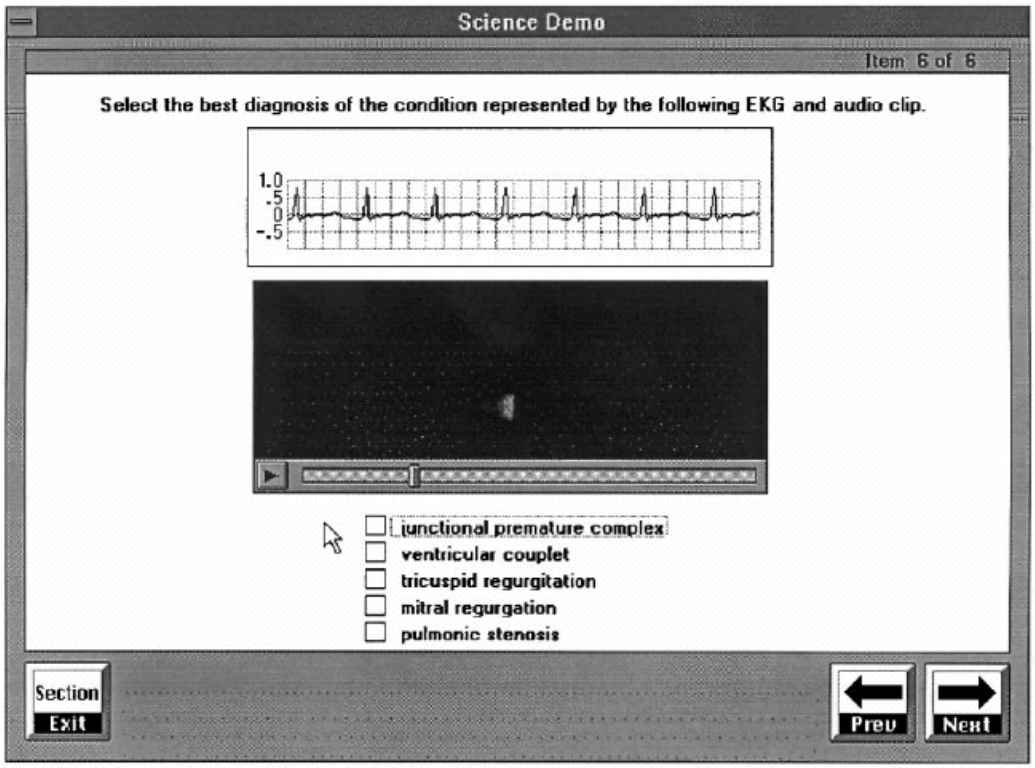


Figure 14.
An example of an ETS analyzing situation item (Bennett, et al., 1999).

Intended for assessment in medical occupation, this item (see Figure 14) stimulus contains three pieces of information: a static electrocardiogram strip, a dynamic trace from a heart monitor (that moves left to right), and a heart beat audio, which is keyed to the monitor display (Bennett et al., 1999). The test-taker must analyze all the medical evidence given to correctly pick the patient's condition from the list provided.

Text-based Response Action

Items have also been designed to accept text-based answers on screen, thus, allowing for the constructed-response format (e.g., short answer and essay paper-and-pencil items) to be transferred to computerized formats. Although items in this category have not produced the variation in design seen in the mouse-based items, the advantage of collecting text-based responses via computer is useful for data

management/scoring purposes (Zenisky & Sireci, 2002). The following represents some of the most commonly used text-based response items.

Essay/Short Answer Item

This item type represents the most directly comparable item to its paper-and-pencil version. As shown in Figure 15, test-takers type their response to the issue presented.

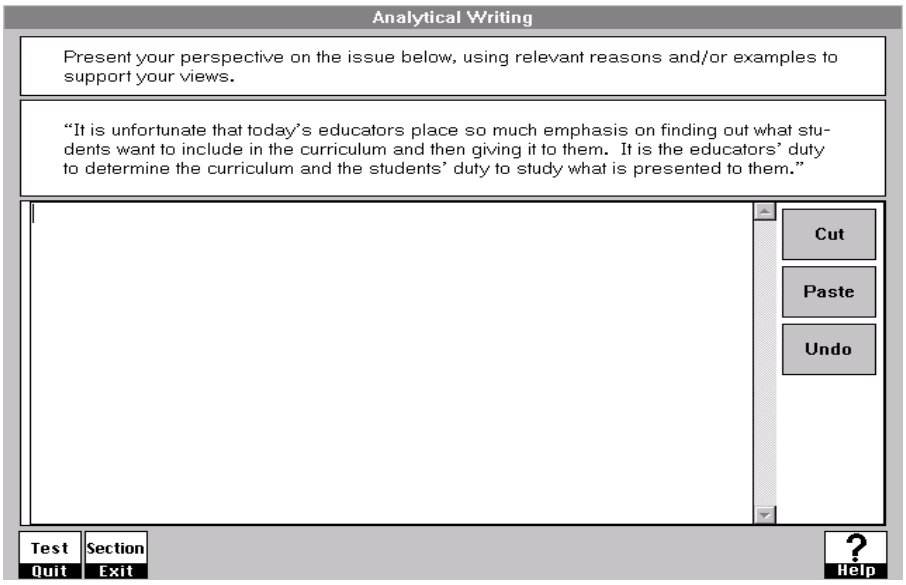


Figure 15.
Example of a GRE essay item format whereby test-takers are required to type a response into a text box that has basic editing functions available (cut, paste, undo) (Microsoft Corporation, 2003).

As will be discussed in the upcoming task constraints section, computer-based free-response items like in Figure 15 require test-takers to display a certain degree of typing proficiency. Further, test-taker familiarity with the basic editing functions often made available on such items is also needed.

Generating Examples Item

Originally created by Norman Frederisken, this task presents a situation, problem, or constraint to the test-taker, whose task it is to pose as many plausible causal reasons for the situation as possible (Bennett & Rock, 1998). The generating examples (GE) item presented in Figure 16 provides an example of an item where arguments and justifications are typed into the item's text box. This item also makes the basic word processing options of "Edit" and "Save" available to the test-taker.

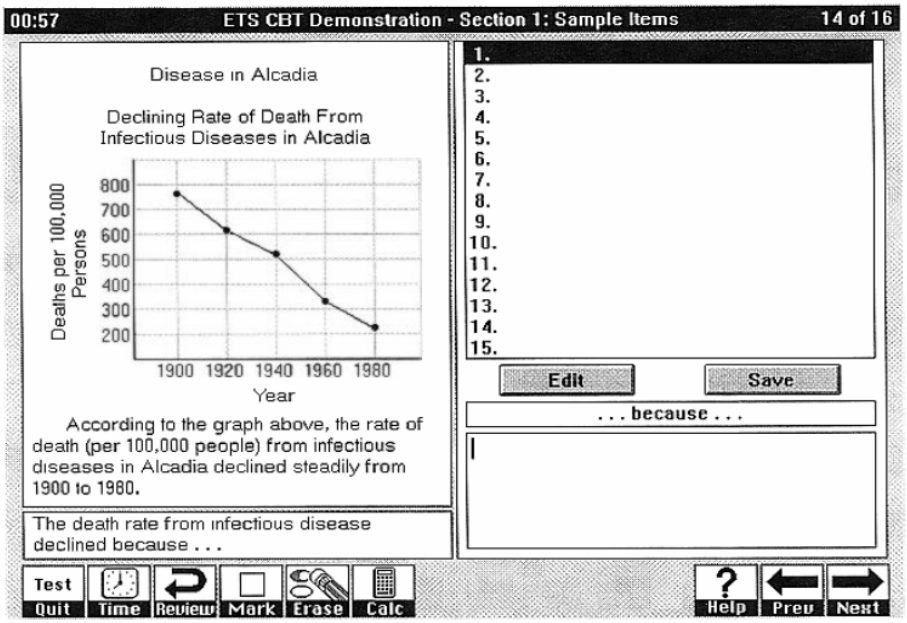


Figure 16.
Example of an ETS generating examples item (Bennett & Rock, 1998).

Bennett and Rock (1998) conducted a study to determine, if prior GE validity results could be generalized to the GRE population, how subgroups performed, and what problems were associated with the operational administration of these items. Validity results showed that GE was found to be reliable but only marginally related to the GRE General Test. With regard to the performance of subgroups, "GE produced smaller gender and ethnic group differences than did the GRE General Test and showed the same relations to outside criteria across groups, suggesting it was

measuring similar skills in each population" (Bennett & Rock, 1998, p. 4). The only concern associated with the administration of these items related to the long and elaborate directions that were given prior to answering the GE items.

Document-analysis Item

A variant on the GE item shown in the previous section is the document-analysis (DA) item. Like the analyzing situation format, test-takers are required to analyze and compare the different pieces of information available. However, for the DA format, responses are constructed and typed into the item's text box. An example of a DA item is shown in Figure 17, which presents four different ways that propaganda was used during WWII.

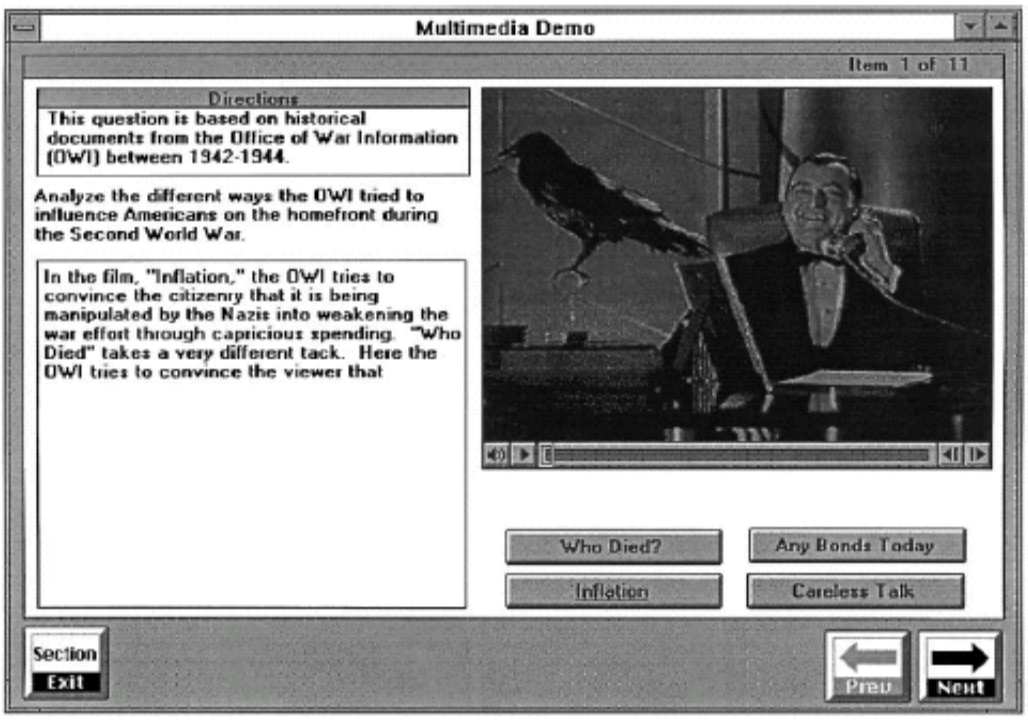


Figure 17.
An example of an ETS document-analysis item that asks the test-taker to analyze and compare the persuasive techniques adopted in wartime propaganda (Bennett et al., 1999).

In this example, the "Who Died" and "Inflation" are sound movies of appropriately 1-3mins in length, and the "Any Bonds Today" and "Careless Talk" are 1-min radio announcements. The text box on the bottom left of the DA item provides an opening paragraph from which test-takers continue the essay.

Text-editing Item

As the name suggests, the text-editing item (see Figure 18) presents test-takers with a passage that contains typically 40 to 50 grammatical and stylistic errors, which are typically enclosed in brackets (Parshall, Davey, & Pashley, 2000).

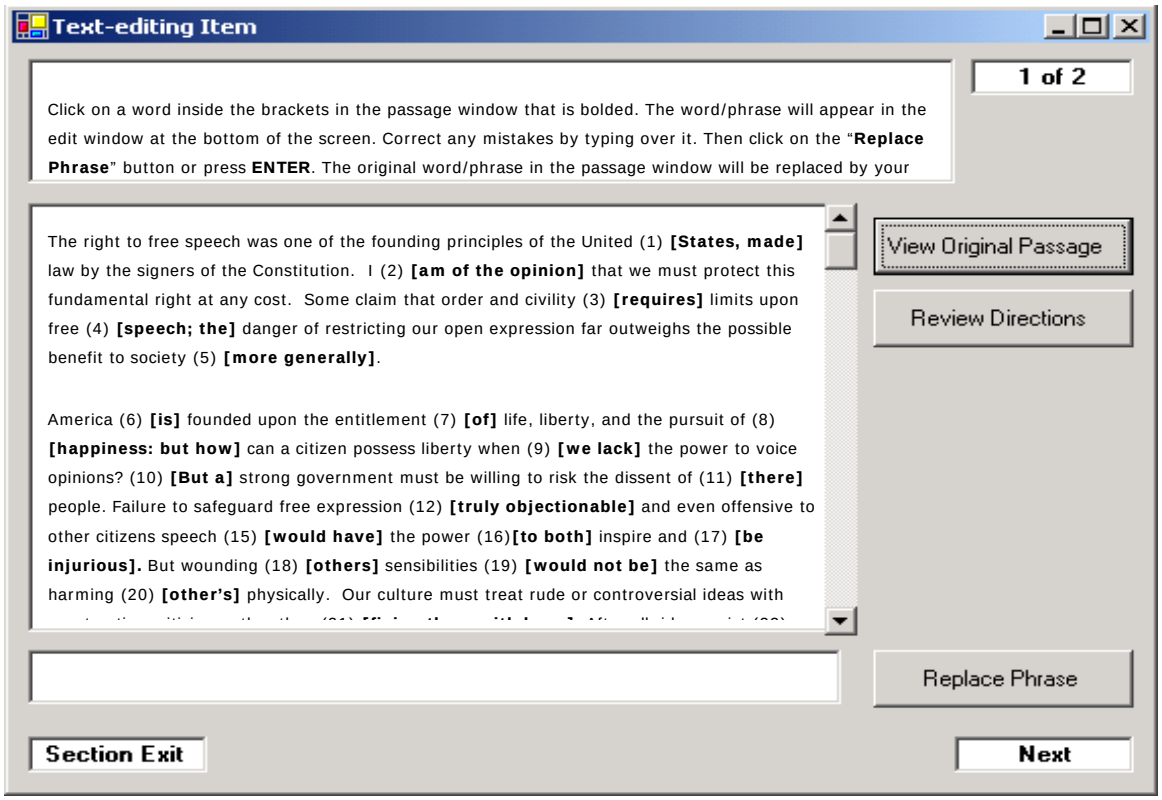


Figure 18.
An example of a text-editing item (Breland, 1999).

After reading the passage, test-takers are required to select the bracketed text that has an error and retype text to correct it, or in the selected response version of this item, select from a list of alternative ways of rewriting the section. Breland (1999) investigated reliability and validity of text-editing (TE) items. Reliability estimates for

the TE items were high (.84) and found to be comparable to other tests of writing such as the Test of Standard Written English (TSWE) (.88) and the SAT II Writing Test (.84). Validity scores correlated well with the self-report inventory as the criterion: test-takers self-assessment of their writing ability (.52), grade-point average (.46), and other writing accomplishments (.30).

Innovative Item Response Formats

As the previous examples showed, a myriad of innovative item design is available which draws on many interactive features that are available in the computerized test environment. Also shown are the various response actions that can be applied to such items, within the current confides of the mouse and keyboard mechanisms. To date, research in this area has focused almost solely on the innovative design and development of item *stimulus* and to a lesser extent on the use of mouse and keyboard mechanisms used to action a response. To date, there has been no research that has focused on the possible innovative item response formats that can be applied to CBTs. Thus, research to date has only utilized the innovation available in the computerized environment for only one aspect of what encompasses an *item*. This lack of development is surprising as the nature of a computerized environment may negate many of the *process* and *comprehension* criticisms leveled at the more elaborate, or administratively involved, response formats applied to PPT tests. Typically process issues relate to *time*, either time involved in one-on-one administration, or time involved in the test-taker responding, for example multiple response steps. Comprehension issues relate to the often lengthy instruction that is needed to inform respondents of the structure of the test and the required understanding of the structure and logic of the response format. An example where

both time and comprehension has proved problematic can be found with the original Self-Perception Profile for Adolescents (SPPA: Harter, 1985). The SPPA is a multi-dimensional self-report instrument that assesses a child's perceived global self-concept and competence across three domains (cognitive, social, and athletic). Although used extensively by social and developmental researchers (Boulton & Smith, 1994; Cairns, McWhirter, Duffy & Barry, 1990; Callaghan & Joseph, 1995; Hoare & Mann, 1994; Hoge & McScheffrey, 1992), and in numerous clinical investigations (Meijer, Sinnema, Bijstra, Mellenbergh, & Wolters, 2000; Schumann, et al., 1999; Veerman, tenBrink, Straathof, & Treffers, 1996), the SPPA's unique two-step response (i.e., test-takers' identify with a group, then provide the intensity of that identification) format has been criticized at three levels. First, lengthy instruction is needed by the test administrator to explain the format, second, the two-step response format doubles the amount of text to be read for each item, and third, the format appears to confuse test-takers resulting in response errors (Marsh, & Holmes, 1990; Trent, Russell, & Cooney, 1994; Wichstrøm, 1995). Although the two-step response format may be problematic in terms of administration time and process, many if not all of the issues raised above can be overcome in a computerized test environment. For example, lengthy instructions can be addressed through effective onscreen design (e.g., using a logical sequence of buttons or screen links that take the test-taker through the steps required) thus, reducing response misunderstanding and need for excessive instruction. Use of transition and special effects (e.g., flashing prompt effect over button or area to signify the action required by the test-taker) aids the understanding of the response required and the progress being made through the computerized test. As these effects result in less instruction text to comprehend and a reduction in response misunderstanding or confusion, a test or scale can take

significantly less time to complete than its paper-and-pencil equivalent. Obviously, a reduction in comprehension load and format confusion will lead to a reduction in the construct-irrelevant variance associated with the observed test score, which is highly beneficial to the application of any measure.

Task Constraints

Regardless of the type of innovative item used, the task constraint attached to an item can impact on performance (Parshall et al., 2002). Task constraints refer to the types of actions that test-takers need to take in order to give a response to an item. In most cases, task constraints arise as a by-product of the test mode environment, therefore, they have the capacity to attach a degree of construct-irrelevant variance to the item (Wise & Kingsbury, 2000). These effects are not restricted to computerized tests. PPTs include a number of construct-irrelevance tasks, such as test-takers needing to keep track of which question they are up to and which answer oval, they need to complete (Wise & Kingsbury, 2000). Although these tasks can become "second-nature" to most test-takers, they are still unintentional, construct-irrelevant aspects of the tests environment that draw test-takers' focus away from the construct of interest (Parshall et al., 2002). The computerized test environment also produces unique task constraints that can impact on the test-takers both positively and negatively. On the positive side, if only one item at a time is being displayed, then there is little, if any concern for the test-taker that they may be responding to the wrong question. Unfortunately, the required use of either a keyboard or mouse for CBT responses might impact negatively on some test-takers given their level of proficiency (Zenisky & Sireci, 2002). Although most computer users are experienced with mouse use, some items require relatively sophisticated manipulations. For example, GM items require

test-takers to sketch out situations, position and move onscreen lines and curve tools using the mouse.

Potentially a greater degree of construct-irrelevant variability might be present when item formats necessitate the use of a keyboard for providing a text-based response. Such responses are typical in construct response formats, where the text answer may range from a short sentence to essay length. For these items, keyboard proficiency can directly impact on the test-taker's ability to produce a full and error free answer. Test-takers who have basic or limited typing expertise may object to the use of the keyboard for providing text-based responses (Parshall et al., 2002). One of the few studies examining possible task constraints resulting from keyboard responses was conducted by Russell (1999), who predicted that keyboard proficiency might have a large effect on students' performance on open-ended items. Test-takers were tested on items from both the Massachusetts Comprehensive Assessment System (MCAS) and the National Assessment of Educational Progress (NAEP). In addition, information on test-takers' prior computer experience and keyboarding speed were collected. Results showed that for students whose keyboarding speeds were at least 0.5 or one-half of a standard deviation above the mean, performing writing items on computer had a moderate positive effect. Conversely, for students whose keyboarding speeds were 0.5 standard deviations below the mean, performance on computer-based writing items had a substantial negative effect, with this effect becoming less pronounced as keyboarding speed increased. Russell (1999) concluded that for the students in this study, who had a keyboarding speed of 20 words per minute or less, performing open-ended questions on computers would substantially underestimate their level of achievement. Further, Russell suggested that students, whose typing speed was greater than 20 words per minute, would be adversely affected if they were

not able to perform the open-ended items on computer. Interestingly, this study also showed that performance on computers for open-ended math items tended to underestimate students' achievement levels, regardless of the level of their typing speed. Russell (1999) stated that this underestimate occurred despite removing graphical items and items that required students to draw a picture.

To address the impacts of task constraints, Parshall et al. (2002) suggested that test developers thoroughly investigate the tasks involved for all the items in the CBT, in order to identify additional unintentional sources of task constraints. Once identified, such constraints should be analyzed to ascertain the potential negative impact on test-takers and their test performance, e.g., either making the task more appropriate or finding ways to prepare the test-takers to deal appropriately with the construct-irrelevant elements.

Discussion

In this review of literature, various participant and technical variables were investigated in order to examine the mode effect that may occur when transferring paper-based tests to computerized versions. As summarized in Table 8, findings confirm that there are many issues to contemplate if equivalence is sought between paper and computer versions of tests, and if optimal computer presentation of items is to occur.

Table 8
Summary of the main human and technological findings

Issues	*Mode Effect*
Participant characteristics	
Race/ethnicity and gender	Only minor performance differences between race/ethnicities, with females showing slightly worse performance on computer-based tests.
Memory and comprehension	Variable results regarding the effects of cognitive processing.
Speededness	Respondents typically take longer to read text from the screen.
Ability	High-ability students' performance appears to be advantaged by CBT.
Computer familiarity	Little difference found in performance of high and low computer users, and this difference can be removed through the use of pretest familiarization tutorials.
Computer anxiety	Inconclusive results in part due to variable operational definitions of "computer anxiety" and methodologies. Predicted that the more familiar users are with computers, the less anxiety they will experience.
User interface – legibility	
Screen resolution	High resolutions result in increased readability and may reduce fatigue.
Screen size	Subjective difference, with participants perceiving that text was easier to read off a larger screen
Font styles and sizes	Affect reading speed and efficiency, with adults performing best on Times New Roman, Arial, and Tahoma, at 12pt, and children preferring Arial and Comic fonts at 14pt size.
Line length	Reading speed optimized for adults when line length falls between 74.8cpl and 100cpl, with children preferring a shorter length of 45cpl.
Number of lines	Four lines of text was read as efficiently as a full screen of text, with reading less efficient when only 1 or 2 lines was presented.
Interline spacing	Typology rules are transferable to screen, where text blocks should be set at approximately 2pts above the size of the font.
Whitespace	Comprehension performance maximized when a one-column block of text is presented.

Issues	*Mode Effect*
User interface – interactive Scrolling	Shown to have a detrimental impact on performance due to a weakening of visual cues and spatial location.
Item review	Although performance has been found to be largely unaffected by whether or not test-takers had the option to review, test-takers prefer to have the option to review available.
Item presentation	Presenting one item per screen tends to increase errors and hurried responses, however, the ability to review items may counter this detrimental effect. Multiple items onscreen may have a facilitating effect allowing test-takers to skip, scan, and build off previous item information.
Task constraints	Both mouse and keyboard response actions have been found to impact on performance when low proficiency levels with these devices are low.

Significant research has been directed towards the characteristics that participants themselves take to the computer testing environment, and the effect that these may have on performance. Unfortunately, cross-cultural and gender comparisons of performance on computerized versions of tests has been largely overlooked. Although Gallagher's et al. study is large and comprehensive, further investigation is required before test administrators can be assured that test delivery is not impacting adversely on the performance of certain sub-groups in society.

Differences in the cognitive processing – in particular speededness, memory, and comprehension, have also been considered between test modes. While early research suggested that test-takers typically took longer to read text from screen, recent studies have found little difference between modes, suggesting that the potential moderating effect of speededness is largely negated by current advanced screen technology. Recall that CRT reading speeds were equivalent to those on paper only under high resolution conditions (1000 x 800), with accommodative accuracy and visual search performance (Harpster et al., 1989), and visual information processing and fatigue

(Ziefle, 1998), improving as a function of screen resolution. In addition, high resolution settings negated the potential advantages gained from using fonts specifically designed for greater onscreen legibility or the readability gained from serifs. Less conclusive have been the investigations of test mode on test-takers' memory and comprehension, with the few studies conducted in this area producing conflicting results. Future research should explore the cognitive processing demands and workload required under each test mode, to establish if in fact there exists a difference in cognitive processing as suggested by the findings of Noyes and Garland (2003), or whether other currently unexplored factors, such as the physiological consequences of each mode, are impacting on performance.

The interaction between participants' ability levels and performance across test modes has also produced contrary findings. While some results found ability not to be a predictor of performance differences (Poggio et al., 2005), other studies found that individuals classified as having high content attainment (Clariana & Wallace, 2002), or high A levels and subject grades (Watson, 2001) were advantaged by a computerized test format. Additional research needs to establish if there is, or the extent of, differential performance across test modes for students with various ability levels.

As computerized and paper-and-pencil versions of tests produce a qualitatively different experience for the test-taker (McDonald, 2002), issues regarding the familiarity of computers and computer anxiety have been the focus of significant research over the last couple of decades. However, empirical evidence regarding the actual impact of these correlated characteristics on CBT performance is largely conflicting. Numerous authors have posited that the wide use of computers throughout

society will diminish the impact of both characteristics, especially when familiarization tutorials or training are administered prior to commencing a CBT.

Parshall et al., (2002) and Booth (1998) have suggested that the user interface encountered by the test-taker is a critical measurement concern for computer-based testing software. Although various interface legibility findings have been discussed in this review (see Table 8 for summary), caution is needed in applying these findings, as many of the typology issues have not been empirically investigated in specific relation to the CBT environment. Instead, research such as the readability of fonts, line length and number, and white space has been investigated from a user interface (UI) perspective, which has typically focused on the application of ergonomic regulation to web sites. This raises questions such as, to what extent are the rules of UI design being applied to CBTs, and further, how transferable are these rules to CBT UI design? Such questions need to be answered in order to evaluate the impact of various UI design options on the reliability and validity of computerized tests.

Related to the presentation of a CBT are the interactive issues of item review and item presentation. Research suggests that performance is largely unaffected by whether or not test-takers had review options available to them, however, qualitative findings suggest that test-takers preferred having the option of item review, believing that they would make use of a review option if it was available. This preference for a reviewing option in CBTs may have the additional benefit of reducing test-takers' anxiety levels and comfort when taking the test. Future research should focus on the impact of reviewing options available in computerized testing, especially in regards to the anxiety levels of low-ability test-takers. In contrast to item review, the effect of item presentation formats on test-takers' performance has received scant attention to date. Of the research found, the most recent study was conducted 14 years ago. In this

study, Dimock and Cormier (1991) suggested that presenting items individually onscreen might be detrimental to performance, as this format caused test-takers to hurry their responses, which resulted in increased errors. Previous to Dimock and Cormier, studies posited that there might be a facilitating effect caused by grouping computer items together, thus allowing test-takers to scan, skip, and build off previous item information (Greaud & Green, 1986; Lee, 1986). However, all of these studies involved the use of out-dated apparatus and technology making conclusions regarding the most optimal item presentation format difficult, if not impossible. Research is needed to establish the presentation format that creates the greatest equivalence across both test modes, and the most optimal presentation format for CBTs specifically.

Poggio et al. (2005) have suggested that a significant confound to the performance of computerized equivalent PPT items, lies in the degree of scrolling that is required to see all of the items stimuli. Though scrolling presents one of the most obvious differences between the test modes, research findings have been largely inconclusive as to the degree that this electronic manipulation affects performance. Some studies have suggested that scrolling detrimentally affects the visual memory of the test-taker, whereas others have argued that scrolling may actually aid in the consolidation of information. Research into the advantages that electronic markers may have on enhancing the comprehension of test-takers and thus, reducing the mode effect of scrolling, is worth further investigation. In addition, future studies should examine the interaction of participants' age and their ability to manipulate and interact with computer-based text with their comprehension of reading material, and performance on text-based passage items.

This literature review also sought to examine innovative items and the response actions required by the test-taker. Zenisky and Sireci's (2002) review of items types

provided the backdrop for the analysis in this Report. This non-exhaustive overview of innovative items alludes to just some of the many creative ways that particular proficiencies can be measured using the computer medium. The selection of items presented in this Report shows that test developers are now not restricted to the static paper-and-pencil approach to item design (Parshall et al., 2002). However, the "bells and whistles of innovative computerized assessments are not without costs…the most omnipresent challenge occurs because these types of assessment are still relatively new; there is not a standard methodology or much available literature that helps practitioners and researchers to develop innovative assessments" (Olson-Buchanan & Drasgow, 1999, p. 3). A question that could be asked is: Are they worth it? The overarching purpose of innovative item types is to improve measurement by either improving the quality of existing measures, or by expanding measures to new areas (Parshall et al., 2002). As Jodoin (2003) found, the multiple-choice items used in his study provided more information per unit time than the innovative DC and CT item types. However, further research may find that this reduced efficiency may be a small cost for increased effectiveness. From the few psychometric evaluations that have been conducted, there have been encouraging findings. For example, Martinez (1993) found that architectural problem-solving FR items were significant predictors of performance, with the majority of test-takers believing this item format had strong face validity. Bennett et al. (2000) found GM items to possess high reliability coefficients, and again high face validity amongst test-takers. Unfortunately, at present there has not been the volume or breadth of research conducted to show that measurement is being fundamentally enhanced by these items. Similarly, there has been a lack of research investigating the innovative approaches to item response formats. It has been posited in this review that many of the process and

comprehension issues that have been attached to previous alternative PPTs response format designs might be overcome through CBT design that makes use of the various effects that can be applied.

Inherent in the design of innovative items are the mouse or keyboard response actions necessary to answer them. As was shown in the presentation of innovative items, many designs require test-takers to be proficient with these input devices, beyond clicking a radio button, or pushing the enter key. By their very nature, innovative items will inherently place task constraints upon the test-taker, whose response-device skill levels will impact positively or negatively on issues such as, test completion time and the amount of errors. While a well-designed preparatory tutorial in basic mouse functions will greatly aid in getting all test-takers to at least a minimal level of competency, keyboard skills, in particular typing speed, is not a skill acquired quickly. Surprisingly, only Russell (1999) has examined the effect of keyboard speed on test-taker performance. Results clearly showed that there was a substantial negative effect for students whose typing speed was below the mean. Parshall et al. (2002) suggested that a task audit is conducted for tests so that all skills required for answering test items, beyond the construct of interest, are identified.

This literature review has focused on some of the key factors associated with the test mode effect, covering both the characteristics of the test-taker and the CBT. Although not exhaustive, a broad range of issues and their associated impacts on performance has been presented. Although some factors related to test mode are still being actively investigated, this review has highlighted many areas for which there is a lack of current research, and where research needs to be addressed specifically towards the design and format of CBTs. Such research is required to ensure that all the factors contributing to the test mode effect can be understood, so attempts can be

made to mitigate its effects. Only when this has been achieved, can test developers have confidence that the test's content is the only stimuli for which the test-taker is responding too, and for which their performance is based.

CHAPTER FIVE

STUDY THREE: WEB-BASED AND PAPER-AND-PENCIL VERSIONS OF THE READER SELF-PERCEPTION SCALE: A COMPARISON OF MEASUREMENT EFFICIENCY AND PARTICIPANTS' PERCEPTIONS

Recent years has seen considerable advancement in the development and delivery of web-based and computer-based standalone testing (Coyne & Bartram, 2006). Of the testing applications that have been web-based, by far the most common have been questionnaire format instruments. Such questionnaires have been designed to measure a multitude of personality constructs such as, self-monitoring (Buchanan & Smith, 1999), self-esteem (Robins, Trzesniewski, Tracy, Gosling, & Potter, 2002), personality development (e.g., Srivastava, John, Gosling, & Potter, 2003), narcissism (Foster, Campbell, & Twenge, 2003), self-focused rumination (Davis, 1999), and motivation (Yost & Homer, 1998), to name a few. This study extends previous research in regards to the specific psychometric analysis undertaken, and the design of innovative items used in this study. First, this study utilized a combination of all of the main multimedia features that are currently available. Specifically, keyed videoed characters were combined in a graphical environment consisting of objects and associated audio. Second, to date, previous interactive innovative items, like the IAV, have been administered via standalone computers, and thus have not utilized the many of the multimedia applications available via the Internet (e.g., Flash). Third, as previous research has shown to date personality measures have dominated web-based testing applications. However, such scales have been typically transferred to screen using equivalent paper-and-pencil form design. The items in this study were the first

application of an interactive scenario-based assessment design, specifically for a non-cognitive ability construct.

Given the innovative design adopted, the primary focus of this study was to compare empirically the measurement efficiency provided by the web-based interactive scenario-based items. This analysis was in order to assess the amount of information derived from a pre-existing paper-and-pencil item that had been transformed into a highly interactive and dynamic state. As Jodoin's (2003) study was concerned with cognitive ability constructs, this investigation is the first to assess the degree of measurement information achieved across administration modes in relation to personality constructs. In addition, this study applied dichotomous scoring across modes to remove any confounds resulting from measurement information deriving from the polytomous nature of the scoring design. Using the 2PL model to generate item parameters, the degree of item information provided at an item level was calculated to establish a gauge of the item measurement efficiency at given theta levels across the continuum. Similarly, at the test level, the degree of test measurement precision for each version of the questionnaire was analyzed across theta levels on the continuum. In order to compare the information functions provided by both web and paper-and-pencil versions of the items, standard error of estimates and relative efficiency statistics were generated.

To date, there have been no previous studies investigating the reactions of school students to either the web-based testing environment, or the use of innovative multimedia item design, in relation to cognitive or non-cognitive ability domains. Thus, this study appraised the test-takers' perceptions to the design of the items administered, and the mode of questionnaire delivery.

Given Jodoin's previous findings, and due to the inherent design properties of the web-based innovative items, e.g., highly visual, engaging, and interactive, it is proposed that the innovative web-based items will provide more information regarding participants' perceptions of their reading ability than the paper-and-pencil item types, and that the test-takers will prefer the innovative items.

The research questions for this study were:

1. Which item type provides the most information (least error) across theta levels at an item and test level?
2. Which version of the questionnaire will be better perceived by the test-takers?

Method

Participants

The sample of this study consisted of 172 intermediate and secondary school students (28 females, 144 males) from two schools in the Auckland area. Students ranged in age from 11-17 years, with a mean age of 12.80 years (*SD* = 1.25) covering school Years 7 to 12.

Design

All participants were administered both the paper-and-pencil and the web-based version of the Reader Self-Perception Scale (RSPS; Henk & Melnick, 1995) Progress items. Two aspects of this design were implemented to prevent memory bias across administration of each version. First, a mixture of 10 additional items from the three other RSPS sub-scales (Observational Comparison, Social Feedback, Physiological States) were incorporated with the eight items from the Progress sub-scale to assist in

disguising the repetition of these items in the web-based version. In addition, there was an interval of three weeks between first and second administrations.

Measures

The measures of focus in this study were the Reader Self-Perception Scale (RSPS) Progress sub-scale in both paper-and-pencil and web-based versions, and a post-assessment feedback measure. Demographic data concerning participants' age and gender were obtained on the opening instruction screen of the web-based questionnaire.

Reader Self-Perception Scale (RSPS) – Paper-and-Pencil

The RSPS is a group-administered self-report instrument for the measurement of how intermediate-level children appraise their reading ability. Based on Bandura's (1977, 1982) theory of perceived self-efficacy and application of such beliefs to specific learning tasks, the RSPS was designed to present items that focus on major elements of reading, such as word analysis and recognition, reading fluency and comprehension. The full RSPS consists of 33 items, where Item 1 is a general item ("I think I am a good reader") and the remaining 32 items represent four sub-scales: Progress, Observational Comparison, Social Feedback and Physiological States. Items were rated on a five-point Likert scale ranging from "strongly disagree" (1) to "strongly agree" (5). For the purposes of this study, only the eight items from the Progress sub-scale were used in *both* paper-and-pencil and web-based questionnaire versions. The Progress sub-scale measures how a participant's perception of their present performance in reading compares with their perceived previous performance (Henk & Melnick, 1995). In addition, a mixture of nine additional items (three from each of the other three sub-scales) was incorporated randomly amongst the eight

progress items (plus the one general item). For this study, the response scale was changed from a Likert system to a dichotomous ("Yes"/"No") response format (see RSPS Progress sub-scale modifications section below for explanation).

Reader Self-Perception Scale (RSPS) – Web-based

Unlike previous transfers of pre-existing paper-and-pencil instruments to screen, the design of the web-based RSPS Progress items was not intended to be constrained by producing a computerized equivalent. Rather, the items designed for the web-based version sought to utilize the rich array of media (e.g., graphics, sound, interactivity etc.) that a web-based administration allows, thus, producing a non-equivalent item format design and a completely novel test-taking experience. The web-based RSPS was based on three distinct design characteristics. First, that the items were embedded within the context of a story. The story used for the web RSPS was based around a reading test that the screen characters were going to be sitting during the course of the day. The storyline was developed through a logical sequence of scenes that progress through pretest, test, and posttest scenes, providing the framework from which the dialogues could be written and Progress items incorporated. Second, the story was written to be as engaging as possible for the test-taker. The development of a virtual school environment was designed to provide a dynamic and vibrant experience for the test-taker. By incorporating some of the interactivity that virtual gaming applications present, the design attempted to create scenes in which the test-taker was fully engrossed in the experience. Last, the storyline and associated scenes were designed to be as credible to the test-taker as possible. Specifically, it was deemed important to provide a situational context for the test-taker that could actually occur in the real world. As the occurrence of a school test is a real and experienced event for all

students, the story's development around the conversations of teachers and students surrounding this event provided the test-takers with a situation that was realistic and believable.

After each item was delivered via onscreen dialogue, the automatic response form appears displaying the item and response options. Attached to each form was a "replay" tab, allowing the test-taker to replay the scene associated with that item, including the lead-in dialogue and item. Once a response had been submitted, the test-taker was automatically taken to the next scene (or next dialogue within the same scene). When back in the virtual environment the test-taker chose the next group of students that they wanted to interact with in that scene. Progression through the scenes was fixed, that is scenes could not be jumped forward or backwards, as scenes did not progress until a response option had been submitted by the test-taker. After finishing all nine items, a final screen appeared thanking them for their participation.

RSPS – Progress sub-scale modifications

Three modifications were made to the Progress items used in this study to facilitate the scenario-based design created for the web version. First, as the web-based items were designed with item stems that were incorporated into the discourse within each scenario, it was decided that the RSPS scales 5-point Likert response format labels, namely, "strongly disagree" (1) to "strongly agree" (5), detracted from the virtual social context that was being created. Instead, the "Yes"/"No" response format was deemed to be better aligned as part of the natural flow of the dialogue. In order to reduce any confounds introduced as a function of response format differences, this dichotomous format was also applied to the paper-and-pencil RSPS items. Second, the scenario questionnaire designed involved a dialogue interaction

with the participant. Scene scripts were created that involved an RSPS item being delivered to the participant in the form of a question. Thus, two forms of each Progress item were used in this study, (a) statement for the paper-and-pencil version (e.g., "I can figure out words better than I could before"), and (b) question for the web-based version (e.g., "Do you find that? Can you figure out words better than you could before?"). Thus, the paper items were as the item statements appear in the original RSPS, whereas items changed into a question (e.g., "I understand what I read better than I could before" to "Do you find that you understand what you read better than you could before?") for the web-based version. In order to maintain a natural sounding dialogue from screen characters, an additional opening question ("What about you?", "Are you the same?", and "Do you find that?") was added prior to six of the web-based items (Items 1 – 6). The last modification again was motivated by the necessity to create a dialogue that was closely aligned with the everyday discourse that participants might experience in a social setting. Therefore, local modes of speech/colloquialisms were incorporated, where necessary, to accentuate the conversational, not formal, aspect of the dialogue. Items were delivered in such a way to accentuate natural discourse and remove as much as possible, the "test feel" of the web questionnaire. For example, the dichotomous response format designed for the web items were written for each item to reflect a response more aligned with the reality of peer conversation (e.g., "Yes, I do", "No, not really"). It is important to note that the web-based Progress items designed for this study were not intended to be the equivalent of their paper counterparts. Instead, the study sought to estimate how much information could be gained by producing a dynamic and innovative approach to construct measurement, rather than assessing the congruity or comparability of these items via web administration.

Post-assessment feedback measure

Upon completion of the web-based questionnaire, participants were assessed on their reactions to both modes of questionnaire administration. A self-constructed post-assessment measure, presented as a paper-and-pencil survey, consisted of 10 items that asked participants' familiarity with computers and their preference between the administration formats. The first item ("If you had a choice of doing a questionnaire either on paper or via the web, which would you choose?") consisted of two parts. First, participants were asked to choose either: (1) "On paper", or (2) "On the web". Second, participants were asked to write down the reasons for their choice. The second item ("How familiar are you with computers?") required participants to indicate their level of familiarity ("Very familiar – I use computers all the time", "Quite familiar – I use computers occasionally", "Not very familiar – I almost never use computers", "Not familiar at all – I never use computers"). Items 3 ("What did you like most about the web questionnaire?") and 4 ("What did you dislike the most about the web questionnaire?") focused on the participants attitudes towards the web-based items specifically. The fifth item ("Which reading questionnaire best reflects what you think about your reading confidence?") consisted of two parts. First participants had to select one of the following two options: (1) Paper questionnaire, and (2) Web questionnaire. Second, based on their first response, participants were asked to write down the reasons supporting their choice. The next five items (6 – 10) asked participants to indicate their level of agreement ("strongly agree", "agree", "disagree", "strongly disagree") with the following propositions: "The onscreen instructions were easy to follow" (Item 6); "I understood what I was expected to do throughout the web questionnaire" (Item 7); "I liked the use of video characters (for example, students and teachers" (Item 8); "I liked the use of graphics (for example,

desks, chairs, plants)" (Item 9); and "I would prefer to do other tests (for example, a science test) in the style of the web questionnaire" (Item 10).

Procedure

Prior to paper-and-pencil administration, teachers were thoroughly briefed regarding the administration of both versions, and the post-assessment feedback measure. Teachers were instructed that the students who volunteered to participate in the research were blind as to the purpose of the study. Thus, teachers were instructed to tell participants that the focus of this research was assessing what students thought about their reading capabilities, and that they would complete two questionnaires that asked them about how they felt about different aspects of their reading ability, and a feedback form.

All participants took both the paper and web versions of the questionnaire. Measures were administered during two separate 20-minute sessions that took place in the classes and computer labs of the schools. The paper-and-pencil version was administered first, in the classroom, and the web-based session was administered after the three week interval in the computer labs. Due to school logistic constraints a randomized counter-balanced (e.g., web-based first and paper-and-pencil second and paper-and-pencil first and web-based second) design could not be adopted. The feedback question was completed after the web-based version had been completed in the second administration. Participants were assured that their results would be kept confidential.

Paper-and-pencil administration

Teachers administered the one page RSPS questionnaire. In all cases, students in the participating classes volunteered to participate in the research. Each paper-and-pencil RSPS had a unique code printed on the questionnaire for matching a participant's paper-and-pencil responses to their web-based responses. Teachers were asked to retain a record of each participant's unique code for use in the introduction screen of the web-based questionnaire.

Web administration

Students were instructed to go to a dedicated web-page to start the web-based questionnaire. A generic password was given to all participants to enter in the login field box. While the application was loading, the instruction screen informed participants to the web questionnaire that they were about to experience, and gave details as to how they would navigate around the screen, interact and respond to the screen characters, and enter their item responses. In addition, participants filled out demographic information about their age and gender. Also participants were required to enter the unique code (from their paper-and-pencil questionnaire), which was supplied by the teacher so that paper-and-pencil forms would not have to be redistributed to participants and potentially confound web-based responses to items.

Throughout the questionnaire, participants were free to replay, and change their response to the current dialogue that they were experiencing. However, once participants had submitted their responses (by clicking on the "submit" button), items could not be replayed, nor could the responses to items be changed.

Results

Data Analysis

Analysis was conducted to assess questionnaire characteristics such as the means, standard deviations and cross-mode correlations across administration modes. Confirmatory factor analysis (CFA) was examined to establish the degree of equivalence between item modes. Item and test information across theta levels were established to examine the psychometric comparability of the scores derived from the paper-and-pencil and web-based versions, and to investigate the questionnaire mode that participants preferred. Responses from both modes were estimated on the IRT theta metric, and modeled using the 2PL model. This IRT model was chosen based on numerous studies that have found it to be the most appropriate model for modeling dichotomous responses from personality constructs (Reise, 1999; Reise & Waller, 1990; Zickar, 2001). IRT person and item parameters were estimated using BILOG 3.1 (Mislevy & Bock, 1990), with program defaults.

Descriptive Statistics

Table 9 shows the means, standard deviations, and the cross-mode correlations for the nine items and total scores for each mode of administration. The correlations between the web-based innovative items and paper-and-pencil although positive, with the exception of Item 8 (r = -0.02), and statistically significant with the exception of Items 7 and 8, were insufficient to suggest cross-mode equivalence between measures. It is worth noting that Items 7 and 8 were two of the three items that did not have a prior question before the question item. This finding might suggest that the vibrancy of the virtual environment requires a prior statement or question to assist in orienting the test-taker to the information that they are about to be asked.

Similarly, the correlation between the web questionnaire and paper-and-pencil totals ($r = 0.50$, $p < .001$) suggests only a low positive correlation across modes. This correlation is not sufficient to suggest that the web-based version of this questionnaire will replicate responses (or rank ordering of participants) from the paper-and-pencil version (Potosky & Bobko, 2004). In addition, paired means *t*-tests were compared to ascertain if there were any mean differences between scores across administration modes. At the test level, *t*-test results indicated that mean scores on the web-based questionnaire ($M = 8.21$, $SD = 1.19$) were not statistically significantly different from the mean scores on the paper-and-pencil version of the questionnaire ($M = 8.04$, $SD = 1.20$). Statistically significant differences in mean scores were found in four of the cross-mode paired items: Item 2 ($t = -2.95$, $p < .01$, $df = 170$), Item 3 ($t = 10.24$, $p < .001$, $df = 170$), Item 4 ($t = -2.89$, $p < .01$, $df = 170$), and Item 7 ($t = -2.80$, $p < .01$, $df = 170$). The mean scores for remaining items were not significantly different across administration modes. The failure to find a significant relationship between paper-and-pencil and web-based responses may be attributed to the low variability of scores on the paper version of these items.

Thus, equivalence between web-based and paper-and-pencil versions of the RSPS Progress sub-scale was achieved by the lack of significant difference reflected in the means and standard deviations across versions. However, equivalence was not reflected by the cross-mode correlations, where high correlations ($r = > 0.70$) of paired items across modes would have reflected a strong positive relationship between the items across modes.

Table 9

Descriptive statistics and cross-mode correlations for the RSPS – Progress items

	Web-based		Paper-and-Pencil											
	M	*SD*	*M*	*SD*	1	2	3	4	5	6	7	8	9	Total
1. What about you, do you think you're a good reader?	0.80	0.40	0.83	0.38	0.64**	0.08	.332	0.22**	0.01	0.01	0.09	-0.05	0.15	
2. What about you, do you think you're getting better at reading?	0.93	0.26	0.99	0.11	0.06	0.19*	.205	0.07	0.03	0.11	0.04	-0.30	0.14	
3. What about you, do you find that you can read faster now than you could before?	0.94	0.24	0.54	0.50	0.02	-0.02	0.17*	0.09	0.04	-0.01	0.06	0.21*	0.05	
4. Do you find that? Can you figure out words better than you could before?	0.89	0.31	0.96	0.19	0.15	0.14	0.03	0.25**	0.15	0.01	-0.07	-0.04	0.18*	
5. Are you the same? Do you find that when you read you can recognize more words than you used to?	0.94	0.25	0.94	0.24	0.01	-0.03	-0.04	.080	0.54**	0.13	0.05	-0.03	0.26**	
6. What about you…do you find reading easier than it used to be?	0.95	0.22	0.88	0.32	-0.04	0.02	0.04	0.09	0.02	0.16*	-0.05	-0.04	0.18*	
7. Do you find that when you read stuff now that you don't have to try as hard as you use to?	0.87	0.34	0.95	0.21	0.02	0.12	0.12	0.02	0.05	0.02	0.08	-0.02	0.07	
8. Do you find that you read better now than you could before?	0.96	0.19	0.99	0.11	-0.01	-0.02	0.02	0.14	0.36**	0.09	0.23	-0.02	0.39**	
9. Do you find that you understand what you read better than you could before?	0.92	0.27	0.95	0.22	-0.01	-0.03	0.04	0.03	0.21**	0.10	-0.02	-0.04	0.23**	
Total	8.21	1.19	8.04	1.20										0.50***

*Note.*****p*< .001 ***p*< .01 **p*< .05

Confirmatory Factor Analysis

Two CFA models, representing each administration were examined. Both reflected a single-factor model structure, based on the notion that that these items represent the items associated with the Progress sub-scale of the RSPS. Thus, the Progress sub-scale was treated as a unidimensional construct. Based on Hoyle and Panter's (1995) recommendation, both absolute and incremental goodness-of-fit indexes for comparing models and analyzing invariance were examined. The absolute fit index was represented by the chi-square statistics. Both the comparative fit index (CFI; Bentler, 1992) the Tucker-Lewis index (TLI; Tucker & Lewis, 1973) and the root-mean-square error of approximation (RMSEA; Steiger & Lind, 1980) were examined for incremental goodness-of-fit indexes. Table 10 provides a summary of the model fit measures observed and item factor loadings for both models for each questionnaire version. None of the nine items showed a consistency in factor loadings across models. Meehl (1990) posits that a factor loading should be at 0.20, and that anything lower may be due to "complex unknown network of genetic and environmental factors" (p. 209). Given that factor loadings are an indication of the amount of variance being captured, it is ideal to have factor loadings of >.60, thus, accounting for at least 50% of the variance underlying the latent variable (Chin, 1998). While both versions show only one item, Item 8 (paper-and-pencil) and Item 3 (web-based), that is <.20, generally factor loadings are low, with only one item in each questionnaire showing loadings of >.60 (Item 9: paper-and-pencil; Item 6: web-based). The biggest difference between factor loadings was found for Item 3 ("What about you, do you find that you can read faster now than you could before?"), where this item accounted for only 5% explanation of variance on the paper version, it accounted for less than half a percent on the web-based version. A post hoc investigation of the response data associated with this item revealed a converse

behavior amongst test-takers where 66.7% responded “Yes” to the web version of Item 3, and 60.9% responded “No” in the paper version. Further theoretical analysis is required to establish why this item provoked such diametric responses across modes.

Results indicated that while two web-based questionnaire fit indexes showed mediocre fit to the data for (CFI = .814; TLI - .752; RMSEA = .077), the paper-and-pencil model fit (CFI = .420; TLI = .254; RMSEA = .191) was poor, indicating misspecification of factor loadings for this version. The better model fit for the web-based questionnaire is in accordance with the findings of other studies (Ployhart et al., 2003; Salgado & Moscoso, 2003; Meade et al., 2007) where web-based scales provided better descriptive statistics than the paper-and-pencil versions. These findings show that the two questionnaire versions did not replicate the same factor structure, thus measurement equivalence was not demonstrated between the administration modes. These results suggest that the lack of significant difference found in the means and standard deviations across questionnaire versions (see Table 9) are not meaningful, confirming the lack of equivalence reflected by the low cross-mode correlations.

Table 10
Item factor loadings, absolute and incremental fit indices for web-based and paper-and-pencil questionnaires

Questionnaire	Factor Loadings	df	χ^2	*CFI*	*TLI*	*RMSEA*
Paper-and-pencil		28	201.37	.420	.254	.191
Item 1	.51					
Item 2	.45					
Item 3	.22					
Item 4	.27					
Item 5	.43					
Item 6	.28					
Item 7	.34					
Item 8	.04					
Item 9	.89					
Web-based		28	54.70	.814	.752	.077
Item 1	.33					
Item 2	.38					
Item 3	.02					
Item 4	.44					
Item 5	.31					
Item 6	.61					
Item 7	.38					
Item 8	.49					
Item 9	.53					

Item Parameter Statistics

Table 11 presents the descriptive statistics for *a* and *b* parameters estimated under the 2PL model. Means and standard deviations for both versions show almost identical discrimination estimations, indicating that under both modes, items provided similar levels of discrimination power. Comparable similarities were found amongst *b* parameters, however, paper-and-pencil items proved on average to be slightly easier when delivered on paper-and-pencil, indicating that participants had increased positive

self-perceptions of their reading progress via this mode. This finding is also reflected in the minimum value for the paper administered questionnaire, however, median and maximum statistics showed that participants were more inclined to be more positive regarding their reading progress in their web-based responses.

Table 11
Descriptive statistics and difficulty and discrimination item parameters for paper-and-pencil and web-based versions

	Paper-and-pencil		Web-based	
	a	*b*	*a*	*b*
Item 1	1.29	-1.80	0.66	-2.38
Item 2	0.97	-4.86	1.16	-2.70
Item 3	0.90	-0.39	0.67	-4.40
Item 4	1.10	-3.49	1.83	-1.67
Item 5	1.49	-2.31	1.10	-2.89
Item 6	1.24	-2.08	1.92	-2.17
Item 7	1.54	-2.67	1.18	-2.09
Item 8	1.26	-3.98	1.42	-2.96
Item 9	2.04	-2.34	2.04	-1.87
Mean	1.32	-2.66	1.33	-2.58
SD	0.35	1.31	0.52	0.82
Minimum	0.90	-4.86	0.66	-4.40
Median	1.26	-2.34	1.17	-2.39
Maximum	2.04	-0.39	2.05	-1.67

Efficiency Analysis

In item response theory, item information functions (IIF) and test information functions (TIF) offer a potent method of describing and comparing the measurement efficiency offered by items and tests. The term *information* in this sense refers to the amount of measurement precision being achieved for the trait or ability of interest. In other words, the degree of measurement efficiency being provided by an item or test. As all derived ability values are estimates of a latent true value, the amount of information

that an item provides about a trait or ability level, dictates the degree of precision under which ability can be estimated (Barker, 2001). The height (or steepness) of information functions relate to the amount of discrimination being provided by an item, with the highest part of the function indicating where, on the trait continuum, the item or test is discriminating most effectively (Donoghue, 1994). Low or relatively flat functions indicate both a low degree of information, and poor discriminating power amongst test-takers' theta levels. Under the 2PL model, the IIF is most symmetrical around the item's associated difficulty parameter (Baker, 2001).

Although, understanding the amount of information being provided by an item is useful, the contribution of a single item on a point on the theta continuum is usually small. However, when IIFs are summed, the resulting test information function (TIF) provides an IRT version of test reliability (Donoghue, 1994). As the amount of measurement error associated with an ability estimate is conversely related to the amount of information provided by an item, the higher the information, the lower the standard error of estimate attached to the trait estimate.

Item Level

Figure 19 presents a plot showing the amount of information produced by each item, under both administration modes. Overall, seven items clearly produced greater measurement precision under different modes. Three items: Items 1 ($I(\theta) = 0.34$), 3 ($I(\theta) = 0.24$), and 5 ($I(\theta) = 0.36$), clearly yielded more precision when administered in the paper-and-pencil mode. Where, four items: Items 2 ($I(\theta) = 0.22$), 4 ($I(\theta) = 0.51$), 6 ($I(\theta) = 0.51$), and 8 ($I(\theta) = 0.22$) showed superior measurement precision in their web format. The remaining two items (Items 7 and 9) produced similarly efficient levels in both questionnaire modes. Interestingly, two items (Items 2 and 8) in the paper-and-pencil mode produced minimal theta information ($I(\theta) = 0.01$ and $I(\theta) = 0.04$

respectively), whereas only one item (Item 3: $I(\theta) = 0.03$) was inefficient in the web format. However, a paired *t* test revealed no significant difference in item information values between paper-and-pencil ($M = 0.25$, $SD = 0.17$) and web-based ($M = 0.29$, $SD = 0.18$) modes. Both versions of Item 9 showed the largest amount of measurement information compared to the other Progress items.

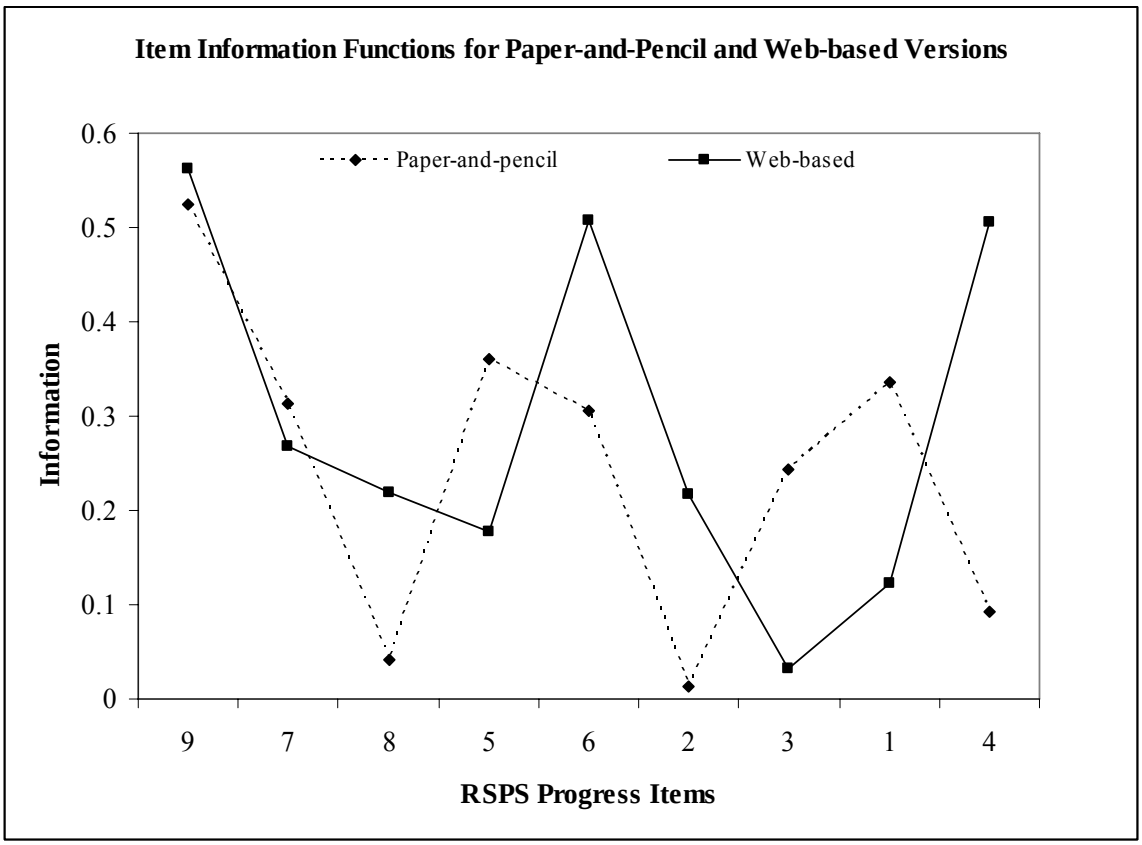


Figure 19.
Item information for the nine RSPS Progress items in both paper-and-pencil and web-based versions.

Test Level

As item information functions are independent of the other items in a test, the individual contribution made by each item towards overall test information can be determined. This aspect of IRT-based reliability is superior to CTT approach, where the discrimination indices and reliability estimates are dependent on the other items in the test. As a result, test information has the advantage of being specific to each theta level

on the continuum, rather than the CTT approach where measurement efficiency is average across all theta levels (Samejima, 1994).

Figure 20 shows the accumulated information being captured by the nine items, across both questionnaire versions. As maximum likelihood estimates were used to attain theta estimates and given that normal approximation is satisfied in most cases with tests as short as nine items (Samejima, 1977), the information from each test can be translated into a standard error of estimation. Within the framework of IRT, this estimate can be obtained to give an indication as to the range with which approximately 68% of the estimates of reading progress self-efficacy fall at each test's point of maximum test information. For the web-based questionnaire, the test information function was symmetric about the theta level of -2.0 (*TIF* = 1.12, *SE* = 0.94). The standard error of 0.94 meant that approximately 68% of the estimates reading self-efficacy fell between -2.94 and -1.06. Test information was captured most effectively around the theta level of -2.4 (*TIF* = 0.94, *SE* = 1.03) for the paper-and-pencil version. Based on the standard error of 1.03, approximately 68% of the estimates reading self-efficacy fall between -3.43 and -1.37. This information showed that the paper-and-pencil mode worked most efficiently amongst participants with higher reading self-efficacy than the web-based measure, whereas, the web-based version represented theta estimates closer to average on the trait continuum (e.g., $\theta = 0$). There was only a minimal difference in the theta range between the two tests (Web_{diff} = -1.89; $Paper\text{-}and\text{-}Pencil_{diff}$ = -2.06), indicating that within their ranges both tests were covering approximately the same range of theta values.

Another useful method for comparing the measurement efficiency of two tests involves comparing statistically, the relative efficiency of each TIF (Lord, 1977). Focusing on the length on the test, this statistic provides an indication of how many

additional items one test may need in order to provide the same degree of precision as another test measuring the same ability (Hambleton et. al, 1991). Based on the web-based TIF, this test was functioning as if it were 19% longer than the paper-and-pencil version. Thus, the paper-and-pencil version would need to be lengthened by 19% to yield the same measurement precision as the web-based version.

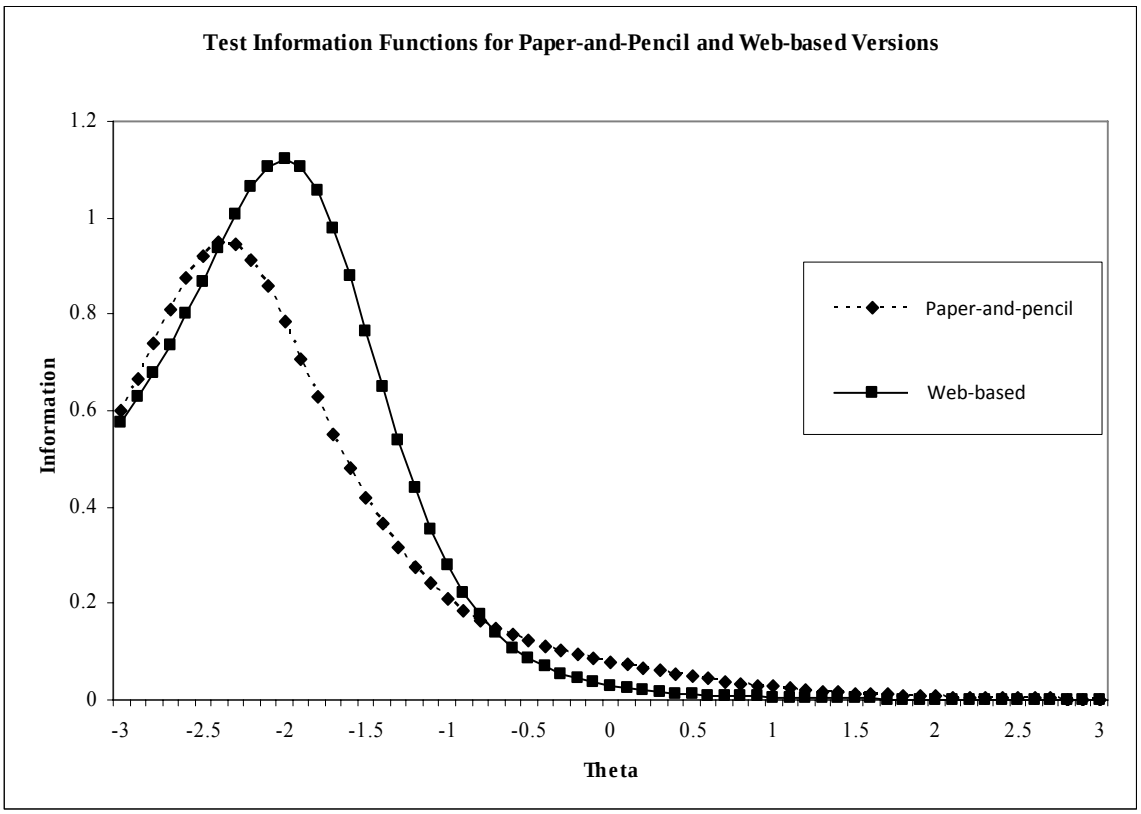


Figure 20.
Average IRT information across theta levels for paper-and-pencil and web-based versions of the RSPS Progress sub-scale.

Mode Preferences

At the end of the second administration, participants were asked to complete a post-assessment feedback instrument regarding which administration mode was preferred, their attitudes to other issues relating to both versions, their level of familiarity with using a computer, and specific questions relating to the web-based version. Of the 133 participants that completed the feedback form, 99% classed themselves as being either "very familiar – I use computers all the time" (48%) or "quite familiar – I use computers occasionally" (51%). Not surprisingly, the chi-squares conducted to test the relationship between computer familiarity with mode preference (Item 1) and mode best reflecting reading confidence (Item 5), were not significant (Item 1: $\chi^2 = 3.700$, $df = 2$, $p = 0.157$); Item 5: $\chi^2 = 2.482$, $df = 2$, $p = 0.289$). Further, computer familiarity did not impact on either web-based or paper-and-pencil questionnaire scores (web: $\chi^2 = 7.618$, df 10, $p = 0.666$; paper-and-pencil: $\chi^2 = 14.587$, $df = 10$, $p = 0.148$).

Participants' responses to the open-ended questions were content analyzed, categorized, and tabulated, with results presented in Tables 12 to 15. As can be seen in Table 12, preference for the web-based version was given by 72% of all participants. The most important reasons for this choice were (in order of frequency): easier on the web (and you make fewer mistakes, easier to follow) (19%); more interactive (18%); more enjoyable (or fun) (17%); more interesting (9%). Conversely, of the minority of participants (28%) that preferred the paper-and-pencil version, most reasons were directed around the ability to review/skip or change answers in the web-based version, e.g., quick to answer (24%), can review backwards and forwards (22%), and can change a previous answer (12%). Only a small proportion of students cited that they found it was easier to read off paper (22%). While the typological differences between

information presented via paper and screen is obvious (e.g., whitespace, line length, font size), generally studies have found that reading speed and comprehension between the two modes are not statistically different, indicating that individuals are comfortable, efficient and effective, reading in both modalities.

Table 12
Frequencies and Percentages for Reasons for Preferred Mode Choice

	Administration Mode					
	Web		Paper-and-pencil		Total	
Item 1	*n*	%	*n*	%	*n*	%
If you had a choice of doing the questionnaire either on paper or via the web, which would you choose?	96	72	37	28	133	100
Why?						
Easier on the web	24	**19**	0	0	24	14
More enjoyable (or fun)	22	17	0	0	22	13
Quick to answer	5	4	11	**24**	16	10
More interactive	23	18	0	0	23	14
More interesting	11	9	0	0	11	7
Can relate to (or see) the people on the screen	7	5	0	0	7	4
Don't like writing (or don't have to write)	6	5	0	0	6	4
Easier to read off paper	0	0	10	22	6	4
Can review backwards and forwards	0	0	10	22	6	4
Better at typing (or clicking) than writing	5	4	0	0	5	3
Can change a previous answer	0	0	5	12	5	3
Voices are easier to understand than words	4	3	0	0	4	2
Better having questions read and spoken to you	4	3	0	0	4	2
Pictures (or graphics) better (or easier) than words	3	2	0	0	3	2
Environment more real than sitting a test paper	3	2	0	0	3	2
Didn't feel like a test	3	2	0	0	3	2
Prefer to read off paper	0	0	3	7	3	2
Understand it better	2	2	0	0	2	1
Prefer using a computer to answer questions	2	2	0	0	2	1
Can't cheat on the web	2	2	0	0	2	1
More exciting	1	1	0	0	1	1
More entertaining	1	1	0	0	1	1
Not good on computers	0	0	2	5	2	1
Used to doing tests on paper	0	0	1	2	1	1
Prefer writing by hand	0	0	1	2	1	1
No distractions	0	0	1	2	1	1
Need to do it again if the computer froze	0	0	1	2	1	1
Don't have to sign in	0	0	1	2	1	1
Total	128	100	42	100	170	100

Note. Bold values represent the item with the highest response percentage for each administration mode. Underlined values represent items that were given as reasons under both modes of administration.

The reason that participants gave when asked which mode best reflected their understanding of their reading progress is shown in Table 13. The most frequent reason given by those who believed that the paper-and-pencil version was best reflecting their progress in reading, was that there were more questions relating to their reading capabilities (47%). This reason is an artifact of the study design where, to minimize memory effects/recognition of items across modes, additional RSPS items were incorporated with Progress items. Obviously, for some participants this gave the impression that more information about their perceived progress in reading was being established. Of the 55% of participants that perceived that the web-based items best reflected their reading progress, 38% believed that the innovative items format gave participants more of an understanding of what was being asked. Specifically, participants felt their comprehension increased as a result of having the questions verbalized to them, rather than simply reading the question /statement. Web-preferred participants also cited that the realistic environment was a factor which assisted in reflecting their reading self-perceptions, adding that, the questions meant more (or were more meaningful) via this mode of administration (15%).

Table 13
Frequencies and Percentages for the Mode that Best Reflected their Self-Perceptions of Reading Progress

	Administration Mode					
	Web		Paper-and-pencil		Total	
Item 5	*n*	%	*n*	%	*n*	%
Which reading questionnaire best reflects what you think about your reading confidence?	72	55	59	45	131	100
Why?						
I understood more of what they were asking me when the students (people) asked me	26	**38**	0	0	26	**31**
Because it was more real, the questions meant more to me (or were more meaningful)	10	15	0	0	10	12
Because I can read and hear the questions	9	13	0	0	9	11
It wasn't boring (or I didn't get bored) so I kept interested in what they were asking me	9	13	0	0	9	11
There were more questions about my reading progress	0	0	8	**47**	8	9
I just had to read the questions	0	0	6	35	6	7
Because I could see the people (their face, expressions)	5	7	0	0	5	6
I thought more about what the questions were asking me (or my reading skills)	4	6	0	0	4	5
I had to think about what I was doing	3	4	0	0	3	4
I like doing interactive things which made me think more about my answers	2	3	0	0	2	2
It was easier to understand what was being asked	0	0	1	6	1	1
Did not like the web questionnaire	0	0	1	6	1	1
I could think about and see my previous answers	0	0	1	6	1	1
Total	68	100	17	100	85	100

Note. Bold values represent the item with the highest response percentage for each administration mode. Underlined values represent items were given as reasons under both modes of administration.

When asked a more specific question relating to the strengths of the web-based innovative items, participants most frequently commented that they liked the interactivity of the scene characters, in particular being an active part of the character's dialogue, and being asked about their reading progress directly by scene characters (e.g.,

students asked you the questions (22%); students talked and interacted with you (18%) (see Table 14).

Participants stated that having the questions verbalized to them made the questions easier to understand (12%). In addition to the interactivity with scene characters, participants liked the ability to interact with the environment, graphics, and scene characters (9%), and liked the visual story-like nature of the questionnaire (9%). Table 14 also presents comments regarding the features most disliked about the innovative items. Most frequent responses were (in order of frequency): nothing (37%); could not review previous questions (26%); took more time (21%).

Table 14
Frequencies and Percentages for Likes and Dislikes of the Web-based Questionnaire

Items	Web *n*	Web %
3. What did you *like* most about the web questionnaire?		
Students asked you the questions	24	22
Students talked and interacted with you	19	18
Easier to understand the questions when they are spoken to you	13	12
Interact with graphics and scene characters	10	9
Like a visual story (movie)	10	9
Fun	9	8
It was like a real (life) school with students	8	7
Graphics	6	6
Understood what you have to do	3	3
Didn't have to write	2	2
Based in a school and classroom	1	1
Students like me	1	1
Questions were spoken	1	1
Total	107	100
4. What did you *dislike* most about the web questionnaire?		
Nothing	7	37
Could not review previous questions	5	26
Took more time	4	21
Could not skip forward to other scenes	1	5
Scene characters should have talked more in each scene	1	5
Total	19	100

The last set of five statements on the post-assessment feedback form assessed participants' opinions of different aspects of the innovative items presented on the web questionnaire. Table 15 shows overwhelmingly that participants either "strongly agree" or "agree" that the onscreen instructions were easy to follow (Item 6: 97%), and that it was clear what they were required to do to progress through the web questionnaire (Item 7: 97%). Regarding the two main design features of the innovative items, namely the use of video characters (Item 8) and scene graphics (Item 9), participants indicated positively that they liked the use of these features, with both items receiving high agreement (Item 8: 86% and Item 9: 88%). A more even distribution of responses was seen from participants' reactions to whether they would like to sit other tests in the style of the web questionnaire (Item 10). Although, this item showed the highest amount of participants strongly disagree or disagree with this statement (n = 14, n = 33 respectively), the majority of participants (64%) displayed a positive attitude towards the possibility of sitting other tests, potentially measuring cognitive-ability constructs, in the design of the web questionnaire.

Table 15
Frequencies and Percentages for Participants Attitudes Regarding Different Features of the Web-based Questionnaire

Items	Strongly Agree	Agree	Disagree	Strongly Disagree
6. The onscreen instructions were easy to follow.	75 (56)	54 (41)	3 (2)	1 (1)
7. I understood what I was expected to do throughout the web questionnaire.	73 (55)	56 (42)	4 (3)	0 (0)
8. I liked the use of video characters (for example, students and teachers).	75 (56)	40 (30)	17 (13)	1 (1)
9. I liked the use of graphics (for example, desks, chairs, and plants).	55 (41)	63 (47)	14 (11)	1 (1)
10. I would prefer to do other tests (for example, a science test in the style of the web questionnaire.	55 (41)	31 (23)	33 (25)	14 (11)

Note. The percentages of the frequencies in parentheses.

Discussion

The primary purpose of this study was to examine whether dynamic, interactive web-based items, created to be part of a scenario test design, provided more measurement precision than their standard format paper-and-pencil versions. Using data examining self-efficacy in reading progress collected from one group of participants, IRT items parameters, and information functions were calculated at item and test level, for both administration modes.

Descriptive item parameter statistics and both item and test information functions indicated that the questionnaires showed strong discrimination across their respective theta ranges. At an item level, four of the items (Items 1, 3, 5, 7) showed more measurement information from the paper version, and five items (Items 2, 4, 6, 8, 9) from the web version. While there appeared to be no substantive reasons for why some items were more efficient in a particular mode, further analysis found that item length (i.e., number of words) was related to the amount of item information across modes. With the exception of one (Item 1), more item information was obtained in the paper version when items were longer in length (>15 words). Conversely, shorter items produced more measurement information from the web-based versions of items. Relating this finding to the item location parameters, with the exception of one item from each mode, the longer paper-and-pencil items required more of the theta being measured than their web versions. Thus, wordier items required lower levels of reading self-perception via the web items, than in their paper form, conversely, shorter items required lower theta levels when administered in paper mode. Although the purpose of this study was not to examine the design differences between the modal presentation of items, if a relationship is made between levels of perceived reading ability and actual

ability, these findings indicate that given the application of optimal user interface legibility features, longer lengths of onscreen text may be more readable (i.e., comprehension, efficiency) for test-takers even with low reading proficiency levels.

Both TIFs were steeply peaked (e.g., not horizontal) resulting in a restricted theta range over which each test was providing the greatest amount of information. It is proposed that this was due in part to the high degree of discrimination that items from each test were providing, tightly clustered difficulty values, and the small number of items in each test. Although both versions covered similar theta ranges, the web-based questionnaire was most precise measuring participants with a lower belief in their reading progress. This finding suggests that due to both the theta position and the steepness of the TIF, the web-based scenario items is a more effective tool for highlighting which individuals are not confident with the progress that they are making with their reading ability. Conversely, as the paper-and-pencil version provided most measurement efficiency amongst participants with a higher positive perception of their reading progress, this suggests that it is less efficient at highlighting poor reading self-perceptions. Given that the questionnaire response means for both versions showed that "Yes" was selected nearly 100% of the time, the web versions TIF theta location suggests that on the occasions when participants did not perceive self-efficacy, this version was more likely to extract this information.

However, it is in regard to measurement precision that differences are shown between web and paper-and-pencil versions. In psychometric terms, information is reciprocal of precision, thus, the precision with which theta levels can be estimated (Fisher, 1932). Obviously, the more precision an instrument offers the better estimation of a test-taker's true theta level, thus, in IRT terms, the more reliable the measure. As proposed, the information produced collectively by the innovative items in this study

provided more precision than the paper-and-pencil version. Analysis of the relative efficiency of each TIF showed that the paper-and-pencil questionnaire would require an additional two items (with comparable parameters to those in the test) to achieve the level of information provided by the web-based questionnaire. This finding indicates that although innovative items, like those designed for this study, might afford more initial outlay in terms of time and development costs, they have the capacity to provide more information regarding test-takers' theta levels, potentially across fewer items.

Although equivalence is of vital import when transferring paper instruments to screen, as highlighted by The Standards for Educational and Psychological Testing (1999, Standard 4.10), this study did not seek to create an equivalent web-based version of a paper-and-pencil instrument. As the web-based items in this study were radically innovative, thus not bearing any resemblance in design format to the paper questionnaire, measurement equivalence was neither anticipated, nor a focus of this study. Rather, out of interest, cross-mode equivalence analysis was conducted to establish the degree of non-equivalence between the two measures. Given the inherit mode differences, it was not unexpected that the item correlations between the two measures were low. An additional examination of the factor structure of each questionnaire showed that while the web-based version displayed mediocre fit to the single-factor model, the paper-and-pencil fit was poor, failing to suggest any evidence of factorial equivalence between questionnaires. Despite these findings, at a response level participants' scores across each mode were not significantly different, suggesting a degree of equivalence between responses. However, the equivalence result here may be spurious given a lack of variance across response scores due to the dichotomous response format.

In addition to the psychometric efficiency of the two questionnaires, another focus of this study related to participants' attitudes regarding the two administration modes of the Progress items. Findings showed that there was an overwhelming preference for the web-based questionnaire, with the most cited reason relating to the perceived easier nature of the items. Although at the test level the paper-and-pencil and innovative versions were psychometrically similar, or in this instance requiring similar levels of reading self-efficacy, innovative items were perceived to be easier than the paper-and-pencil items. In conjunction with statements regarding the easiness of the web-based items, participants also perceived that they would make fewer mistakes when responding to the innovative items. This may be due to the fact that the web-based scenario questionnaire was very linear in progression (e.g., no reviewing, changing previous answers, skipping items/scenes) which may have provided a sense of security throughout the testing process given that certain requirements had to be met (e.g., response chosen, all possible interactions completed in each scene) before progression could occur. This argument can also be extended to the frequent comment that the web version was easier to follow.

Beyond this reason, the other most cited reason related to the overall enjoyment and fun that was experienced, especially in relation to the interactivity accompanying each scene and dialogue interaction. The issue of interactivity was a major motivation behind the design of the innovative items adopted for the web version. To date, *interactivity* has been related to a broad spectrum of user actions on computer-based tests. At a base level, interaction occurs when a test-taker selects radio buttons for responses, at the more sophisticated level, when the test-taker requires tools (e.g., mouse) to grasp objects for onscreen manipulation, such as drag-and-drop type exercises. This study chose to design innovative items that were more aligned with the interactive nature offered by the

gaming industry, particularly where human characters react behaviorally to commands and interaction by the user. Further, such characters were set within a dynamic and vibrant environment that replicated, in a virtual sense, a reality. Interestingly, these themes were the reasons most frequently cited by participants when responding to what they liked most about the web questionnaire. For example, reasons relating to the interaction given by the student characters in each scene dominated responses, with participants citing they liked the way that the characters in each scene engaged in a dialogue with each other, which included the participant as a passive observer, and focused the questions directly to the participant. Furthermore, many participants felt that the questions were easier to understand when they were spoken to them.

The environment of the questionnaire was frequently cited as a reason for web-based preference. Reasons such as, the interactivity of the graphics and the environment, the virtual reality created of school and classroom scenes, and the visual story type design, correlated strongly with the typical design and inherit appeal of the computer games designed for and marketed primarily at the cohort sampled in this study.

As noted in previous studies investigating test-takers' attitudes towards computerized cognitive ability tests, participants have disliked both the inability to review/skip forwards and backwards through items, and change previous answers (see Luecht, Hadadi, Swanson, & Case, 1998; Revuelta, Ximénez, Olea, 2003; Vispoel, 1998; Vispoel, 2000). In addition, participants tend to find paper-and-pencil versions quicker to complete. These issues were dominant amongst the reasons that participants in this study disliked the web-based version of the Progress items. However, given the problem of error introduced by the often aberrant and irreverent response behavior of participants, particularly amongst non-adult populations, a more directed, controlled, and engaging response environment may result in a more thoughtful, genuine, and less

hurried response set. Interestingly, when responding to which mode they thought best reflected what they thought about their reading abilities, participants cited reasons that indicated that more thoughtful response behavior was produced by the web-based items. Most cited was participants' opinion that they understood more of what was being asked of them regarding their reading progress, when compared to the paper statements. It is proposed that greater understanding was experienced by participants as a result of the delivery of these items by people, set in an engaging environment. As frequently cited by participants, the realistic life-like environment resulted in them experiencing a more meaningful dialogue by, as one participant wrote, "…kids just like me". It is argued that as participants would be less likely to give random, meaningless responses during in situ peer interaction, it may be similarly less likely that such response behavior would occur during a virtual peer interaction. If participants did find the web-based items more meaningful and engaging, this may account for the increased information found collectively across the theta levels captured by this version.

While this study is unique in that it is the first to empirically test the information acquired from multimedia interactive innovative items, it must be noted that only measurement information has been investigated in this study. While the exact reasons why this innovative scenario-based design produced more information were not isolated in this study, several possibilities can be proposed. First, given the high positive response by participants to the interaction with peer-like scene characters in the web-based questionnaire, it could be argued that this interactivity and level of engagement facilitated the increased theta information. It is of interest that the items that showed the greatest measurement precision, in comparison to their paper-and-pencil versions, were items that had more lead-in dialogue preceding the item. It is possible that this increase lead-in scripting and contextualizing of the events prior, during, and posttest, engaged

the participants to an extent that impacted on the focus and understanding of the question, hence clarity of the response. Again, this is speculatory and requires further examination under a specific experimental design to isolate the impact that these aspects of innovative item design may have on measurement efficiency. Another interesting finding related to the perceived easiness of the web-based questionnaire. Further research is required to see if this dominant reason for web-based preference would extend to cognitive ability based assessment. If so it could be that the typically difficult (or perceived difficult) concepts may be more effectively, from a learning perspective, delivered via a multimedia format.

The findings of this study must be interpreted in the light of some potential limitations, and from which future research should be directed. First, although a strength of this study lies in the exposure of the same sample to both administration modes, the participants in this study were from schools where the use of technology is emphasized, and where reading achievement levels are typically above national averages. Thus, it is advisable for the sake of generalizability, that these items should be administered to a wider student population that is more reflective of the schools throughout the country. Second, this study did not focus on the validity of the construct being captured by the innovative items. As comparison to a criterion might prove difficult given the innovative design of the items, a multi-trait multi-method analysis whereby responses are matched to both convergent (e.g., teacher feedback, reading achievement scores) and discriminant evidence would provide useful information, if methods confounds can be controlled.

In general, research on web-based testing has lagged behind web-based testing practice (Lievens & Harris, 2003), and further work is required to establish more psychometric properties of multimedia driven innovative items. This study provides promising empirical evidence, at both a psychometric and test-taker level, for the

benefits in leveraging technology to develop dynamic interactive innovative items for use in assessing personality-based psychological constructs. The results suggest that in addition to providing a dynamic and engaging platform for tests and questionnaires to be delivered, multimedia items, such as those examined in this study, may provide more information about the participant than traditional paper-and-pencil measures.

CHAPTER SIX

DISCUSSION

The research presented in this thesis examined two interrelated measurement problems that have been associated with assessment of non-cognitive ability constructs, namely the way personality data is modeled and the mode by which we administer these scales. Within these two separate, yet complementary measurement issues, the overarching purpose of the studies in this thesis has been concerned with exploring issues that relate to the maximizing of information derived from personality measures. Over three studies, the research explored the different explanations of how test-takers respond towards personality scales, and which theoretical position, and thus modeling approach, best represents this behavior. Furthermore, modal issues were examined in relation to the human and technological issues that arise when computerizing tests, and some of the innovations that have been applied to item design and test development. Finally, this research developed nine innovative items, designed to utilize the full gambit of multimedia features available at the time of their development. Based on the literature review's finding that there exists a scant amount of research examining the psychometric properties of innovative items, particularly in relation to measurement efficiency, the information provided by the innovative items were compared to their paper-and-pencil versions. Driven also by the review finding that some test-takers experience anxiety, or are impacted negatively because of a lack of computer familiarity, feedback was obtained in relation to what version of the scale test-takers preferred and the reasons for this choice.

The first section of this chapter provides a summary of the methods adopted and findings from each of the studies in this thesis in relation to the research questions and propositions discussed in Chapter One. The second section will discuss the implications of these findings and suggests future research possibilities. Finally, the contribution made across these studies will be reviewed.

The first empirical study related to the nature with which response behavior associated with personality measures has been traditionally viewed and based on these assumptions, the typical modeling and scale construction approaches that have been adopted. Developing earlier arguments (Coombs, 1964; Cronbach, 1949; Thurstone, 1928, 1931), recent research (e.g., Chernyshenko et al., 2001; Maydeu-Olivares, 2005; Roberts et al., 1999) has suggested that personality responses might be more appropriately handled based on assumption of typical performance criteria, where there is no specific knowledge that needs to be recognized or recalled. In order to represent this assumption to personality scales, an ideal point approach is necessary to model this type of response data. The ideal point response process suggests that a test-taker will positively agree with an item if the test-taker is located in close proximity to the item on the trait continuum. While previous research has produced mixed results regarding the model-data fit achieved by ideal point modeling, with the exception of Chernyshenko (2001, 2002) who dichotomized polytomous data, estimates have been generated from response data derived from scales constructed under dominance assumptions. Further, no study to date, has investigated the fit provided by a dominance IRT model from responses derived from a scale constructed given ideal point assumptions. As such, the first proposition suggested was that an ideal point modeling of response data would provide better fit to responses derived from an ideal point constructed personality scale, than the dominance-based GRM. Thus, due to the theoretical alignment between the

construction and estimation procedures, it was proposed that the GRM would struggle to model the ensuing response patterns.

This study aimed to answer the question: Which measurement model provides the best fit to the ideal point constructed Academic Self-Worth Scale? Participants from four secondary schools completed the Academic Self-Worth scale, which was modified from pre-existing sub-scales from the Academic Self-Handicapping Scale (Midgley, Arunkumar, & Urdan, 1996), Life Orientation test (Scheier, Carver, & Bridges, 1994), Academic Process Questionnaire (Martin, 1998), and Defensive Pessimism Questionnaire (Norem & Cantor, 1986). The 5-point Likert scale was developed using ideal point construction guidelines provided by Chernyshenko (2002), whereby experienced item writers identified where items were positioned on the trait continuum. Additional neutral and extreme reflections of the three sub-scales were written by experienced item writers to ensure that the entire trait continuum was represented. Using exploratory maximum likelihood factor analysis with an oblique rotation, it was possible to determine the items association with the theoretical underlying dimensions of each sub-scale. Further IRT item parameter estimations, cross-validation analysis, and model-data fit were conducted to evaluate which model provided the best fit to the response patterns.

With the current trend towards increased use of computers for test delivery and development, important issues relating to the evaluation of equivalence of tests under both paper and computer modes is especially important. Thus, the second study focused on issues relating to the transference of paper-and-pencil tests to screen, and how these might have an adverse impact on test-taker performance. The review also examined some of the various innovative item types that have been designed specifically for computer- and web-based administration. In addition to the psychometric properties of

such items, focus was also given to literature relating to the task-related skills that are necessary for test-takers to answer these items. Thus the research question for this study was: What are the factors that produce a mode effect between paper-based and computerized tests?

This review incorporated literature from a diverse range of disciplines and sectors. First, psychological and educational journals were examined in relation to the influence that various participant characteristics, such as race, ethnicity, and gender might have on mode performance. In addition, literature from these disciplines provided valuable research relating to the cognitive processing differences (e.g., memory, comprehension, reading speed) and the impact of prior ability/academic levels between modes, and test-takers' familiarity or anxiety in relation to computer use. However, the user interface variables posited by Muter (1996) relating to the non-equivalence between paper and screen, required reviews of research relating to issues surrounding usability and human-computer interaction. Information relating to these areas was typically presented in literature aimed at the screen layout and design of web sites or computer application, and rarely, was this literature specific to computerized tests. Therefore, such information was included into the review based on both the import given by Muter, and perceived current relevance to testing applications. For clarity, the mode effect literature was presented in two distinct sections. The first section related to the human and technological issues surrounding impact of computerized testing, where the second section focused on the test-taker interactivity issues relating to existing computer-based items. Where available, the psychometric properties and empirical findings for current computerized items were presented, in particular those where statistical mode effects (e.g., measurement equivalence) had been examined.

In light of the many mode effect and innovative item design findings presented in the literature review, the second empirical study of this thesis was developed. History shows that the field of assessment is no stranger to technological advancement. This was evidenced by the rise of multiple-choice methodologies in the middle of the last century, which was driven initially by the development of high-speed scanners (Parshall, Spray, Kalohn, & Davey, 2002). A more recent advance has been the computer scoring of essays and innovative item formats (Clariana & Wallace, 2002); and the use of adaptive tests where the computer program, based on previous responses, chooses the next best item to administer to the candidate (Jones, 2000). Because of the increased affordability and the computational ability of modern computers, computerized assessment has gained a more prominent role (Tonidandel, Quiñones, & Adams, 2002). This impact has been particularly evident in computer-based tests (CBTs) where the use of computers has provided test developers with the opportunity to re-envision what test items look like and how they are scored (Zenisky & Sireci, 2002). However endless the design possibilities are in relation to computerized items, it is essential that the psychometric integrity of these items are retained (or established) in order for confident application. As the literature review highlighted, there is limited research available offering information relating to the psychometric properties (e.g., validity and reliability) of innovative items. The research that has been conducted in this area typically relates to the issue of equivalence (i.e., validity), where various analyzes (e.g., factorial and cross-mode correlations) establish the degree of difference between a computerized and paper version of a test, at a psychometric level. While this type of examination offers invaluable information regarding any structural and/or dimensional changes in the construct occurring across modes, equivalence analysis does not provide any indication of changes in the measurement efficiency (i.e., reliability) of each test. In the only study

of measurement efficiency Jodoin (2003) found that the various cognitive-based innovative items (e.g., drag-and-drop, create-a-tree) provided more measurement information across all theta levels, than their paper version. However, this finding may have been confounded by the polytomous nature of the items, where previous studies (e.g., Donoghue, 1994; Samejima, 1976; Thissen, 1976) had found that polytomously scored items typically provided more information than dichotomously scored multiple-choice items. As such, the impact that the test mode and item design had on the measurement information found is uncertain. Therefore, based in part on Jodoin's findings, and on the assumption that dynamic and interactive scenario-based items might elicit greater construct information, the second proposition in this thesis proposed that, when compared to the paper version, innovative items will provide more measurement efficiency across all proficiency levels captured by the measure. In addition to establishing psychometric information, gaining feedback from a scale is especially important given the still unique aspects surrounding onscreen items. This is even more important when test-takers have been exposed to highly dynamic and interactive test experiences that, due to the design's uniqueness, they may not have had previous experience or exposure. Given the innovative items engaging and interactive design, the final proposition of this thesis asserted that test-takers would prefer the web-based innovative version of the RSPS Progress scale. Thus, this study aimed to answer the following three research questions, first, which item type provides the most information (least error) across theta levels at an item level, second, which test type provides the most information (least error) across theta levels at a test level, and lastly, which version of the questionnaire will be better perceived by the test-takers?

A dichotomous ("Yes"/"No") answer response was used, with cross-mode analysis, paired t-test means, and confirmatory factor analysis examined in order to establish the

degree of equivalence across both scale versions. Item parameters from both versions of the scale were generated from the dichotomous response data and used to assess the measurement efficiency of the administered items.

Findings

The proposition for Study One suggesting that the ideal point GGUM modeling approach would provide superior fit to the ideal point constructed Academic Self-Worth Scale was not supported across the total distributions of responses. Instead, the GRM was found to provide response modeling comparable to the GGUM across all levels of the trait continuum. Although graphical analysis of the theoretical and expected response functions indicated that the GGUM estimations did represent the data well, the GRM showed less discrepancies between empirical proportions and estimated option response functions in nearly half of the items. Statistical fit analysis generally supported these model-data fit results, with chi-square fit analysis showing that the GGUM response modeling was similar to that of the GRM, and only producing better fit for a quarter of the items in the scale. Thus, the findings indicated that the while the GRM applies a monotonic response modeling approach to non-cognitive data, this may be valid and appropriate despite the theoretical supposition of this estimation model being based on the assumption of dominance responding behavior, and despite estimates being derived from a scale constructed using an ideal point methodology.

The literature review investigating the mode effect of delivering tests onscreen found many important human and technological issues that need to be considered when either transferring paper tests, or developing innovative items. Of the mode effect variables relating to human issues, the most prominent focus had been applied to race/ethnicity and gender, memory and comprehension, speededness, ability, computer

familiarity, and computer anxiety. Amongst these, test-taker ability levels and cognitive processing issues (e.g., memory, comprehension, and speededness) had the greatest potential for adverse impact from onscreen test delivery, although results were variable. For example, Mead and Drasgow (1993) argued that their meta-analysis showed that test mode had its greatest impact in timed tests, and concluded that this might relate to a screen readability (speed and efficiency) issue. Other studies have also found that text read from screen typically takes longer than that from paper. While confounded with test-takers' ability/academic performance, memory awareness patterns and associated cognitive processing appears to be different when processing information from screen. Other research has suggested that low ability students are more adversely impacted by the mode effect than their high ability peers. However, these studies have often been confounded by the test-takers' familiarity with computers; that is, high-ability students have tended to be more frequent computer users. This result suggests that as computer use becomes more mainstream, its moderating effect in relation to performance will reduce.

User interface issues were found to be related to either the legibility of screen features (e.g., font, resolution, whitespace), or specific to the actions involved in interacting with aspects of the onscreen application (e.g., scrolling, item review and presentation). This review found that high screen resolution and a 12pt font size in Arial, Times New Roman, or Tahoma (or 14pt font size in Arial or Comic) enhanced the readability of onscreen text. In addition, the length, number, and spacing of lines have been found to impact readability, particularly reading speed. Here, adults' readability was optimized when line lengths fell between 74.8cpl and 100cpl, whereas children preferred a shorter length (45cpl). Reading efficiency was reduced when only one to two lines of text were presented, whereas, four lines or more (to a full screen) produced

greater and similar readability findings. Only interline spacing research showed that existing paper typology rules were applicable for onscreen text. The interactive features of onscreen applications produced results that both objectively and subjectively related to test-takers' performance. Various studies have indicated that scrolling has a detrimental impact on performance, although the degree of this impact has been inconclusive. Proposed reasons for this impact have been the visual memory of test-takers is affected due to the temporary disconnect of presented information. The way in which items are presented to the test-taker were found to increase errors and produced hurried responses. These issues were particularly apparent when items were presented one at a time onscreen. There was less impact on test-taker performance when several items were presented onscreen at a time. It was proposed that this might be because of the test-takers' ability to review or skip items, in addition to using the information presented in previous items to assist responses. Although the opportunity to review items in a test has not shown to impact on performance, test-takers overwhelmingly prefer having this option available for computerized tests.

The findings from both the ideal point and mode effect studies in this thesis influenced the design and development of the web-based innovative items administered in the final study. First, given the unique response behavior proposed by Coombs (1964) and Thurstone (1928, 1931) to occur during personality assessment, innovative items were designed to mitigate, as much as possible, the traditional "cognitive-based" testing condition. Using an interactive scenario-based design, items presented a virtual experience to the test-taker. It is proposed that this highly engaging and dynamic design would promote and enhance a typical response behavior from test-takers, thus, negating the test-like "right/wrong" environment from which a dominance response behavior might be exhibited. Therefore, while an ideal point model (e.g., GGUM) was not used in

the third study to model response patterns, due to the strong performance of the dominance approach (i.e., GRM) in the first study and lack of the sample size ideal point estimations, the theoretical foundation of ideal point was employed in the design of the assessment conditions presented to test-takers.

It is unfortunate that little is known currently regarding the design and user interface issues related to optimizing innovative items. However, the literature review of the key factors associated with test mode effects revealed some design fundamentals that should be taken into account when designing computerized tests, and these were applied in the design of the innovative items in Study Three. Given findings regarding the potential adverse impact of familiarity and anxiety on test-takers' performance and test experience (e.g., McDonald, 2002; Taylor, Kirsch, Eignor, & Jamieson, 1999), clear onscreen instructions were provided at the beginning of the test. Test instructions outlined both what the test-taker will experience and what the test-taker was required to do in order to interact with the onscreen characters, and progress within and between each screen. Based on Powers and O'Neill's (1992) findings showing the benefits of providing pretest training/practice, a practice question was presented prior to the beginning of the test, enabling the test-taker to become familiar with the response form and mouse action required to submit their response. Optimal design aspects were also applied in relation to the user interface of the innovative items. In relation to minimizing impacts from onscreen readability and legibility of text, suggested font size and style, line length and spacing, and whitespace were applied. All text was presented using Comic font style in 16pt font size given the review's findings that this style and size was preferred by children, who found this combination less test-like and formal (Bernard, Mills, Frank, & McKown, 2001). Line length of text (e.g., instructions, item questions) were kept to within 45cpl (Bernard, Fernandez, & Hull, 2002) and interline spacing set

at 18pt (e.g., 2pt above the size of the font) to reduce eye fixation points, thus enhance legibility and readability (Lynch & Horton, 2002). Given the essentially graphical design of the innovative items, whitespace was only an issue in relation to the response form presented at the end of each item scenario. As the response form presented only one item at a time, the form object was reduced to consume only one-third of the screen. This resulted in little whitespace surrounding the response form text, thus maximizing the comprehension of the information being presented (Bernard, Chaparro, & Thomasson, 2000). Given the interactive nature of the innovative items, the mode issues relating to scrolling and item review features, were of particular import to the design considerations. As scrolling has been found to have a detrimental impact on performance (Choi & Tinkler, 2002), the innovative items were specifically designed so that no scrolling was required for any of the user functionality. In scenes requiring the test-taker to move (e.g., side-to-side) beyond the initial visible screen in order to interact with the screen characters, the mouse pointer and onscreen hotspots (i.e., arrows) moved test-takers automatically to these scene areas.

Due to the sequential scenario-based nature of the design, test-takers could not skip and revisit items. However, given that research has found the ability to item review is a preference for test-takers (Luecht, Hadadi, Swanson, & Case, 1998), the response form presented the option of replaying the entire scene relating to the current item.

The last two propositions in this thesis relate to the second empirical study that compared the amount of measurement precision provided by web-based and paper-and-pencil items. The main proposition of this study asserted that the innovative scenario-based items would provide more measurement efficiency across the proficiency levels captured by the measure. Both graphical and statistical fit analysis confirmed this proposition at a test level, and for over half of nine innovative items, but only amongst

students showing low theta (i.e., reading self-perception). At the item level, the IIF plots for both item types revealed a mixture of findings, where four items showed superior measurement information in their paper-and-pencil versions, and five items displayed more information in the web-based version. While there appeared to be no substantive reason for this finding in relation to item content or structure, a difference was found in the amount of words in each item. The items that showed more measurement efficiency in the paper version were, with one exception, longer on average by six words. Conversely, greater measurement precision was achieved by shorter items (≤14 words) when delivered in their innovative web-based version. Further analysis showed that typically location parameters were higher (e.g., moving towards the positive end of the continuum) under the mode that produced the greater measurement efficiency. Specifically, the longer items (≥16 words) provided more measurement information via paper administration, but required more of the theta being measured to respond positively. Thus, the wordier items required less perceived proficiency in reading to respond positively on the web, but were less discriminatory in this mode. Conversely, responses to shorter paper-and-pencil items required less perceived reading proficiency, but provided less discrimination amongst this area on the trait continuum. Thus, a hypothesis might be that when following the optimal legibility issues of font size and style, line length and spacing, and whitespace conventions specifically for children, onscreen presentation of longer chunks of text required lower levels of perceived reading efficacy, than the paper version of those items. In contrast the smaller font (12pt) and perceived more formal font (Times New Roman) in the paper version, required higher levels of perceived reading efficacy amongst test-takers when items were longer. As test-takers with higher perceived reading proficiency almost certainly had greater reading efficiency and comprehension skills, this finding indicates that less

reading proficiency/verbal ability is required when reading larger chunks of test onscreen, if user interface issues are optimized for the specificity (e.g., age) of the test-takers.

A more conclusive finding was apparent when information was summated and analyzed at a test level. Here, the TIF plot for each test version was compared along with their respective TIF values. An overlay of both TIFs showed graphically that the innovative items collectively, were providing more measurement information across a wider theta range than the paper items. While both measures showed maximum precision for lower levels of theta, the innovative items showed more measurement precision over a wider lower theta range than the paper-and-pencil version. The salient finding here is that despite the loss of discrete information resulting from the "Yes"/"No" response format and the high positive response to both sets of items, the innovative format was able to elicit a higher degree of discrimination from the occasions when test-takers' perceived reading progress was poor.

This graphical finding was supported by the relative efficiency statistic of each TIF. These statistical results showed that the web-based test was functioning, across the lower theta range, as though it was nearly 20% longer than the paper version.

The final proposition proposed that, given the highly interactive and engaging design of the web-based test, test-takers would prefer this mode to the traditional paper-and-pencil version. Participants' responses to 10-item post-assessment feedback measure revealed a distinct positive response to the web-based test. Findings showed that if offered the choice, 72% of test-takers would choose to sit the web-based version of the test. Of the reasons given for this choice, the most frequent referred to the perception that the web-based test was easier, more enjoyable, interesting, and interactive. Further,

test-takers perceived that the web-based test reflected better, what they thought of their reading self-efficacy, with the majority stating that they had more of a clearer understanding of what was being asked, and that the questions appeared more meaningful and real to them. The most referred to reason why the test-takers liked the web-based test were the preference for being asked the questions by the onscreen characters.

Implications

The findings from the three studies in this thesis present some significant implications regarding how personality constructs should be designed and how they are assessed. In relation to the response behavior assumed when choosing the approach with which to model personality response patterns, the findings from this study suggest that both ideal point and dominance positions appear to be equally applicable. The model-data fit provided by the dominance-based GRM to data derived from an ideal point scale suggests that this model can provide acceptable fit to items across the full trait continuum being measured, including items positioned during scale construction at neutral and extreme theta levels. An alternative argument is that the ideal point scale constructed for this measure did not, despite diligent adherence to the methodology, actually represent a true ideal point scale. Interestingly, few of the fit plots produced by GGUM showed a nonmonotonic response function. This issue was also found by Chernyshenko (2002) whose data from an ideal point scale similarly showed monotonic tendencies. Had Chernyshenko applied a dominance-based IRT model to his data, he too may have found good model-data fit was achieved. As the scale construction guidelines for Study One were from Chernyshenko's study (which was based on the Thurstonian

approach), it might be deduced that the construction approach detailed may not actually be producing truly ideal point (nonmonotonic) measures.

In terms of the modal differences between computer and paper administration of items, the review in this thesis highlighted the various human and technological issues that must be taken into account when developing or transferring tests to screen. While it is proposed that due to exposure, the psycho-emotional perception of adverse impact experienced by some test-takers will gradually diminish, computer familiarity and associated anxiety still appears to have a significant impact on performance. Although general usage will increase the familiarity related to response actions via keyboard and mouse, it may not remove anxiety associated with usage related to often high-stake specialized testing applications.

While the review listed a varied array of legibility and interactive mode issues, all relate to the overarching issue of onscreen readability. It is essential that screen layout and environment is designed to optimize reading from the screen, and reduce impact on performance. For example, if a difference between measurement information across modes is a factor of the length of items, rather than the primary dimension being measured, then these findings might indicate that the test-takers' verbal/reading ability to comprehend and respond to each item's complexity may be an important consideration in onscreen item design.

In relation to the user interaction aspects of onscreen testing, Sweller's (1988) cognitive load theory is particularly relevant given its focus on the impact of onscreen information elements, design, and degree of interactivity, which results in a load on test-takers' working memory and impact on performance. Given that all that individuals "know" about the world and themselves is held in their long-term memory (Feinberg &

Murphy, 2000), the application of this theory as a baseline for effective computer or web-based test design would be relevant for both cognitive ability and non-cognitive ability measures. While there are very high levels of measurement potential offered by innovative items, it is imperative that the possibilities are not comprised or confounded by mode-based construct-irrelevant variance introduced from legibility and interaction features.

The last study of this thesis investigated the measurement efficiency provided by web-based innovative items that were designed to assess participants' perceived progress in their reading abilities. Amongst test-takers who have a lower perception of their reading proficiency, the web-based innovative items provided 20% more measurement information than the paper-and-pencil versions. It is possible that due to the interactive, engaging, and one-on-one nature of the innovative items, test-takers may have provided responses with less associated response bias (e.g., social desirability), resulting in finer discrimination and less error than paper-and-pencil responses to the test. Thus, innovative items might be particularly useful in providing information from items that induce response bias (e.g., social desirability, impression management, self-presentation) which are typically associated with item content that has a degree of sensitivity, deprivation, or vulnerability (e.g., self-concept) for the test-taker.

Feedback from test-takers indicated that most preferred the interactive web experience to the traditional paper-and-pencil approach. Given the age of these test-takers, it is more than likely that they have had significant exposure to computers (i.e., Internet, gaming applications); therefore, it is not unexpected that they preferred this mode of test delivery. This finding reinforces the argument that issues such as computer familiarity and anxiety will have less adverse impact given the wide exposure to computers (Levine & Donitsa-Schmidt, 1998; Todman & Lawrenson, 1992). However,

attention should also be given to the fact that 25% of the test-takers preferred the paper version of the measure. Amongst these test-takers, the majority of reasons for paper-and-pencil preference related to the fact that test-takers found the paper-and-pencil items quicker to complete. Thus, preference for the paper version was not related to a lack of confidence regarding completing the web-based version, rather, the design of the innovative items forced a speed of progression on the test-takers. From a psychometric perspective, the hurried response permitted by the paper version potentially introduces the aberrant response error that is inherently difficult to control especially amongst non-cognitive ability measures. Potentially, the smaller amount of error associated with the innovative items at the lower end of the theta, might be a function of the directed and controlled design in this version.

Future Research

Further research is required to establish how valid the current method is for constructing ideal point scales. Although Chernyshenko's (2002) approach is true to the ideal point theory espoused by Thurstone and Coombs, it is important that it be assessed for its own merit. Obviously, this process is an important prerequisite before categorical assumptions can be made as to the superiority and appropriateness of an ideal point approach to non-cognitive data. Until this scale construction method can be validated, the implication from this study is that the GRM provides an adequate representation of personality response patterns, even when the scale has not been developed with an aligned theoretical framework. In addition to scale construction, future research should investigate the relationship between innovative design and ideal point response patterns. Specifically, does a highly engaging measure, which does not resemble in design or function a traditional "right/wrong" test, induce the type of typical response behavior

proposed to occur when responding to personality/attitude measures? Given that studies investigating manipulations in font style and size has found an impact on individuals impressions and behavior regarding the type of the text they are reading (e.g., formal/informal), it may be possible that in addition to the item content, the "look and feel" of a virtual or scenario-based design accentuates typical response behavior.

Future research should explore further the impact of pre-familiarization tutorials and pretest training on performance differences and perceived familiarity/anxiety. Particularly, it should not be presumed that increased exposure and use of computers would negate the anxiety associated with completing computerized tests. While computer use is becoming a large and essential part of everyday life, most users do not regularly participate in completing psychological testing onscreen. Therefore where possible, the anxiety that is produced from "being tested" needs to be isolated from any anxiety resulting from completing measures onscreen.

As mentioned earlier, much of the research sourced in the review relating to user interface has not been examined or conducted in relation to a computerized testing environment. Focus needs to be given to the user interface issues that relate specifically to onscreen testing environments, particularly where there may be the use of multiple split screens, multimedia sections, and embedded windows. In addition, further research needs to establish how many multiple forms of information (e.g., text, video, and graphics) can be displayed onscreen before cognitive load, comprehension, or readability might be compromised.

Of the user interface research that has been related directly to CBTs, scrolling has received primary focus. Many studies have found scrolling to produce a significant negative impact on test-taker performance. Such findings have suggested that scrolling

impacts on the cognitive processing and reading efficiency of the test-taker, where it is proposed that the visual memory is adversely impacted upon, thus impacting on readability. Future research needs to establish the size of this impact on performance. If scrolling is confirmed to be a significant confound, research then needs to investigate the advantages that text location aids (e.g., electronic markers), or alternative text presentation options might have in reducing the impact of this mode effect. Given the inevitability of some degree of scrolling when presenting onscreen text, and the potential impact on performance, it is important that such research is conducted and possible solutions developed.

Similar issues surround item presentation and item review, which has received considerably less research attention, but impact significantly on the test environment and test-taker's experience. In relation to item presentation, the review found only three studies focusing on issues relating to the optimal presentation of items, with the most recent conducted 17 years ago. It is neither sufficient nor appropriate to relate the findings from these studies to current technologies and test environments. In order to ascertain which test design offers the least impact on test-takers' cognitive load and performance, future research should examine the impact of several different item presentation approaches, in addition to investigating reviewing and skipping options *within* these test designs.

In sum, Lonsdale, Dyson, and Reynolds (2006) posited that assumptions regarding how to optimize test design should not be based only on experience; instead, investigations need to show how effective design and typographic features can be manipulated to present test material that is optimal in regards to conceptual and perceptual processing, and legibility. Additional research into the numerous typology issues highlighted in the review would assist in the creation of onscreen user guidelines

that are specifically aimed at computer-based testing systems. Such studies would provide insight into the impact of different test interfaces and designs on test performance, thus aiding in the design of innovative items that provide high measurement precision and discrimination, and which are not confounded by nuisance or secondary dimensions.

From the few investigations on cognitive processing issues, such as mode impacts on memory and comprehension, findings have been largely inconclusive. The cognitive processing demands under each mode should be explored further to establish if indeed there is a difference as suggested by Noyes and Garland (2003). Future investigations may benefit from applying a foundational theory, such as cognitive load theory (Sweller, 1988) to examining potential differences in cognitive processing. Based on working memory capacity, cognitive load theory proposes that optimum learning or comprehension is achieved only when the cognitive load on the working memory is minimal, thus permitting successful transference to long-term memory. Understanding the cognitive load generated from graphical user interfaces and multimedia designs, like the innovative items developed in this thesis, is vital in ascertaining and controlling construct-irrelevant variability in performance.

This review highlighted that significant research is required to establish the psychometric properties of onscreen innovative item types. Only after such investigations can a progression be made from the current focus between mode comparisons (i.e., paper vs. computer) to within mode comparisons across different classes of innovative items (i.e., text and graphics-based vs. multimedia-based). Such research is vital for the establishment of the validity and reliability estimates for innovative measures. For example, there may be unique properties associated with multimedia items that are not found, for example, amongst text and graphic items. Thus,

conclusions from one type of innovative design may not be generalizable to another, albeit they have been developed for the same medium.

The establishment of the measurement efficiency given by various computerized item design types, will help to establish whether the findings in this thesis are reproducible with other designs (e.g., simple online response form, video-only or graphic-only content, multimedia), and what designs provide superior measurement precision. Such investigations will establish whether pre-existing measures are being enhanced by the development of innovative versions, and if so, by what degree. Based on the outcomes of these studies, test developers can make informed decisions regarding whether the extra cost (e.g., design and development) and necessary equivalence and efficiency examinations are worth the amount of information being captured.

Although the measurement equivalence between the two versions was not the focus of the last study, the lack of comparable factor structure found adds to the inconsistent findings in previous investigations of validity between the modes. Future research should establish if factorial stability is occurring only across modes when there has been a direct transference of paper design to screen, or whether factor structures hold even across vastly different designs, such as the innovative items design in this thesis. Such investigations would aid in establishing both what innovative tests measure and, if different underlying constructs are being tapped, how these should be interpreted. Only when paper and innovative items show that the same underlying construct is being measured, can the same interpretations be applied across both modes. Obviously, if each mode is capturing different constructs, then these need to be identified and different interpretations applied.

Furthermore, given that the development, design, and construct relevancy of various innovative items is still very much in its infancy, it is important that in addition to psychometric investigations, the feedback from test-takers be sought to establish issues relating to the perceived relevancy (i.e., face validity) and impact of interactivity requirements (i.e., task/response actions) of items.

Contribution

While personality research has received an increased focus over the last decade, two areas in particular have recently challenged both the way we model personality responses, and the mode by which personality items are presented. This thesis contributed substantially to both of these areas. The first area centered on the theory of an ideal point response process, which presumes that there are specific motivating factors that dictate response behavior to non-cognitive ability items. Given that respondents are not motivated to maximize their attitude or personality scores, it has been proposed that test-takers display behavior that is typical of the individual at the time of the test. Thus, the first study compared two major models for building and analyzing non-cognitive ability scales – the ideal point and the dominance approach. It has been argued that the ideal point model is inappropriate for the "neither right nor wrong" non-cognitive ability responses. No previous research had compared the modeling capabilities of a dominance IRT model to response data generated from an ideal point constructed scale. Further, this is the first research to maintain the original polytomous nature of the personality scales responses for cross model comparison.

The second area investigated the issues and design options related to the transfer of pre-existing paper-and-pencil measures to screen, specifically, the adverse impacts of responding onscreen, and the psychometric properties of such items. In relation to the

mode effects on test-taker performance, this thesis presents the first literature review of research that focused on both the human and technological factors. Because of the scarcity of research directed at the technological factors and user issues relating to computerized tests, research from various areas, such as human-computer interaction (HCI), user interface (UI), and typography were reviewed. Most research has focused on the human factors associated with the mode effect (such as computer familiarity and anxiety), with little research directed at the technological user interface (UI) and presentation issues. In addition, of the research that has been conducted, almost all has been focused on cognitive-ability tests. In addition, this review has provided an extensive collection of innovative item examples and related psychometric information.

As significant additional resources are associated with the development and design of innovative items, the measurement precision gained from these approaches is particularly important to establish. Study Three developed and empirically assessed the measurement efficiency of a series of web-based innovative items. Using the latest multimedia development tools available at the time of this study, nine innovative items were developed that combined videoed objects within a graphical environment. This is the first innovative interactive web-based measure to be developed to assess non-cognitive ability data. In addition, the innovative items were the first to use a scenario-based test design which presented a story-like structure, and test characters (i.e., screen actors) interacting directly with the test-taker. From an analysis perspective, this research provided only the second study to compare the amount of measurement information being offered by an innovative item and its traditional paper-and-pencil version. In addition, it is the first time that this type of analysis has been applied to a personality measure.

In the context of personality domains, the two related areas of data modeling and modal presentation are concerned with maximizing the information that is derived from personality-based measures. Hence, where ideal point modeling seeks to accurately reflect individuals' positions on all areas on the trait continuum rather than misaligning respondents with divergent or polar traits, computerized item designs seek to not only assess traits in a more engaging and enriched manner, but to also increase the amount of construct information derived. Thus, this thesis contributed to both of these areas substantially through the creation of a thorough review of paper to screen issues, and the empirical examination of both modeling and mode approaches.

The three studies in this thesis developed recent arguments surrounding non-cognitive ability data, reviewing test administration effects, and empirically measuring competing modeling approaches, and innovative item measurement efficiency. It is proposed that given the substantial increase in the use of attitude and personality measurement tools across many areas of industry and academia, the areas focused in this thesis are highly relevant for the development of more sophisticated and arguably more psychometrically appropriate attitudinal and personality measurement approaches.

REFERENCES

AERA, APA, & NCME (1999). Standards for educational and psychological testing. Washington, D.C.: Author.

Andrich, D. (1996). A general hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology, 49*, 347–365.

Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement, 17*, 253–276.

Ashworth, S. D., & McHenry, J. J. (1992, April). *Developing a multimedia in-basket: Lessons learned.* Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.

Baker, F. B. (2001). The basics of item response theory. In *ERIC clearinghouse on assessment and evaluation*. College Park, MD: University of Maryland. Available http://ericae.net/irt/baker.

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, *37*, 122–147.

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist, 28*, 117-148.

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Multifaceted

impact of self-efficacy beliefs on academic functioning. *Child Development, 67*, 1206-1222.

Barlow, R. E., & Proschan, F. (1996). *Mathematical theory of reliability.* SIAM: Philadelphia.

Bartram, D., & Brown, A. (2004). Online test: Mode of administration and the stability of OPQ32i scores. *International Journal of Selection and Assessment, 12,* 278-284.

Baumeister, R. F., & Scher, S. J. (1988). Self-defeating behavior patterns among normal individuals: Review and analysis of common self-destructive tendencies. *Psychological Bulletin, 104*, 3-22.

Bennett, R. E., & Rock, D. A. (1998). *Examining the validity of a computer-based generating-explanations test in an operational setting*. (GRE Board Professional Report No. 93-01P). Princeton, NJ: Educational Testing Service.

Bennett, R. E., Goodman, M., Hessinger, J., Kahn, J., Ligget, G., Marshall, H., & Zack, J. (1999). Using multimedia in large-scale computer-based testing programs. *Computers in Human Behavior*, 15, 283-294.

Bennett, R. E., Morley, M., & Quardt, D. A. (2000). Three response types for broadening the conception of mathematical problem solving in computerized tests. *Applied Psychological Measurement, 24*, 294-309.

Bennett, R. E., Morley, M., Quardt, D., & Rock, D. A. (2000). Graphical modeling: A new response type for measuring the qualitative component of mathematical reasoning. *Applied Measurement in Education, 13,* 303–322.

Bentler, P. M. (1992). *EQS structural equations program manual*. Los Angeles, Calif. BMDP Statistical Software.

Bernard, M., & Mills, M. (2000). So. What size and type of font should I use on my website? *Usability News, 2.* Retrieved January 10, 2006, from http://www.surl.org/usabilitynews/22/font.asp

Bernard, M., Chaparro, B., & Thomasson, R. (2000). Finding information on the web: Does the amount of whitespace really matter? *Usability News, 2.* Retrieved January 2, 2006, from http://www.surl.org/usabilitynews/21/whitespace.asp

Bernard, M., Fernandez, M., & Hull, S. (2002). The effects of line length on children and adults' online reading performance. *Usability News, 4.2*. Retrieved February 3, 2006, from http://psychology.wichita.edu/surl/usabilitynews/42/text_length.asp

Bernard, M., Lida, B., Riley, S., Hackler, T., & Janzen, K. (2002). A comparison of popular online fonts: Which size and type is best? *Usability News, 4*. Retrieved January 2, 2006, from http://www.surl.org/usabilitynews/32/font.asp

Bernard, M., Mills, M., Frank, T., & McKown, J. (2001). Which fonts do children prefer to read online? *Usability News, 3*. Retrieved January 10, 2006, from http://www.surl.org/usabilitynews/31/fontJR.asp

Bernard, M., Mills, M., Peterson, M., & Storrer, K. (2001). A comparison of popular online fonts: Which is best and when? *Usability News, 3*. Retrieved January, 10, 2006, from http://www.surl.org/usabilitynews/32/font.asp

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item

parameters: Application of an EM algorithm. *Psychometrika, 46*, 443–459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35*, 179–197.

Boodoo, G. M. (1998). Addressing cultural context in the development of performance-based assessments and computer-adaptive testing: Preliminary validity considerations. *The Journal of Negro Education, 67*, 211-219.

Booth, J. F. (1998). The user interface in computer-based selection and assessment: Applied and theoretical problematics of an evolving technology. *International Journal of Selection and Assessment, 6*, 61-81.

Boulton, M., & Smith, P. (1994). Bully/victim problems in middle school children: stability, self-perceived competence, peer perceptions and peer acceptance. *British Journal of Developmental Psychology, 12*, 315-329.

Boyarski, D., Neuwirth, C., Forlizzi, J., & Regli, S. H. (1998). *A study of fonts designed for screen display.* Proceedings of CHI' 98, 87-94.

Breland, H. M. (1999). *Exploration of an automated editing task as a GRE writing measure.* (GRE Board Professional Report No. 96-01R). Princeton, NJ: Educational Testing Service.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2002). *Effects of screen size, screen resolution, and display rate on computer-based test performance.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Buchanan, T. (2003). Internet-based questionnaire assessment: appropriate use in

clinical contexts. *Cognitive Behaviour Therapy, 32*, 100–109.

Buchanan, T., & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology, 90*, 125-144.

Buchanan, T., Johnson, J. A., & Goldberg, L. R. (2005). Implementing a five-factor personality inventory for use on the Internet. *European Journal of Psychological Assessment, 21*, 116-128.

Byrne, J. (2004). *Accessible web typology – an introduction for web designers.* [Electronic version]. Retrieved May 7, 2006, from http://www.scotconnect.com/webtypography/

Cairns, E., McWhirter, L., Duffy, U., & Barry, R. (1990). The stability of self-concept in late adolescence: Gender and situational effects. *Personality and Individual Differences, 11*, 937-944.

Callaghan, S., & Joseph, S. (1995). Self-concept and peer victimization among schoolchildren. *Personality and Individual Differences, 18*, 161-163.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 629-637.

Chernyshenko, O. S. (2002). *Applications of ideal point approaches to scale constructions and scoring in personality measurement: The development of a six-faceted measure of conscientiousness.* Unpublished manuscript, University of Illinois at Urbana-Champaign.

Chernyshenko, O. S., Stark, S., Chan, K.-Y., Drasgow, F., & Williams, B. (2001).

Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*, 563–562.

Chernyshenko, O. S., Stark, S., Chan, K-Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*, 523-562.

Chin, W.W. (1998). *The Partial Least Squares Approach for Structural Equation Modeling*. Lawrence Erlbaum Associates, 295-336.

Choi, S. W., & Tinkler, T. (2002). *Evaluating comparability of paper-and-pencil and computer-based assessment in a K-12 setting.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Chua, S. L., Chen, D. T., & Wong, A. F. L. (1999). Computer anxiety and its correlates: A meta-analysis. *Computers in Human Behavior, 15*, 609-623.

Clariana, R. B., & Wallace, P. E. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology, 33*, 595-904.

Cliff, N., Collins, L. M., Zatkin, J., Gallipeau, D., & McCormick, D. J. (1988). An ordinal scaling method for questionnaire and other ordinal data. *Applied Psychological Measurement, 12*, 83–97.

Conway, M. A., Gardiner, J. M., Perfect, T. J., Anderson, S. J., & Cohen, G., (1997). Changes in memory awareness during learning: The acquisition of knowledge by psychology undergraduates. *Journal of Experimental*

Psychology: General, 126, 393-413.

Coombs, C. H. (1964). *A theory of data.* New York: Wiley.

Covington, M. V. (1992). *Making the grade: A self-worth perspective on motivation and school reform*. Cambridge: Cambridge University Press.

Coyne, I., & Bartram, D. (2006). Design and development of the ITC guidelines on computer-based and Internet-delivered testing. *International Journal of Testing, 6*, 133-142.

Cronbach, L. J. (1949). *Essentials of psychological testing*. New York: Halpern.

Davis, R. N. (1999). Web-based administration of a personality questionnaire: Comparison with traditional method. *Behavior Research Methods, Instruments, & Computers, 31,* 572-577.

de Bruijn, D., de Mul, S., & van Oostendorp, H. (1992). The influence of screen size and text layout on the study of text. *Behaviour and Information Technology, 11*, 71-78.

de Rossi, L.C. (2002, January). What are "Serif" and "Sans-serif" fonts? *MasterView International, 8*. Retrieved June 15, 2004, from http://masterview.ikonosnewmedia.com002F2002/01/15/what_are_serif_and_sansserif.htm

DeGree, C. E., & Snyder, C. R. (1985). Adler's psychology (of use) today: Personal history of traumatic life events as a self-handicapping strategy. *Journal of Personality and Social Psychology, 48*, 1512-1519.

Desarbo, W. S., & Hoffman, D. L. (1986). Simple and weighted unfolding threshold

models for the spatial representation of binary choice data. *Applied Psychological Measurement, 10*, 247–264.

Desmarais, L. B., Dyer, P. J., Midkiff, K. R., Barbera, K. M., Curtis, J. R., Esrig, F. H., & Masi, D. L. (1992, May). *Scientific uncertainties in the development of a multimedia test: Trade-offs and decisions.* Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Montreal Quebec.

Desmarais, L. B., Masi, D. L., Olson, M. J., Barbera, K. M., & Dyer, P. J. (1994, April). *Scoring a multimedia situational judgment test: IBM's experience.* Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Nashville, TN.

Dillon, A. (1992). Reading from paper versus screens: A critical review of the empirical literature. *Ergonomics, 35*, 1297-1326.

Dillon, A. (1994). *Designing usable electronic text*. London: Taylor & Francis.

Dimock, P. H., & Cormier, P. (1991). The effects of format differences and computer experience on performance and anxiety on a computer-administered test. *Measurement & Evaluation in Counseling & Development, 24*, 119-126.

Donoghue, J. R. (1994). An empirical investigation of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement*, *31*, 295–311.

Doran, H. C. (2005). The information function for the one-parameter logistic model: Is it reliability? *Educational and Psychological Measurement, 65*, 759-769.

Drasgow, F., & Mattern, K. (2006). New tests and new items: Opportunities and issues. In D. Bartram & R. Hambleton (Eds.), *Computer-based testing and the Internet: Issues and advances*. London: Wiley.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B. A., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19*, 143-165.

Drasgow, F., Olson-Buchanan, J. B., & Moberg, P. J. (1999). Development of an interactive video assessment: Trials and tribulations. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 177-196). Mahwah, NJ: Lawrence Erlbaum Associates.

Duchnicky, R. L., & Kolers, P. A. (1983). Readability of text scrolled on a visual display terminal as a function of window size. *Human Factors, 25*, 683-692.

Durndell, A., & Lightbody, P. (1994). Gender and computing: Change over time? *Computers & Education, 21*, 331-336.

Dyer, P. J., Desmarais, L. B., Midkiff, K. R., Colihan, J. P., & Olson, J. B. (1992, May) *Designing a multimedia test: Understanding the organizational charge, building team, and making the basic research commitments.* Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Montreal, Quebec.

Dyson, M. C., & Haselgrove, M. (2001). The influence of reading speed and line length on the effectiveness of reading from screen. *International Journal of Human-Computer Studies, 54*, 585-612.

Dyson, M. C., & Kipping, G.J. (1998). The effects of line length and method of

movement on patterns of reading from screen. *Visible Language, 32*, 150-181.

Eaves, R.C., & Smith, E. (1986). The effect of media and amount of microcomputer experience on examination scores. *Journal of Experimental Education, 55*, 23-26.

Ellis, B. B., Becker, P., & Kimmel, H. D. (1993). An item response theory evaluation of an English version of the Trier Personality Inventory (TPI). *Journal of Cross-Cultural Psychology, 24,* 133-148.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272-299.

Feinberg, S., & Murphy, M. (2000). *Applying cognitive load theory to the design of Web-based instruction.* Paper presented at the IEEE Professional Communication Society International Professional Communication Conference, September 24-27, 2000, Cambridge, MA.

Fisher, R.A. (1932). *Statistical methods for research workers* (5th ed.). Edinburgh, UK: Oliver & Boyd.

Fletcher, R. B., & Hattie, J. A. (2004). An examination of the psychometric properties of the physical self-description questionnaire using a polytomous item response model. *Psychology of Sport and Exercise, 5*, 423-466.

Foster, J. D., Campbell, W. K., & Twenge, J. M. (2003). Individual differences in narcissism: Inflated self-views across the lifespan and around the world. *Journal of Research in Personality, 37*, 469-486.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item-response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology, 78*, 350-365.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2000). *The effect of computer-based tests on racial/ethnic, gender, and language groups.* (GRE Board Professional Report No. 96-21P). Princeton, NJ: Education Testing Service.

Garcia, T. and Pintrich, P. R., (1994). Regulating motivation and cognition in the classroom: The role of self-schemas and self-regulatory strategies. In D. H. Schunk and B. J. Zimmerman (Eds.), *Self-regulation of learning and performance: Issues and educational applications* (pp. 127–153). Lawrence Erlbaum Associates, Hillsdale, NJ.

Garcia, T., Matula, J. S., Harris, C. L., Egan-Dowdy, K., Lissi, M. R., & Davila, C. (1995). *Worriers and procrastinators: Differences in motivation, cognitive engagement, and achievement between defensive pessimists and self-handicappers*. Paper presented at Annual American Educational Research Association Conference, San Francisco, CA.

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53*, 525-546.

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. J. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe Vol. 7* (pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

Goldsmith, R. E. (1986). Dimensionality of the Rosenberg Self-Esteem Scale.

Journal of Social Behavior and Personality, 1, 253-264.

Gould, J. D., & Grischkowsky, N. (1984). Doing the same work with hard copy and cathode-ray tube (CRT) computer terminals. *Human Factors, 26*, 323-337.

Gould, J. D., Alfaro, L., Finn, R., Haupt, B., & Minuto, A. (1987). Reading from CRT displays can be as fast as reading from paper. *Human Factors, 29*, 497-517.

Greaud, V., & Green, B. F. (1986). Equivalence of conventional and computer presentation of speed tests. *Applied Psychological Measurement, 10*, 23-34.

Habing, B., Finch, H., & Roberts, J. S. (2005). A Q3 statistic for unfolding item response theory models: Assessment of unidimensionality with two factors and simple structure. *Applied Psychological Measurement, 29*, 457-471.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis (*5th ed.). Upper Saddle River, NJ: Prentice-Hall.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 147-200). New York: Macmillan.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Harmes, J. C. (1999). *Computer-based testing: Towards the design and use of innovative items.* Retrieved February, 23, 2006, from http://www.coedu.usf.edu/itphdsem/eme7938/ch899.pdf

Harpster, G. L., Freivalds, A., Shulman, G. L., & Leibowitz, H. W. (1989). Visual

performance on CRT screens and hard-copy displays. *Human Factors, 31*, 247-257.

Hart, S. G. and Straveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of experimental and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139-183). Amsterdam: North-Holland.

Harter, S. (1988). *Manual for the self-perception profile for adolescence*. Denver, CO: University of Denver.

Harvey, R. J., & Murry, W. D. (1994). Scoring the Myers-Briggs Type Indicator: Empirical comparison of preference score versus latent-trait methods. *Journal of Personality Assessment, 62*, 116-129.

Henk, W. A., & Melnick, S. A. (1995). The Reader Self Perception Scale RSPS: A new tool for measuring how children feel about themselves as readers. *The Reading Teacher, 48*, 470-482.

Hensley, V. R. (1988). Australian normative study of the Achenbach Child Behavior Checklist. *Australian Psychologist, 23*, 371-382.

Hirt, E. R., Deppe, R. K., & Gordon, L. J. (1991). Self-reported versus behavioral self-handicapping: Empirical evidence for a theoretical distinction. *Journal of Personality and Social Psychology, 61*, 981-991.

Hoare, P., & Mann, H. (1994). Self-esteem and behavioural adjustment in children with epilepsy and children with diabetes. *Journal of Psychosomatic Research, 38*, 859-869.

Hofer, P. J., & Green, B. F. (1985). The challenge of competence and creativity in computerized psychological testing. *Journal of Consulting and Clinical Psychology, 53*, 826-838.

Hoge, R. D., & McScheffrey, R. (1992). Performance within an enriched program for the gifted. *Child Study Journal, 22*, 93-102.

Hoijtink, H. (1990). A latent trait model for dichotomous choice data. *Psychometrika, 55*, 641–656.

Hoijtink, H. (1991). The measurement of latent traits by proximity items. *Applied Psychological Measurement, 15*, 153–169.

Horan, P. M., DiStefano, C., & Motl, R. W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling, 10*, 435-455.

Horton, W. (1989). *Designing and writing online documentation: Help files to hypertext*. New York: John Wiley & Sons.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle, (Ed.), *Structural equation modeling: Comments, issues, and applications* (pp. 158–176). Thousand Oaks, CA: Sage.

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement, 40,* 1-15.

Johnson, J.A. (2000, March). *Web-based personality assessment*. Poster session presented at the 71st Annual Meeting of the Eastern Psychological Association, Baltimore, MD.

Jones, E. E., & Berglas, S. (1978). Control of attributions about the self through self-handicapping strategies: The appeal of alcohol and the role of underachievement. *Personality and Social Psychology Bulletin, 4*, 200-206.

Jones, J. P. (2000). Promoting stakeholder acceptance of CBT. *Journal of Testing Technologies, 1.* Retrieved January, 8, 2006, from http://www.testpublishers.org/Documents/journal02.pdf.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Kak, A. V. (1981). Relationships between readability of printed and CRT-displayed text. *Proceedings of Human Factors Society 25th Annual Meeting* (pp. 137-140). Human Factors Society Santa Monica, CA.

Keenan, P. A., & Olson, J. B. (1991, April). *A model-based multi-media technique for assessing conflict management skills.* Paper presented at the annual conference of the Society for Industrial and Organizational Psychology Conference, St. Louis, MO.

Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. (1998). *Computer familiarity among TOEFL examinees.* (TOEFL Research Report 59). Princeton, NJ: Educational Testing Service.

Kolers, P., Duchnicky, R. L., & Ferguson, D. C. (1981). Eye movement of readability of CRT displays. *Human Factors, 23*, 517-527.

Kruk, R. S. & Muter, P. (1984). Reading of continuous text on video screens. *Human Factors, 26*, 339-345.

Lee, J. (1986). The effects of past computer experience on computer aptitude test performance. *Educational and Psychological Measurement, 46*, 727-733.

Levesque, M. J., Lowe, C. A., & Mendenhall, C. (2001). Self-handicapping as a method of self-presentation: An analysis of costs and benefits. *Current Research in Social Psychology, 6*, 221–237.

Levine, M. V. (1984). *An introduction to multilinear formula score theory* (Measurement Series 84–4). Champaign: University of Illinois, Model Based Measurement Laboratory.

Levine, T., & Donitsa-Schmidt, S. (1998). Computer use, confidence, attitudes, and knowledge: A causal analysis. *Computers in Human Behavior, 14*, 125-146.

Lievens, F. and Harris, M. M. (2003). Research on Internet recruiting and testing: Current status and future directions. In, C. L. Cooper &, I. T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology, Vol. 18*. (pp. 131–165). Chichester, UK: John Wiley,

Lonsdale, M. S., Dyson, M. C., & Reynolds, L. (2006). Reading in examination-type situations: The effects of text layout on performance. *Journal of Research in Reading, 29*, 433-453.

Lord, F. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14*, 117-138.

Lovelace, E. A., & Southall, S. D. (1983). Memory for words in prose and their locations on the page. *Memory and Cognition, 11*, 429-434.

Luecht, R. M., Hadadi, A., Swanson, D. B., & Case, S. M. (1998). Testing the test:

A comparative study of a comprehensive basic science test using paper-and-pencil and computerized formats. *Academic Medicine, 73*, 51-53.

Lynch, P. J., & Horton, S. (2002). *Web style guide: Basic design principles for creating web sites* (2nd ed.). New Haven, CT: Yale University Press.

Lynch, R. (2000). Computer-based testing: The test of English as a foreign language (TOEFL). *The Source*, Fall 2000. Retrieved January, 13, 2006, from http://www.usc.edu/dept/education/TheSource/>Fall2000.

Marsh, H. W. (1986). The bias of negatively worded items in rating scales for young children: A cognitive-developmental phenomenon. *Developmental Psychology, 22*, 37-49.

Marsh, H. W., & Holmes, I. W. M. (1990). Multidimensional self-concepts: Construct validation of responses by children. *American Educational Research Journal, 27*, 87-117.

Marsh, H. W., Martin, A. J., & Debus, R. (2001). Individual differences in verbal and math self-perceptions: One factor, two factors, or does it depend on the construct. In R. Riding & S. Rayner (Eds.), *International perspectives on individual differences*. London: Greenwood Publishing.

Marsh, H. W., Richards, G. E., Johnson, S., Roche, L., & Tremayne, P. (1994). Physical self-description questionnaire: Psychometric properties and a multitrait-multimethod analysis of relations with existing instruments. *Journal of Sport and Exercise Psychology, 15*, 270-305.

Martin, A. J. (1998). *Self-handicapping and defensive pessimism: Predictors and consequences from a self-worth motivation perspective.* Unpublished doctoral

dissertation. University of Western Sydney, Macarthur.

Martin, A. J., Marsh, H. W., & Debus, R. L. (2001). Self-handicapping and defensive pessimism: Exploring a model of predictors and outcomes from a self-protection perspective. *Journal of Educational Psychology, 93*, 87–102.

Martinez, M. E. & Jenkins, J. B. (1993). *Figural response assessment: System development and pilot research in cell and molecular biology.* GRE Board Professional Report No. 89-02; ETS Research Report 92-50. Princeton, NJ: Educational Testing Service.

Martinez, M. E. (1993). Problem-solving correlates of new assessment forms in architecture. *Applied Measurement in Education, 6*, 167-180.

Mathewson, G, C. (1994). Model of attitude influence upon reading and learning to read. In H. Singer, R. B. Ruddell, & M. I. Ruddell (Eds.), *Theoretical models and processes of reading* (4th ed., pp. 1131-1161). Newark, DE: International Reading Association.

Maydeu-Olivares, A. (2005). Further empirical results on parametric versus non-parametric IRT modeling of Likert-type personality data. *Multivariate Behavioral Research, 40*, 261-279.

Mayes, D. K., Sims, V. K., & Koonce, J. M. (2001). Comprehension and workload differences for VDT and paper-and-pencil reading. *International Journal of Industrial Ergonomics, 28*, 367-378.

McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers &*

Education, 39, 299-312.

McHenry, J. J., & Schmitt, N. (1994). Multimedia testing. In M. G. Rumsey & C. B. Walker (Eds.), *Personnel selection and classification* (pp. 193-232). Hillsdale, NJ: Lawrence Erlbaum Associates.

McMullin, J., Varnhagen, C. K., Heng, P., & Apedoe, X. (2002). Effects of surrounding information and line length on text comprehension from the web. *Canadian Journal of Learning and Technology, 28*, 19-29.

Mead, A. D., & Coussons-Read, M. (2002, April). *The equivalence of paper- and web-based version of the 16PF Questionnaire.* Paper presented at the annual Conference of the Society for Industrial and Organizational Psychology, Toronto, Ontario, Canada.

Mead, A. D., & Drasgow, G. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449-458.

Meade, A. W., Michels, L. C., & Lautenschlager, G. J. (2007). Are internet and paper-and-pencil personality tests truly comparable? An experimental design measurement invariance study. *Organizational Research Methods, 10*, 322-345.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*, 108-141.

Meijer, S. A., Sinnema, G., Bijstra, J. O., Mellenbergh, G. J, & Wolters, W. H. G. (2000). Social functioning in children with a chronic illness. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 41*, 309-317.

Michell, J. (1994). Measuring dimensions of belief by unidimensional unfolding. Journal of Mathematical Psychology 38, 244-273.

Microsoft Corporation. (2003, September). *Microsoft Certified Professional (MCP) Exam Demos*. Retrieved February, 14, 2007, from http://www.microsoft.com/downloads/search.aspx?displaylang=en.

Midgley, C., & Urdan, T. (1995). Predictors of middle school students' use of self-handicapping strategies. *Journal of Early Adolescence, 15*, 389-411.

Midgley, C., Arunkumar, R., Urdan, T. (1996). "If I don't do well tomorrow, there's a reason": Predictors of adolescents' use of academic self-handicapping strategies. *Journal of Educational Psychology, 88*, 423-434.

Midkiff, K. R., Dyer, P. J., Desmarais, L. B., Rogg, K., & McCusker, C. R. (1992, May). *The multimedia test: Friend or foe?* Paper presented at the annual conference of the Society for Industrial and Organizational Psychology Conference, Montreal, Quebec.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3* (2nd ed.). Mooresville IN: Scientific Software Inc.

Morrison, R. E., & Inhoff, A. W. (1981). Visual factors and eye movements in reading. *Visible Language, 15*, 129-146.

Mullet, K., & Sano, D. (1995). *Designing visual interfaces: Communication oriented techniques*. New York: Prentice Hall.

Muter, P. (1996). Interface design and optimization of reading of continuous text. In H. van Oostendorp, & S. de Mul (Eds.), *Cognitive aspects of electronic text*

processing (pp. 161-180). Norwood, NJ: Ablex.

Muter, P., & Maurutto, P. (1991). Reading and skimming from computer screen and books: The paperless office revisited? *Behavior & Information Technology, 10*, 257-266.

Muter, P., Latremouille, S. A., Treurniet, W. C., & Beam, P. (1982). Extended reading of continuous text on television screens. *Human Factors, 24*, 501-508.

Norem, J. K., & Cantor, N. (1986). Defensive pessimism: Harnessing anxiety as motivation. *Journal of Personality and Social Psychology, 51*, 1208–1217.

Norem, J. K., & Illingworth, K. S. S. (1993). Strategy-dependent effects of reflecting on self and tasks: Some implications of optimism and defensive pessimism. *Journal of Personality and Social Psychology, 65*, 822–835.

Noyes, J. M., & Garland, K. J. (2003). VDT versus paper-based text: Reply to Mayes, Sims, and Koonce. *International Journal of Industrial Ergonomics, 31*, 411-423.

Noyes, J. M., Garland, K. J., & Robbins, E. (2004). Paper-based versus computer-based assessment – Is workload another test mode effect? *British Journal of Educational Technology, 35*, 111-113.

O'Neill, K., & Folk, V. (1996, April). *Innovative CBT item formats in a teacher licensing program*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Olson-Buchanan, J. B., & Drasgow, F. (1999). Beyond bells and whistles: An introduction to computerized assessment. In F. Drasgow & J. D. Olson-

Buchanan (Eds.), *Innovations in computerized assessment* (pp. 1-5). Hillsdale, NJ: Erlbaum.

Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., Keenan, P. A., & Donovan, M. A. (1998). Interactive video assessment of conflict resolution skills. *Personnel Psychology, 51,* 1-24.

Parshall, C. G., & Kromrey, J. D. (1993, April). *Computer testing versus paper-and-pencil testing: An analysis of examinee characteristics associated with mode effect.* Paper presented at the annual Meeting of the American Educational Research Association. Atlanta, GA.

Parshall, C. G., Davey, T., & Pashley, P. J. (2000). Innovative item types for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice*, (pp. 129-148). Dordrecht, The Netherlands: Kluwer.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer-Verlag.

Pasveer, K. A., & Ellard, J. H. (1998). The making of a personality inventory: Help from the WWW. *Behavior Research Methods, Instruments & Computers, 30*, 309-313.

Pettit, F. A. (2002). A comparison of World-Wide Web and paper-and-pencil personality questionnaires. *Behavior research Methods, Instruments, & Computers, 34*, 50-54.

Piolat, A., Roussey, J. Y., & Thunin, O. (1997). Effects of screen presented text on reading and revising. *International Journal of Human-Computer Studies, 47*,

565-589.

Ployhart, R. E., Weekley, J. A., Holtz, B. C. and Kemp, C. (2003). Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, bio data, and situational judgment tests comparable? *Personnel Psychology, 56*, 733-752.

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. F. (2002, April). *Web-based vs. paper-and-pencil testing: A comparison of factor structures across applicants and incumbents*. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Toronto, CA.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper-and-pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment, 3*. Available from http://www.jtla.org

Potosky, D., & Bobko, P. (2004). Selection testing via the internet: Practical considerations and exploratory empirical findings. *Personnel Psychology, 57*, 1003-1034.

Powers, D. E., & O'Neill, K. (1992). *Inexperienced and anxious computer users: Coping with a computer-administered test of academic skills.* The Praxis Series: Professional assessments for beginning teachers. Princeton, NJ: Educational Testing Service.

Rabinowitz, S., & Brandt, T. (2001). *Computer-based assessment: Can it deliver on it's promise? WestEd Knowledge Brief*. Retrieved January, 3, 2004, from http://www.wested.org.

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading*. Englewood Cliffs, N.J.: Prentice Hall.

Reise, S. P. (1999). Personality measurement issues viewed through the eyes of IRT. In S. E. Embretson & S. Hershberger (Eds.), *The new rules of measurement: What every psychologist should know* (pp. 219-241). Mahwah, NJ: Erlbaum.

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7*, 347-364.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14*, 45-58.

Revuelta, J., Ximénez, M. C., & Olea, J. (2003). Psychometric and psychological effects of item selection and review on computerized testing. *Educational and Psychological Measurement, 63*, 791-808.

Rhodewalt, F., & Davison, J. (1986). Self-handicapping and subsequent performance: Role of outcome valence and attributional certainty. *Basic and Applied Social Psychology, 7*, 307-323.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*, 3-32.

Roberts, J. S., Fang, H-R, Cui, W., & Wang, Y. (2006). GGUM2004: A windows-based program to estimate parameters in the generalized graded unfolding model. *Applied Psychological Measurement, 30*, 64-65.

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the Likert

and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement, 59,* 211-233.

Robins, R. W., Trzesniewski, K. H., Tracy, J. L., Gosling, S. D., & Potter, J. (2002). Self-esteem across the lifespan. *Psychology and Aging, 17*, 423-434.

Rouse, S. V., Finger, M. S., & Butcher, J. N. (1999). Advances in clinical personality measurement: An Item Response Theory analysis of the MMPI-2 Psy-5 scales. *Journal of Personality Assessment, 72*, 282-307.

Rowe, K. S., & Rowe, K. J. (1997). Norms for parental ratings on Conners' Abbreviated Parent-Teacher Questionnaire: Implications for the design of behavioral rating inventories and analyses of data derived from them. *Journal of Abnormal Child Psychology, 25*, 425-451.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7*. Retrieved February, 23, 2006, from http://epaa.asu.edu/epaa/v7n20/

Salgado, J. F. (2003) Predicting job performance using FFM and non-FFM personality measures. *Journal of Occupational and Organizational Psychology, 76*, 323–346.

Salgado, J. F., & Moscoso, S. (2003). Internet-based personality testing: Equivalence of measures and assesses' perceptions and reactions. *International Journal of Selection and Assessment, 11*, 194-205.

Samejima, F. (1969). Calibration of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 17*.

Samejima, F. (1977). A method of estimating item characteristic functions using the maximum likelihood estimator of ability. *Psychometrika*, *42*, 163–192.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18*, 229–244.

Sandoval, J. (1977). The measurement of the hyperactive syndrome in children. *Review of Education Research, 47*, 293-318.

Sandoval, J. (1981). Format effects in two teacher rating scales of hyperactivity. *Journal of Abnormal Child Psychology, 9*, 203-218.

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the life orientation test. *Journal of Personality and Social Psychology, 67*, 1063-1078.

Scherbaum, C.A., Finlinson, S., Barden, K., & Tamanini, K. (2006). Applications of item response theory to measurement issues in leadership research. *Leadership Quarterly, 17*, 366-386.

Schumann, B. C., Striegel-Moore, R. H., McMahon, R. P., Waclaawiw, M. A., Morrison, J. A., & Schreiber, G. B. (1999). Psychometric properties of the self-perception profile for children in a biracial cohort of adolescent girls: The NHLBI growth and health study. *Journal of Personality Assessment, 73*, 260-275.

Shotland, A., Alliger, G. M., & Sales, T. (1998). Face validity in the context of personnel selection: A multimedia approach. *International Journal of Selection*

and Assessment, 6, 124-130.

Smedshammar, H., Frenckner, K., Nordquist, C., & Romberger, S. (1989). *Why is the difference in reading speed when reading from VDUs and from paper bigger for fast readers than for slow readers?* Paper presented at the WWDU 1989, Second International Scientific Conference, Montreal, CA.

Smith, A., & Savory, M. (1989). Effects and after-effects of working at a VDU: Investigation of the influence of personal variables. In E.D. Megaw (Ed.), *Contemporary ergonomics* (pp. 252–257). London: Taylor & Francis.

Smith, T. W., Snyder, C. R., & Perkins, S. C. (1983). The self-serving function of hypochondriacal complaints: Physical symptoms as self-handicapping strategies. *Journal of Personality & Social Psychology, 44*, 787-797.

Spencer, H. (1968). *The visible word*. London: Royal College of Art.

Spool, J., Scanlon, T., Schroeder, W., Snyder, C., & De Angelo, T. (1997). *Web site usability: A designer's guide*. North Andover, MA: User Interface Engineering.

Spray, J. A., Ackerman, T. A., Reckase, M. D., & Carlson, J. E. (1989). Effect of the medium of item presentation on examinee performance and item characteristics. *Journal of Educational Measurement, 26*, 261-271.

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology, 84*, 1041-1053.

Stark, S. (2002). *MODFIT* [Computer program]

Stark, S., Chernyshenko, S., Chuah, D., Lee, W., & Wadlington, P. (2001). *Selecting*

a dichotomous IRT model. [On-line tutorial]. Available: http://work.psych.uiuc.edu/irt/modeling_dich1.asp

Steiger, J.H., & Lind, J.C. (1980). *Statistically-based tests for the number of common factors.* Paper presented at the annual spring meeting of the Psychometric Society in Iowa city. May 30, 1980.

Steinberg, L., & Thissen, D. (1995). Item response theory in personality research. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A Festschrift honouring Donald W. Fiske* (pp. 161-182). Hillsdale, NJ: Erlbaum.

Sternberg, R. J., & Kaufman J. C. (1998). Human abilities. *Annual Review of Psychology, 49*, 479-502.

Strube, M. J. (1986). An analysis of the self-handicapping scale. *Basic and Applied Social Psychology, 7*, 211-224.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 185-233.

Taylor, C., Kirsch, I., Eignor, D., & Jamieson, J. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks.* (Research Reports 61). Princeton, NJ: Educational Testing Service.

Taylor, C., Kirsch, I., Eignor, D., & Jamieson, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning, 49*, 219-274.

Thissen, D. (1976). Information in wrong responses to the Raven Progressive Matrices. *Journal of Educational Measurement, 13*, 201–214.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567–577.

Thissen, D., Chen, W.-H., & Bock, D. (2003). *Multilog for Windows (Version 7.0)* [Computer Software]. Lincolnwood, IL: Scientific Software International.

Thompson, T. (1994). Self-worth protection: Review and implications for the classroom. *Educational Review, 46*, 259–274.

Thompson, T. (1994). Self-worth protection: Review and implications for the classroom. *Educational Review, 46*, 259-275.

Thurstone, L. L. (1928). Attitudes can be measured. American Journal of Sociology, 23, 529-554.

Thurstone, L. L. (1931). Measurement of social attitudes. *Journal of Abnormal and Social Psychology, 26*, 249-269.

Tinker, M. (1963). *Legibility of print*. Ames, IA: Iowa State University Press.

Todman, J., & Lawrenson, H. (1992). Computer anxiety in primary school children and university students. *British Educational Research Journal, 18*, 63-72.

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.

Trent, L. M. Y., Russell, G., & Cooney, G. (1994). Assessment of self-concept in early adolescence. Australian Journal of Psychology, 46, 21-28.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1–10.

Tullis, T. S., Boynton, J. L., & Hersh, H. (1995). Readability of fonts in the Windows environment. *CHI 95 Extended Abstracts*, 127-128.

Van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47*, 123-140.

van Nes, F. (1986). Space, colour and typography of visual display terminals. *Behaviour and Information Technology, 5*, 99-118.

van Schuur, W. H. (1984). *Structure in political beliefs: A new model for stochastic unfolding with application to European party activists*. Amsterdam: CT Press.

Veerman, J. W., tenBrink, L. T., Straathof, M. A. E., & Treffers, P. D. A. (1996). Measuring children's self-concept with a Dutch version of the "self-perception profile for children": factorial validity and invariance across a nonclinic and a clinic group. *Journal of Personality Assessment, 67*, 142-154.

Verhelst, N. D., & Verstralen, H. H. F. M. (1993). A stochastic unfolding model derived from the partial credit model. *Kwantitativea Methoden, 42*, 73–92.

Vispoel, W. P. (1998). Reviewing and changing answers on computer-adaptive and self-adaptive vocabulary tests. *Journal of Educational Measurement, 35*, 328-345.

Vispoel, W. P. (2000). Reviewing and changing answers on computerized fixed-item vocabulary tests. *Educational and Psychological Measurement, 60*, 371-384.

Wakeman, L. (2000). *Fonts and accessibility in web pages*. Retrieved March, 11, 2006, from http://lois.co.uk/web/articles/font-access.shtml.

Waller, N. G., Tellegen, A., McDonald, R. P., & Lykken, D. T. (1996). Exploring nonlinear models in personality assessment: Development and preliminary validation of a negative emotionality scale. *Journal of Personality, 64*, 545-576.

Watson, B. (2001). Key factors affecting conceptual gains from CAL material. *British Journal of Educational Technology, 32*, 587–593.

Wichstrøm, L. (1995). Harter's self-perception profile for adolescents: Reliability, validity, and evaluation of the question format. *Journal of Personality Assessment, 65*, 100-116.

Wise, S. L., & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica, 21*, 135-155.

Woolhouse, L., & Myers, S. (1999). *Factors affecting sample make-up: Results from an Internet-based personality questionnaire*. Paper presented at the 1999 British Psychological Society Social Psychology Section Conference.

Wright, P., & Lickorish, A. (1983). Proof-reading: VDU and paper text compared for speed, accuracy, and fatigue. *Behavior & Information Technology, 2*, 227-235.

Yost, P. R., & Homer, L. E. (1998, April). *Electronic versus paper surveys: Does the medium affect the response?* Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Dallas, TX.

Zenisky, A. L., & Sireci, S. (2001). *Feasibility review of selected performance assessment for the computerized Uniform CPA Exam* (Laboratory of

Psychometric and Rep. No. 405). Amherst: School of Education, University of Massachusetts.

Zenisky, A.L., & Sireci, S.G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*, 337-362.

Zickar, M. J. (2001). Conquering the next frontier: Modeling personality data with item response theory. In B. Roberts & R. Hogan (Eds.), Applied personality psychology: The intersection of personality and I/O psychology (pp. 141-158). Washington, DC: American Psychological Association.

Ziefle, M. (1998). Effects of display resolution on visual performance. *Human Factors, 40*, 554-568.

Zumbo, B. D., Gelin, M. N., & Hubley, A. M. (2001, February). *Psychometric study of the CES-D: Factor analysis and DIF*. 29th annual meeting of the International Neuropsychological Society (INS), Chicago, IL, U.S.A.