EFFECTS OF ITEM PRESENTATION AND USER FLEXIBILITY ON RATER PERCEPTIONS

by

RYAN EDWARD PHILLIPS

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN PSYCHOLOGY

THE UNIVERSITY OF TEXAS AT ARLINGTON

August 2008

ACKNOWLEDGEMENTS

The successful completion of my thesis requirements and the master's degree program would not have been possible without the guidance and support of several individuals. First and foremost, I would like to express my sincere gratitude and appreciation to the members of my thesis committee. Many thanks are owed to my graduate advisor and committee chair, Dr. Mark Frame, for his creative insight, flexibility, and expertise. I am truly thankful for your dedication to helping students succeed. The contributions of my other committee members, Dr. Nicolette Lopez and Dr. Jared Kenworthy, were also invaluable throughout the entire process.

I also wish to thank my friends within the graduate program for their fresh perspectives and encouragement throughout this process, as well as my friends and family outside the bubble of graduate school who have been equally supportive of my efforts over the last few years. I would like to thank my friends at Leadership Worth Following, LLC for their constant dedication to helping me succeed. To my parents and grandparents, your support and encouragement made this journey possible. My greatest thanks go to my wife, Renee, who sacrificed so much during this challenging time. I could never have made it this far without your unwavering love, patience, and understanding.

July 16, 2008

ABSTRACT

EFFECTS OF ITEM PRESENTATION AND USER FLEXIBILITY ON RATER PERCEPTIONS

Ryan Edward Phillips, MS

The University of Texas at Arlington, 2008

Supervising Professor: Mark C. Frame

The purpose of this study was to examine the effects of item presentation and user flexibility on performance ratings and rater perceptions of the performance rating process. The participants were 252 undergraduate students enrolled in one or more psychology courses at the University of Texas at Arlington. Participants rated the performance of their psychology instructor using four different rating formats: (a) Multiple-item presentation, high user-flexibility, (b) Multiple-item presentation, low user-flexibility, (c) Single-item presentation, high user-flexibility, and (d) Single-item presentation, low user-flexibility. Dependent measures were administered within each performance rating format to obtain ratings of (a) instructor performance, (b) rater confidence in the accuracy of performance ratings, and (c) rater satisfaction with the performance rating process. At the end of the study, participants were asked to choose their most preferred rating format. Item presentation and user flexibility did not significantly affect performance ratings or rater perceptions; however rater preference were found to significantly differ across the four rating formats. The proportion of raters selecting the multiple-item presentation, high user-flexibility rating format was significantly higher than the

other three rating formats, and the proportion of raters selecting the single-item presentation, low user-flexibility rating format was significantly lower than the other three rating formats.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS…… iii

ABSTRACT…… iv

LIST OF ILLUSTRATIONS…….. viii

LIST OF TABLES……... ix

Chapter Page

1. INTRODUCTION..... 1

1.1 Impact of Computers on Psychometric Theory..... 2

1.2 The Mode Effect……..... 4

1.3 Computer-Adaptive Measures..... 5

1.4 Mode Effects Associated with User Flexibility…..... 8

1.5 Mode Effects Associated with Item Presentation..... 10

1.6 Integrated Analysis of the Mode Effects Associated with CAMs..... 11

1.7 Computer-Adaptive Rating Scales..... 12

1.8 Study Objectives..... 12

1.9 Study Hypotheses..... 13

1.9.1 Hypothesis One..... 13

1.9.2 Hypothesis Two..... 13

2. METHODS..... 14

2.1 Experimental Design..... 14

2.2 Participants..... 15

2.3 Measures..... 15

2.3.1 Student Evaluation of Instructor Competence..... 16

2.3.2 User Evaluation of the Performance Rating Process…..… 16

2.3.3 Supplemental Demographic Questionnaire.....…..… 17

2.4 Procedure.....…….…….. 17

2.5 Exclusion Criteria.....…….…….. 19

2.6 Preliminary Data Analysis.....…….…….. 19

2.7 Hypothesis Testing.....…….…….. 20

2.8 Exploratory Analysis.....…….…….. 20

3. RESULTS….....…….…… 22

3.1 Demographic Characteristics.....…… 22

3.2 Preliminary Data Analysis….....…… 23

3.3 Hypothesis Testing.....…… 24

3.4 Exploratory Analysis.....…… 25

4. DISCUSSION.....…….. 27

4.1 Limitations of the Study.....…… 28

4.2 Implications.....…… 29

4.3 Future Research Directions.....…… 30

4.4 Concluding Comments.....…… 31

APPENDIX

A. FIGURES.....…….…… 32

B. TABLES.....…….…… 39

C. DEPENDENT MEASURES.....…….…… 44

REFERENCES.....…….…… 51

BIOGRAPHICAL INFORMATION.……..…… 55

LIST OF ILLUSTRATIONS

Figure		Page
A.1	Structural Design of the Doubly-Multivariate Analysis of Variance.....	33
A.2	Sequence of Rating Formats across the Four Rating Conditions.....…….	34
A.3	Profiles of Performance Ratings.....……	35
A.4	Profiles of Rater Confidence.....…….	36
A.5	Profiles of Rater Satisfaction.....…….	37
A.6	Frequency of Preferred Rating Formats.....…….	38

LIST OF TABLES

Table		Page
B.1	Summary of Participant Demographics.....	40
B.2	Summary of Statistics of Observed Variables.....	41
B.3	Doubly-Multivariate Analysis of Variance……	42
B.4	Trend Analysis of the Group by Session Interaction.....	43

CHAPTER 1

INTRODUCTION

Scientific progress has often been characterized by its ebb and flow between periods of steady advancement that excite a flurry of noteworthy contributions and gradual retreats into stagnation where popular research directions seem to reach a plateau. While almost all fields of science can be viewed through such a lens, these transitions do not necessarily take place in uniform fashion from one area of research to the next. Within the field of psychology, many personality theorists identify the 1970s as the "Decade of Doubt" (McAdams, 1997) and may even remember Carlson's (1971) article appearing in the *Psychological Bulletin* entitled, "Where is the Person in Personality Research?" calling on theorists in personality psychology to solidify their identity. Carlson (1984) published a follow-up article in the midst of a similar period for social psychologists which appeared in the *Journal of Social and Personality Psychology* under the fittingly predictable title, "What's social about social psychology? Where's the person in personality research?"

There remain a handful of catalysts which precipitate a transformation that extends to almost all scientific endeavors and cause researchers to wonder how they managed to function under their previous circumstances. Whether it is Nicolas Copernicus's *De revolutionibus obium coelestium* ushering in The Scientific Revolution or James Watt's steam powered engine energizing The Industrial Revolution, such events seem to force researchers to rethink previous conceptual theories and innovate new approaches to scientific research. One of the most recent examples of such circumstances is occurring today in what is commonly being termed The Technological Revolution (or "The Computer Age"). Within the last two decades, the progression of the personal computer and related technologies have led to advances in the

collection and analysis of data that were hardly imaginable less than fifty years ago. For example, highly sophisticated computer-based statistical applications have replaced the need for researchers to painstakingly trudge through handwritten equations. Nunnally and Bernstein (1994) point out:

> It is very easy to think that the main role of a computer is to expedite analyses that one would have performed anyway. This is certainly important. Anyone who has used computers for a long time appreciates the increasing flexibility and user-friendliness of major computer packages such as BMDP, SAS, SPSSX, SYSTAT, and UniMult. However, one additional point must be stressed – computers now allow fundamentally different kinds of analyses to be performed, i.e., open form analyses that are effectively impossible to do by hand.

As with the Industrial Revolution, these changes are hardly unique to the academic sciences. It can be argued that technology has impacted the business world more than any other context, and the pace of these changes is only expected to increase throughout the next century as well (Cascio, 1995). Within today's world of work, project teams collaborate across the globe on a real-time basis, products and services are rolled out to market in days instead of months, and vast amounts of continually updated information remain at an organization's fingertips.

Industrial and organizational (I/O) psychology lies within an area of considerable overlap between the business world and the scientific research community. As an area of psychology that is primarily concerned with the study of individuals in the workplace, it should come as no surprise that technological advances have impacted researchers within the field of I/O psychology and employees within the workplace in some very similar ways. One example of this commonality lies within the realm of psychometric theory.

1.1 Impact of Computers on Psychometric Theory

Both researchers and employers have relied on tests, surveys, and questionnaires to perform a variety of functions long before students were able to obtain a graduate degree in I/O

psychology. The first issue of the *Journal of Applied Psychology* includes a study published by Terman (1917) examining the potential benefits of using the Binet-Simon Intelligence Scale in employment testing. Since that time considerable debate has taken place regarding the use and application of intelligence tests in the workplace. While the advent of the personal computer and rise of the Internet do not seem to have resolved the debate surrounding the use of intelligence tests in the workplace, recent technological advances have caused researchers to look at testing and assessment from entirely different perspectives (Drasgow & Mattern, 2006).

Traditional applications of computer technology to psychometric theory have primarily focused on the development of computerized measures (CMs) that can replicate the functionality of paper-and-pencil measures (PPMs). Most CMs display multiple items on the computer screen at a time, similar to the presentation of items on a single page of a PPM, and participants are generally able to review previous responses, make changes to item responses, or skip forward to future items within the CM. As with their paper-and-pencil counterparts, CMs tend to display exactly the same set of items in exactly the same order to each participant (Mead & Drasgow, 1993). Both CMs and PPMs also rely upon more conventional test theories, such as classical measurement theory (also termed “classical test theory”), to obtain estimates of a participant’s true score. In other words, obtained scores are typically calculated by summing item responses together to arrive at an estimate of a true score.

The computer-based component of CMs combined with the increasing prevalence of the Internet provides a number of benefits for both the researcher and the participant. This mode of administration provides more convenience than PPMs, because participants are able to complete the measure on their own time and, in many cases, item responses can be scored automatically or imported directly into statistical packages for future data analysis. While the initial costs associated with developing CMs may exceed the costs associated with the development of PPMs, the ongoing costs are often reduced by eliminating the need for printing

copies, paying individuals to administer the measure, and avoiding the postage expenses associated with mailing measures to and from participants.

1.2 The Mode Effect

The potential benefits provided by CMs are understandably appealing to researchers, however added convenience and reduced cost become much less noteworthy if CMs reduce the validity and reliability of obtained scores. The vast amount of research published to date which has examined the similarity between CMs and PPMs seems to underscore the importance of this issue within the academic community. Previous research has found that seemingly minor modifications to a measure can yield significant changes in obtained scores (Schuman & Presser, 1996). Therefore, it should come as no surprise that many researchers continue to remain sensitive to the factors which may be responsible for differences in obtained scores depending on the unique qualities of the medium being used to administer the measure (commonly referred to as "medium effects" or "mode effects") (Mead & Drasgow, 1993; Leeson, 2006).

Mead and Drasgow (1993) published the first and only meta-analysis pertaining to mode effects. In this study, Mead and Drasgow (1993) meta-analyzed 29 studies published between 1978 and 1993 which examined the equivalence between computerized and paper-and-pencil cognitive ability tests. It was found that there was no mode effect for carefully constructed power tests of cognitive ability (i.e., tests concerned with the participant's cognitive ability in some content area with little or no regard for processing speed), however the researchers did find the existence of a significant mode effect with speed tests of cognitive ability (i.e., timed tests designed to measure processing speed). Mead and Drasgow (1993) speculated that different motor skills might be required by CMs compared to PPMs which may account for the differences in performance on measures where participants are under significant time constraints. While these findings provide evidence that mode effects may only exist in special contexts (e.g., speed tests), a primary limitation of this study relates to the limited

number of variables coded for moderator analysis (Mead & Drasgow, 1993). This limitation prevented the researchers from being able to identify specific factors which may lead to larger mode effects (e.g., amount of text within the measure, font size, etc.).

Leeson (2006) conducted a literature review of factors which may potentially moderate item response differences between CMs and PPMs and found that each of these factors typically fits into one of two categories: participant issues and technological issues. Participant issues refer to individual differences that may contribute to the mode effect. Leeson (2006) outlined five core participant issues: (1) race, ethnicity, and gender, (2) cognitive processing, (3) ability, (4) familiarity with computers, and (5) computer anxiety. Technological issues refer to user interface characteristics that may contribute to mode effects. Leeson (2006) separated user interface characteristics into two classes. The first class was termed legibility and referred to six characteristics: screen size and resolution, font characteristics, line length, number of lines, interline spacing, and white space. The second class was termed interactive and referred to three characteristics: scrolling, user flexibility (or "item review"), and item presentation (for complete review, see Leeson, 2006).

Many of these factors are variables that the participant brings to the test setting (i.e., participant issues), and the researcher has very little recourse beyond providing time for pretesting to become comfortable with the computer-based medium. As it relates to technological issues, previous research demonstrates that these effects can be mitigated by developing CMs that mirror the functionality and appearance of PPMs as closely as possible (Spray, Ackerman, Reckase, & Carlson, 1989). But what about those instances when it is impossible to replicate the functionality and appearance of PPMs?

1.3 Computer-Adaptive Measures

Examples of these instances would appear to be few and far between when relating the list of potential mode effects to CMs as they have previously been outlined in this literature review. However, recent applications of less conventional psychometric theories (termed "item

response theories" [IRTs]) to computer-based methods for obtaining information about individuals have created a new breed of computer-adaptive measures (CAMs) that are quickly becoming both popular and prevalent in today's workplace. CAMs are designed to tailor the item content of the measure to the ability level of the participant based on his or her previous item responses (Wainer, 2000). CAMs share many of the benefits associated with more traditional CMs, with some important additions.

As previously mentioned, traditional CMs rely upon more conventional psychometric theories, such as classical measurement theory, whereby each participant receives the same item content and item responses are summed together to arrive at an estimate of the participant's true score. Classical measurement theory tends to provide the most accurate estimate of a true score for participants falling nearest to the central tendency of obtained scores within the distribution curve, and this precision becomes poorer as participants' obtained scores fall closer to the extremities of the curve. By relying upon IRTs, rather than classical measurement theories, CAMs are able to arrive at an estimate of a participant's true score at the item level. This characteristic of IRT allows CAMs to arrive at a more precise estimate of the participant's true score with fewer items. Weiss and Kingsbury (1984) estimate that CAMs typically require 50 percent fewer items to arrive at an equally precise estimate of a participant's true score.

Despite the utility of CAMs in selected contexts, these methods are far from a panacea for all challenges related to psychometric theory. CAMs require a much larger pool of potential items and pilot test populations typically range from 5,000 to 50,000 individuals (Bartram, 2006). The majority of commercially available CAMs are also designed for selected-response measures which instruct participants to choose the correct answer from a list of answer choices (e.g., ability tests). Researchers have only recently begun to apply CAMs to contexts beyond the realm of ability testing.

Within the last decade, researchers have begun to examine the potential for using computer-adaptive rating scales (CARS) to evaluate employee performance in the workplace (Borman et al., 1998). The application of CARS to the evaluation of performance in the workplace is particularly appealing, because computer-adaptive methods have demonstrated the capacity to decrease the time it takes to complete a measure and obtain more accurate estimates of true scores. Within the realm of personality testing, computer-adaptive methods have also demonstrated the potential to mitigate socially desirable responses to items within the measure by identifying cases where item responses consistently differ from the predicted item response curve (Rothstein & Goffin, 2006).

Taken together, these three benefits (i.e., ability to decrease time to completion, ability to increase accuracy of ratings, and ability to identify suspicious response patterns) address some of the primary shortfalls of performance appraisal systems as they are currently being used. For example, employees in today's workplace can be asked to rate the performance of fifty or more co-workers per year (Salopek, 2004). The time that it takes these employees to complete numerous questionnaires represents a labor cost that employers would presumably prefer to see decreased in the future. As it relates to the accuracy of ratings, the validity and reliability of performance ratings provided by co-workers is still lacking in many respects (Conway & Huffcutt, 1997; Greguras & Robie, 1998). Efforts to increase the accuracy of performance ratings have produced negligible improvements over the years, and caused some researchers to call for a moratorium on research efforts trying to increase their accuracy (Landy & Farr, 1980). Lastly, as it relates to identifying suspicious response patterns, formal performance appraisal systems have long been accused of facilitating collusion among co-workers (or "mutual back patting") and tit-for-tat games where co-workers get even with those thought to be responsible for negative ratings in their own performance reviews. In a field study conducted by Longenecker and Ludwig (1990) more than 70 percent of supervisors indicated that they have intentionally inflated or deflated performance ratings at one time or another in

order to protect a colleague or send a message to a poor performer. From the organization's perspective, each of the benefits associated with CAMs has the potential to address some of the most serious deficiencies inherent to the performance rating process.

With this said the administration of CAMs is considerably different from traditional CMs and PPMs, in that, most CAMs rely on single-item presentation and restrict participants from being able to review and make changes to previous responses. These restrictions are necessary, because CAMs rely on previous item responses in order to tailor future items to the participant's ability level (or level of the construct being measured). Vispoel (1998) examined the effects of allowing item review on CAMs and found that allowing item review increased testing time by 41 percent, reduced measurement precision, complicated item administration algorithms, and artificially inflated ability estimates for participants. It is also quite illogical for participants to be able to skip forward to items which, theoretically, have not been selected prior to obtaining responses from earlier items.

1.4 Mode Effects Associated with User Flexibility

User flexibility refers to the extent to which a measure allows individuals responding to items to move freely from item to item and review or make changes to responses provided. Prior to the development of CMs, most psychometric research focused on the common tendencies of users to review and change previous responses to items (Vispoel, 1998). Benjamin, Cavell, and Shallenberger (1984) outlined four primary findings in their review of 33 published studies that pertain to item review. They found that only a small percentage of items are actually changed, more answers are changed from wrong to right than from right to wrong on selected-response measures, most people change responses to some items, and most people who change answers on selected-response measures increase their score (Benjamin, Cavell, & Shallenberger, 1984).

The available research examining the mode effects associated with user flexibility has been lacking in many respects (Vispoel, 1998). Only a handful of research studies have

examined these effects, and results have been mixed to date (Leeson, 2006). Eaves and Smith (1986) randomly assigned individuals to receive a selected-response measure using one of two administration procedures: paper-and-pencil test and fixed-item, computer-based test. The paper-and-pencil test allowed the test taker to move freely from item to item, review previous answers, and change answers provided. The computer-based test presented the test taker with one item at a time, restricted the test taker from moving freely from item to item, prevented review of previous answers, and did not allow test takers to change answers provided. No significant differences were found in test performance between the two administration procedures (Eaves & Smith, 1986).

Luecht, Hadadi, Swanson, and Case (1998) took the Eaves and Smith (1986) study one step further by randomly assigning individuals to receive a selected-response measure using one of three administration procedures: paper-and-pencil test, fixed-item, computer-based test without item review, and fixed-item, computer-based test with item review. The paper-and-pencil test and the fixed-item, computer-based test without item review were similar to both administration procedures within the Eaves and Smith (1986) study, while the fixed-item, computer-based test with item review was designed to mimic the capabilities of the paper-and-pencil test. Results were consistent with Eaves and Smith (1986) and demonstrated that preventing item review did not lead to significant differences in test scores (Luecht et al., 1998).

Conversely, Moreno, Lee and Sympson (1985) conducted a study to examine the effects of transferring items from a paper-and-pencil arithmetic reasoning to an early version of computer-based testing. This early version displayed one item at a time and prevented test takers from being able to skip items, review answers, and make changes to previous responses. Results from this study demonstrated that individuals using the paper-and-pencil format scored significantly higher than those individuals using the CMs (Moreno, Lee, & Sympson, 1985).

While much of the literature pertaining to user flexibility focuses on changes to the psychometric properties of a measure after restricting the user's ability to review item responses or skip over items, there is a growing body research that examines the psychological reactions associated with user flexibility. For example, most participants report frustration with the testing process when they are restricted from being able to review previous responses (Luecht et al., 1998; Vispoel, Wang, de la Torre, Bleiler, & Dings, 1992). Luecht and colleagues (1998) surveyed participants completing an ability test which prevented them from being able to review previous responses and found that 20 percent of these participants identified lack of item review as the factor that they liked least about the testing process. This negative reaction appeared independent from obtained scores on the ability test, because there were no differences in test performance based upon whether or not item review was permitted (Luecht et al., 1998).

1.5 Mode Effects Associated with Item Presentation

Item presentation refers to the number of items presented to participants at one time (i.e., the number of items displayed on a computer screen). Dimock and Comier (1991) attempted to measure the effects of item presentation on obtained scores by using index cards to present verbal reasoning items one at a time and comparing these scores to a paper-and-pencil format with identical item content. Results demonstrated that individuals scored significantly higher on the paper-and-pencil format than when presented with one item at a time via index cards. Dimock and Comier (1991) also compared the single-item presentation via index cards to the single-item presentation via computer, and found that individual scores still differed significantly. These results suggested that novelty of the single-item presentation via index cards may have confounded comparisons to paper-and-pencil formats (Leeson, 2006).

A review of previous literature focusing only on computer-based methods suggests that grouping items together on the display screen facilitates performance on ability tests (Hofer & Green, 1986; Lee, 1986). When using non-ability measures such as organizational surveys, Rosenfeld, Booth-Kewley, and Edwards (1993) speculate that single-item presentation may

cause individuals to focus on each item being presented to a greater degree, although the absence of empirical research on the subject to date makes this hypothesis largely anecdotal.

The effect of item presentation on user perceptions have been studied to a much lesser degree than user perceptions associated with user flexibility. Some researchers suggest that participants completing single-item presentation measures may have more difficulty tracking the number of items left to complete which may prove frustrating (Rosenfeld, Booth-Kewley, & Edwards, 1993), however no research has been conducted to test this hypothesis. Further research is needed within this area in order to determine whether or not item presentation affects user perceptions.

1.6 Integrated Analysis of the Mode Effects Associated with CAMs

Taken together, CAMs provide the potential for researchers and employers to obtain more precise information with greater efficiency, however CAMs are also associated with two additional mode effect factors (as identified by Leeson, 2006) which are quite unavoidable as things currently stand: user flexibility and item presentation. Mead and Drasgow (1993) included CAMs in their meta-analysis of cognitive ability tests and found no evidence of a mode effect associated with CAMs. In contrast to these findings, previous research has found the existence of mode effects with more traditional CMs that mirror the functionality of CAMs by presenting one item at a time and restricting participants from being able to skip past items, review previous responses, and make changes (Leeson, 2006). These contradictions seem to suggest that the benefits of the computer-adaptive component might mitigate the mode effects which may be present due to the differences in the administration of items between CAMs and more conventional measures.

These assumptions can only be made in the context of ability testing, however, because very few researchers have compared CAMs to more conventional measures outside the realm of ability testing (e.g., personality testing, performance appraisal, etc.). Even fewer researchers have examined the mode effects associated with item presentation and user

flexibility outside the realm of ability testing. Considering the recent applications of CAMs to contexts other than ability testing, it would appear that this lack of research should be of particular interest to both researchers in the field of I/O psychology and employers considering the use of CAMs for purposes other than ability testing.

1.7 Computer-Adaptive Rating Scales

Performance appraisals obtain estimates of a true score by asking employees in contact with the participant to rate his or her performance in the workplace. Researchers hope that CARS will provide the performance appraisal process with similar benefits to those found in the realm of ability testing by obtaining more accurate estimates of an employee's job performance. This is a significant departure from other applications of CAMs that should lead employers to question the generalizability of previous research examining the mode effects associated with CAMs.

Looking beyond the psychometric properties of CARS, the success of any performance appraisal system is highly dependent upon the acceptance of performance ratings by employees within the organization (Tziner & Kopelman, 2002). Previous research on the mode effects associated with item presentation and user flexibility suggests that participants are less confident in their responses and less satisfied with the measurement process when they are restricted from moving freely from item to item, reviewing previous responses, and making changes to previous responses. It remains to be seen whether or not these negative reactions have the potential to undermine the utility of CARS in the performance appraisal process.

1.8 Study Objectives

Within the last five years, researchers have begun looking for ways to apply computer-adaptive methods to obtain performance ratings within the workplace (Borman, Buck, & Hanson, 2001; Schneider, Goff, & Anderson, & Borman, 2003). No known research has been published to date which examines the effects of the functional properties associated with CARS (i.e., single-item presentation and low user flexibility) on performance ratings provided by others

and rater perceptions of the performance rating process. The purpose of this study was to bridge this apparent gap within the research literature pertaining to CARS. The aims of this study were as follows:

1. The first aim of this study was to determine the extent to which performance ratings are affected by user flexibility (high user-flexibility vs. low user-control) and item presentation (multiple-item presentation vs. single-item presentation).
2. The second aim was to determine the extent to which user flexibility and item presentation affect rater confidence in the accuracy of performance ratings.
3. The third aim was to determine whether or not user flexibility and item presentation affect rater satisfaction with the performance rating process.

1.9 Study Hypotheses

1.9.1 Hypothesis One

Rater confidence in the accuracy of performance ratings will differ across administration procedures, such that presenting one item at a time and restricting user flexibility will decrease rater confidence in the accuracy of performance ratings.

1.9.2 Hypothesis Two

Rater satisfaction the performance rating process will differ across administration procedures, such that presenting one item at a time and restricting user flexibility will decrease perceptions of satisfaction with the functioning of the performance rating process.

CHAPTER 2

METHODS

2.1 Experimental Design

The primary purpose of this study was to evaluate the effect of item presentation and user flexibility on performance ratings and rater perceptions of the performance rating process. The two manipulated factors were item presentation (single-item presentation vs. multiple-item presentation) and user flexibility (high user-flexibility vs. low user-flexibility). The single-item presentation format displayed one item at a time on the computer screen. The multiple-item presentation format displayed eight items at a time on the computer screen. The high user-flexibility format allowed participants to navigate forward/backward within the measure and allowed participants to make changes to previous responses. The low user-flexibility format restricted participants' ability to navigate forward/backward within the measure and restricted participants' ability to make changes to previous responses. Three DVs were included within this study: (a) student ratings of instructor performance, (b) rater confidence in the accuracy of performance ratings, and (c) rater satisfaction with the performance rating process.

The manipulated factors were factorially combined to create four IVs ("rating formats"): (a) multiple-item presentation, high user-flexibility (MH), (b) multiple-item presentation, low user-flexibility (ML), (c) single-item presentation, high-user flexibility (SH), (d) single-item presentation, low user-flexibility (SL). The three DVs were repeatedly measured across all four IVs in order to evaluate the within-subjects effects of item presentation and user flexibility on performance ratings and rater perceptions of the performance rating process. This procedure produced a within-subject design with multiple DVs.

It was a concern within this study that order effects would produce differences in the DVs. For example, it was believed that participants may experience rater fatigue as they experienced more levels of the IV. A Latin square technique was used to arrange the order of presentation of the four IVs, so that order effects were distributed evenly across repeated measurements. This procedure produced a Latin square design with multiple DVs.

To evaluate the between-subjects effects of item presentation and user flexibility on performance ratings and rater perceptions of the performance rating process, the order of presentation of the four IVs (i.e., Latin square arrangement) was treated as a between-subjects grouping variable. This procedure produced a between-within-subjects design with multiple DVs.

It was also a concern within this study that within-subjects effects would violate the assumption of sphericity. In order to circumvent this assumption, both the within-subjects part of the design and the multiple DVs were analyzed multivariately. This procedure produced a doubly-multivariate experimental design. The between-subject effect was singly multivariate; the within-subject effects and interactions were doubly-multivariate. Figure 1 displays the structure of the doubly-multivariate design used within this study.

2.2 Participants

Participants were undergraduate students enrolled in one or more psychology courses at the University of Texas at Arlington (UTA). Students participating in this study received one credit hour towards fulfillment of course requirements. Participants were recruited to participate in this study from February 21, 2008 until May 1, 2008. Informed consent was obtained from all students prior to their participation in this study.

2.3 Measures

As indicated in the Experimental Design subsection, three DVs were included within this study: (a) student ratings of instructor performance (“performance ratings”), (b) rater confidence in the accuracy of performance ratings (“rater confidence”), and (c) rater satisfaction with the

performance rating process ("rater satisfaction"). Two dependent measures were developed prior to data collection to measure these three DVs. Performance ratings were measured using the Student Evaluation of Instructor Competence (SEIC). Rater confidence and rater satisfaction were measured using the User Evaluation of the Performance Rating Process (UEPRP). A supplemental demographic questionnaire (SDQ) was also administered to participants to identify the demographic characteristics of the sample population and control for potential confounding variables. Appendix A provides a copy of all the measures used in this study.

2.3.1. Student Evaluation of Instructor Competence

The Student Evaluation of Instructor Competence (SEIC) is a 33-item performance evaluation that asks students to rate their instructor's performance on a variety of competency behaviors believed to contribute to effective or ineffective performance. Students rate their level of agreement with each competency behavior on a 5-point response scale where 1 = Strongly Agree and 5 = Strongly Disagree. Students were also provided with the opportunity to select 'Don't Know' from the response choices if they felt they were unable to rate their instructor on a specific competency behavior. Of the 33 items included within the SEIC, 16 of these items illustrate core behaviors that most raters were considered to have ample opportunity to observe during interactions with their instructor. These 16 items (1, 3, 5, 6, 14, 20, 21, 23, 25, 27, 28, 29, 30, 31, 32, and 33) were averaged together to compute a dependent measure of instructor performance for this study. Within this study, the 16-item SEIC was found to have a Cronbach's alpha of .90.

2.3.2. User Evaluation of the Performance Rating Process

The User Evaluation of the Performance Rating Process (UEPRP) is a 7-item self-report measure used to evaluate rater perceptions of the performance rating process. The first six items in the UEPRP are fixed-response items believed to target positive rater perceptions or negative rater perceptions. Students rate their level of agreement with each of these items on a

6-point-response scale where 1 = Strongly Disagree and 6 = Strongly Agree. Item 1 through 3 were developed to target rater confidence, and items 4 through 6 were developed to target rater satisfaction. Item 2 was reverse-scored so that lower scores yielded higher confidence in the accuracy of performance ratings. Items 1 through 3 were averaged together to compute a dependent measure of rater confidence for this study. Within this study, the three items pertaining to rater confidence were found to have a Cronbach's alpha of .76. Items 4 through 6 were averaged together to compute a dependent measure of rater satisfaction for this study. Within this study, the three items pertaining to rater satisfaction were found to have a Cronbach's alpha of .73. Item 7 was an optional open-response item where participants were encouraged to provide additional feedback on the performance rating process.

2.3.3. Supplemental Demographic Questionnaire

The Supplemental Demographic Questionnaire (SDQ) is a 17-item self-report questionnaire used to collect demographic information and control for potential confounding variables. Item 1 through 7 were developed to collect basic demographic information and target potential confounding variables such as class attendance, current grade point average, and expected grade point average for the course taught by the instructor they are rating. Items 8 through 17 were taken from Loyd and Gressard's (1984) 30-item Computer Attitudes Scale (CAS). The 10 items used within this study represent the Computer Confidence subscale (α = .91). This subscale asks respondents to rate their level of agreement with each computer confidence item on a 6-point response scale where 1 = Strongly Disagree and 6 = Strongly Agree. Items 3, 4, 5, 7, and 8 were reverse-scored so that lower scores yielded higher confidence with computers. Item 8 through 17 were summed together to arrive at an overall computer confidence rating.

2.4 Procedure

Details of this study were posted by the UTA Department of Psychology to its online Sona system. The Sona system is an online network used to track student participation in

departmental research activities. Students navigating within the Sona system were able to review the details of this study and decide whether or not to participate.

Students who decided to participate in this study were able to sign up through the Sona system. Study information posted on the Sona system notified students who signed up for this study that they would receive an e-mail message with a hyperlink to the online performance rating sessions within 48 hours of signing up. After receiving the e-mail message and clicking on the enclosed hyperlink, participants were directed to one of the four randomly assigned experimental groups described in the Experimental Design subsection.

All participants completed a total of four performance rating sessions. Each rating session utilized a different rating format: (i) multiple-item presentation, high user-flexibility (MH), (ii) multiple-item presentation, low user-flexibility (ML), (iii) single-item presentation, high user-flexibility (SH), and (iv) single-item presentation, low user-flexibility (SL). Participants within each experimental group completed the four rating formats in a different order: (a) MH, SL, ML, SH (Condition 1), (b) ML, SH, MH, SL (Condition 2), (c) SH, ML, SL, MH (Condition 3), (d) SL, MH, SH, ML (Condition 4). Figure 2 details the survey administration procedures used for each of the experimental groups within this study.

Participants were asked to rate the performance of the same instructor across all four rating sessions. Participants were asked to provide the name of the instructor whose performance they were rating at the beginning of each session. During each rating session, participants rated their instructor's performance using the SEIC. After each performance rating session, participants completed the UEPRP as a measure their confidence in the accuracy of their performance ratings and satisfaction with the performance rating process. The SDQ was administered to participants after the fourth rating session.

While there were only four performance rating sessions in this study, participants were informed that there would be five sessions. The presence of the additional faux-rating session was communicated to participants in order to determine which of the rating formats they most

preferred. After completing the SDQ, participants were instructed to select the rating format that they most preferred, and participants were led to believe the rating format that they chose would be the format utilized to administer items during the fifth rating session. After the participants chose their preferred rating format, a message was displayed on the screen indicating that the previous question was used to determine their preferred rating format and that they would not need to complete a final rating session (i.e., they were informed that their participation in the study was complete).

2.5 Exclusion Criteria

Participants were selected for inclusion within this study if they met three minimum criteria: (1) Completion of all four rating sessions within a period of 24 hours, (2) Completion of at least 50% of survey items for each dependent measure, and (3) Minimum levels of response agreement across the four rating sessions. The SEIC contains four reverse-coded items developed to identify participants who provide logically discrepant responses. For example, a participant who strongly agrees with the statement, “Makes self accessible and readily available to students,” and strongly agrees with the statement, “Rarely make self accessible and readily available to students,” would be classified as providing a single instance of response discrepancy. Participants could be classified as providing discrepant responses on four occasions within the SEIC, and a total of 16 occasions within the entire study (i.e., four instances on the SEIC multiplied by four rating sessions). If participants were classified as having more than eight instances of response discrepancy, then they were identified as failing to meet the minimum levels of response agreement across the four rating sessions.

2.6 Preliminary Data Analysis

Preliminary data analysis focused on screening collected data for any characteristics that might pose problems for interpreting results from multivariate statistics. Issues of primary importance within this study pertained to the potential for outliers to influence obtained results, heterogeneity of variance-covariance matrices, and deviations from multivariate normality.

Repeated-measures statistical designs are typically quite robust to multicollinearity, and the choice to use a doubly-multivariate design circumvents most issues pertaining to sphericity (Tabachnik & Fidell, 2007); therefore these assumptions were less of a concern within this study.

2.7 Hypothesis Testing

A doubly-multivariate analysis of variance was conducted to test hypotheses included within this study. Within a doubly-multivariate analysis of variance the between-subject effect is treated as a singly multivariate analysis, and the within-subject effects and interaction effects are treated as a doubly multivariate analysis (for complete review, see Tabachnick & Fiddell, 2007). The between-subjects grouping variable was the experimental group to which each participant was randomly assigned (i.e., Condition 1, Condition 2, Condition 3, and Condition 4). The three noncommensurate (i.e., not all measured on the same scale) dependent variables were (a) student ratings of instructor performance measured using the SEIC, (b) rater confidence in the accuracy of performance ratings measured using the UEPRP, and (c) rater satisfaction with the performance rating process measured using the UEPRP. The within-subjects variable was the rating session, and this variable contained four repeated measurements. Figure 1 reviews the doubly-multivariate design used within this study.

2.8 Exploratory Analysis

As previously stated in the Procedure subsection, participants were notified at the beginning of this study that they would complete a total of five rating sessions. After completing the fourth rating session, participants were instructed to select the rating format that would be used to administer items in a final rating session. Participants were not actually required to complete a fifth rating session; however their rating format selections were recorded to evaluate rater preferences between the four rating formats. A one-sample chi-square test was conducted on the frequency of rating formats selected by participants. Follow-up pairwise comparisons were conducted between item presentation (multiple-item vs. single-item) and user flexibility

(high vs. low) to further explore rater preferences.

CHAPTER 3

RESULTS

3.1 Demographic Characteristics

A total of 282 undergraduate students consented to participate in this study. 19 participants were excluded from the sample population, because they did not complete all four rating sessions within a period of 24 hours (Exclusion Criterion 1). Four participants were excluded from the sample population, because they did not respond to 50% or more of the performance evaluation items (Exclusion Criterion 2). Seven participants were excluded from the sample population, because they did not demonstrate minimum levels of response agreement on the SEIC (Exclusion Criterion 3). Therefore, 30 participants were excluded from the sample population.

The sample population (*N* = 252) was represented by 171 female participants (67.9%), 80 male participants (31.7%), and one participant (0.4%) with no gender reported. The modal age range of the sample population was 18 to 20 years old (56.0%), and more than 95% of the sample population fell below 30 years old. Three participants (1.2%) did not report their age category. Nearly half of all participants identified themselves as White / Caucasian American (49.6%) with Asian / Asian American representing the second most frequently identified response option (16.7%) followed by Hispanic/Latino American (15.%), Black / African American (14.7%), and American Indian / Native American (0.8%). Eight participants (3.2%) identified themselves more closely with a race or ethnicity that was not provided as a response option. Table 1 provides a summary of the distribution of demographic characteristics across the four experimental groups included within this study.

3.2 Preliminary Data Analysis

Univariate outliers were identified by separating participants according to their experimental group and calculating standardized scores (z-scores) on each of the dependent measures to be used during hypothesis testing. Participants with a z-score > 3.29, $p < .001$ were considered univariate outliers, and a total of 7 participants were classified as outliers when applying this criterion. Multivariate outliers were identified by separating participants according to their experimental group and calculating Mahalonobis distance values ($\chi 2$) on the combination of dependent measures to be used during hypothesis testing. Participants with $\chi 2 > 39.25$, $p < .001$ were considered multivariate outliers, and a total of 8 participants were classified as outliers when applying this criterion. Upon closer examination, it was determined that most of these outliers tended to rate their instructor's performance lower than other participants within their experimental group. A decision was made not to delete any cases, because it was believed that deleting low performance ratings would decrease rather than increase the generalizability of the sample population. Transformation of variables with outlying cases was also not considered to be an appropriate choice, because doing so would increase the difficulty associated with interpreting results from an already complex statistical design. Therefore, the decision was made to analyze the collected data with and without outlying cases in order to determine whether or not outlying cases influenced the obtained results.

Within this study, the ratio of participants in the smallest experimental group ($n = 61$) to total DVs ($n = 16$) was 3.8 to 1.0, and sample sizes were generally equivalent among all experimental groups; therefore deviations from normality of the sample distributions and heterogeneity of variance-covariance matrices were not expected. The DVs used within this study tended to demonstrate moderate levels of negative skewness and positive kurtosis; however the same properties were apparent among all variables. Also, the variances for all 16

DVs fell at or below a 2 to 1 ratio. These observations appeared to confirm the expectation that deviations from multivariate normality and heterogeneity of variance-covariance matrices would not pose problems when interpreting obtained results during hypothesis testing.

High correlations among DVs were expected within this study; however tolerance levels were computed on each DV in order to determine whether or not statistical multicollinearity might pose problems when interpreting results during hypothesis testing. All tolerance levels were found to be equal to or greater than .035; therefore statistical multicollinearity was not expected to pose problems when interpreting results from hypothesis testing.

3.3 Hypothesis Testing

The doubly-multivariate analysis of variance yields three different significance tests: (a) flatness test, (b) levels test, and (c) parallelism test. The doubly-multivariate flatness test is whether, with experimental groups combined, the slope of the profiles for the repeatedly measured DVs deviate from zero (i.e., within-subject effect). The singly-multivariate levels test is whether, with rating sessions combined, the means of the DVs differ between experimental groups (i.e., between-subject effect). The doubly-multivariate parallelism test examines whether there is an interaction between the profiles for the repeatedly measured DVs and the experimental groups (i.e., group by session interaction effect). Table 2 displays the means and standard deviations of the observed variables within this study as a function of experimental group, rating session, and rating format.

Multivariate tests of flatness and levels effects analyze mean differences by combining experimental groups (flatness) or rating session (levels) resulting in the loss of information pertaining to rating format. Therefore, significant deviations from flatness and significant mean differences between experimental groups do not necessarily imply mean differences between rating formats. Because each experimental group was presented with a different rating format in each rating session, effects attributable to the rating format would be evidenced by a significant interaction between experimental group and rating session. Therefore, deviations from

parallelism were of primary interest within this study. Using Wilks' criterion, the doubly-multivariate test of parallelism effects indicated the presence of a significant interaction between experimental group and rating session, $F(27, 702) = 1.56$, $p < .05$, partial $\eta 2 = .06$. Table 3 summarizes the results of the doubly-multivariate analysis of variance. Profile plots for each dependent measure can be found in Figures 3 through 5.

Deviations from parallelism were explored by decomposing univariate effects into trend analysis. This procedure was considered preferable, because the assumption of sphericity that was circumvented when examining multivariate effects (using a doubly-multivariate design) could still be avoided when examining univariate effects (Tabachnick & Fidell, 2007). An experimentwise error rate of 5% was achieved by setting $\alpha = .004$ for each the 12 trends to be evaluated. None of the trends for the experimental group by rating session interaction effect reached significance at $\alpha = .004$. Table 4 summarizes the results of the trend analysis for the interaction effect.

3.4 Exploratory Analysis

Despite null findings during hypothesis testing, exploratory analyses revealed significant differences in the frequency of rating formats selected as most preferred by participants, $\chi 2 = 149.56$, $p < .001$. The effect size of .20 indicates that the observed frequencies deviate moderately from the expected frequencies. The proportion of raters who selected the multiple-item presentation, high-user flexibility format ($P = .57$) was much greater than the expected proportion of .25, while the proportion of raters who selected the single-item presentation, low user-flexibility format ($P = .05$) was much lower than the expected proportion. The proportion of raters who selected the multiple-item presentation, low-user flexibility format ($P = .21$) and single-item presentation, high user-flexibility format ($P = .17$) were approximately the same value and less than the expected proportions of .25.

In order to address the possibility that rater preferences might be affected by order effects, a two-way contingency table analysis was performed to evaluate the relationship

between experimental group and preferred rating format. Experimental group was included as a variable within this analysis, because rating formats were presented in a different order to each experimental group. Therefore, a significant relationship between experimental group and preferred rating format would indicate a relationship between order of presentation and preferred rating format. Experimental group and preferred rating format were not found to be significantly related, Pearson $\chi 2 = 13.14$, $p = .16$; therefore rater preferences did not appear to be affected by order effects.

A follow-up test indicated that the proportion of raters who preferred rating formats with multiple-item presentation was significantly higher than the proportion of raters who preferred rating formats with single-item presentation, $\chi 2 = 80.02$, $p < .001$, and size of this effect was approximately .56 ($P2 – P1$). Another follow-up test indicated that the proportion of raters who preferred rating formats with high user-flexibility was significantly higher than the proportion of raters who preferred rating formats with low user-flexibility, $\chi 2 = 55.25$, $p < .001$, and the size of this effect was approximately .46 ($P2 – P1$). Figure 6 provides a bar chart depicting the frequency that each of the four rating formats was selected by raters.

CHAPTER 4

DISCUSSION

Results from hypothesis testing did not yield any significant effects attributable to rating formats. Post hoc pairwise comparisons of marginal means were performed in order to confirm null findings obtained during hypothesis testing. Bonferroni's correction was applied to all pairwise comparisons in order to control for inflated Type I error rates. Consistent with null findings obtained during hypothesis testing, no significant differences were found between rating formats on the three dependent measures. Hypothesis testing was also conducted after excluding multivariate outliers, and the obtained results did not differ from reported findings. Therefore, it was concluded that item presentation and user flexibility did not significantly affect performance ratings (Research Aim 1), rater confidence in the accuracy of performance ratings (Research Aim 2, Hypothesis 1), or rater satisfaction with the performance rating process (Research Aim 3, Hypothesis 2).

Regardless of the rating format being administered to participants, raters remained quite confident in the accuracy of their performance ratings and were generally satisfied with the overall functioning of the online performance rating process. Despite these positive rater perceptions, raters did appear to have distinct preferences for certain rating formats. Results from exploratory analysis indicated that participants (a) preferred multiple-item presentation formats with a much higher frequency than single-item presentation formats, and (b) preferred high user-flexibility formats with a much higher frequency than low user-flexibility formats. The effects sizes associated with each of these differences were quite large; however rater preferences pertaining to item presentation were slightly stronger than rater preferences pertaining to user flexibility.

4.1 Limitations of the Study

One possible limitation of this study relates to the generalizability of the experimental setting to the workplace. Previous research suggests that performance ratings provided by raters can be influenced, to some degree, by characteristics of jobs and organizations, the purpose of the ratings, and characteristics of raters and ratees (Landy & Farr, 1980). The current study utilized a sample population composed of undergraduate students being asked to rate the performance of their psychology instructor. It was also communicated to these participants that the instructor being rated would not have access to the performance ratings provided. These circumstances may limit the generalizability of findings within this study to rating conditions found within the workplace.

Another possible limitation relates to the use of dependent measures that were not previously validated prior to data collection. While estimates of internal consistency for each of the dependent measures were found to be adequate for the purposes of this study, the results obtained within this study were reliant upon the face validity of the dependent measures. The possibility exists that the scale items did not adequately target constructs related to variables of interest to this study.

Finally, the current study did not examine all of the functional differences associated with CARS. For example, the CARS formats appearing within previous research utilize a forced-choice response scale (Borman et al., 2001). Whereas item presentation and user flexibility do not appear to significantly affect rater confidence and rater satisfaction, forced-choice response scales might yield significant differences on these outcome variables. Another characteristic of computer-adaptive rating scales that was not examined within this study relates to the variability in the item content being presented to raters. The item content within most computer-adaptive measures is not fixed; therefore it is quite unlikely that two raters would evaluate a target individual's performance across identical behavioral items. The fact that different raters are

being asked to rate different behaviors during the performance rating process may significantly affect rater perceptions of the performance rating process.

4.2 Implications

The results of this study have implications for both research and practice. A major theme within performance rating research over the past half century has been the need to improve the quality of performance ratings being obtained within the workplace (Murphy, 2008). While research examining the potential to increase the quality of performance ratings by using different rating formats was quite prevalent within the 1960s and 1970s; this area of research has decreased substantially since Landy and Farr's (1980) call for a moratorium on rating format research. As reported by Landy and Farr (1980) in their often cited review of research pertaining to performance ratings, the percent of variance in performance ratings that can be attributed to rating formats generally falls within the range of 4 to 8 percent. These effect sizes were considered to be of such a trivial nature to warrant a moratorium on research into rating formats altogether (Landy & Farr, 1980).

In the midst of this decrease in rating format research, recent applications of computer-adaptive methodologies to performance rating measures have yielded improvements in the quality of performance ratings not usually witnessed within this area of research (Borman et al., 2001; Schneider et al., 2003). Borman and colleagues (2001) compared CARS to other popular performance rating tools, and found incremental increases in validity (d = .18), incremental increases in accuracy (d = .07), and 23% to 37% lower standard errors of measurement. Some of these effects are two to four times greater than the range of effect sizes identified by Landy and Farr (1980) in their review of research pertaining to rating formats.

Despite the potential for CARS to improve the quality of performance ratings, very little research has been conducted to evaluate the extent to which the functional differences between CARS and more traditional rating scales affect rater perceptions of the performance rating process. The utility of performance rating systems are highly dependent upon the acceptance of

rating process by employees within the organizations (Tziner & Kopelman, 2002); therefore studies examining this underrepresented area of research represent a welcome contribution to the field of I/O psychology.

This study advances recent research pertaining to CARS by demonstrating that two prominent functional differences between CARS and more traditional rating scales (i.e., item presentation and user flexibility) do not significantly affect two important rater perceptions inherent to the performance rating process (i.e., rater confidence and rater satisfaction). This study also demonstrates that raters, if given the choice, generally do not prefer some of the functional properties associated with CARS (i.e., single-item presentation and low user flexibility). In fact, raters demonstrated just the opposite by reporting preferences for the functional properties associated with more traditional rating tools (i.e., multiple-item presentation and high user-flexibility). These finding have implications for the utility of computer-adaptive methodologies applied to performance rating measures. While computer-adaptive methodologies generally require the restriction of user flexibility, methodologies have been developed which do not require single-item presentation formats. Therefore, the development of algorithms that allow CARS to utilize multiple-item presentation formats may mitigate some of the differences in rater preferences demonstrated within this study.

4.3 Future Research Directions

Future research could expand or improve upon findings within this study by obtaining estimates of actual performance so that the effects of item presentation and user flexibility on performance ratings can be evaluated on factors such as rater accuracy or sensitivity to rater errors. The analysis of rater perceptions using qualitative data is another area of research that warrants further exploration. Over 150 comments were collected from participants responding to an option feedback response item included within this study. Some of the information provided by participants was not relevant to the rating format being administered, but several comments were found to be directly related to item presentation and user flexibility. A number of participant

comments identified the single-item presentation format as being overly cumbersome and adding too much time the performance rating process; however some participant comments identified the single-item presentation format as allowing them to more easily focus on each behavioral item that they were rating. Participant comments endorsing the multiple-item presentation format generally identified the decreased time to completion as a positive aspect of the rating format. Some participant comments identified the restricted functionality of the low user-flexibility format as a limiting factor. In general, 400 participant comments provide a representative sample for the analysis of qualitative data (Cozby, 2000). While this study falls well below the number of participant comments needed to produce a generalizable sample, the content analysis of qualitative data may yield information related to item presentation and user flexibility that fixed response scales are unable to measure.

4.4 Concluding Comments

The current research examined the effects of item presentation and user flexibility on performance ratings and rater perceptions of the performance rating process. While item presentation and user flexibility did not significantly affect rater perceptions of the performance rating process, raters were found to prefer rating formats related to traditional performance rating measures over rating formats related to computer-adaptive performance rating measures. This finding does not represent a limitation that should forestall future research into CARS. It appears to be an unfortunate coincidence that few researchers have initiated independent explorations of CARS since their initial development, and no research has been published on the subject within the last five years. In light of the fact that, during this lull, researchers increasingly question the utility of performance rating measures, the potential benefits of CARS almost certainly outweigh the limitations.

APPENDIX A

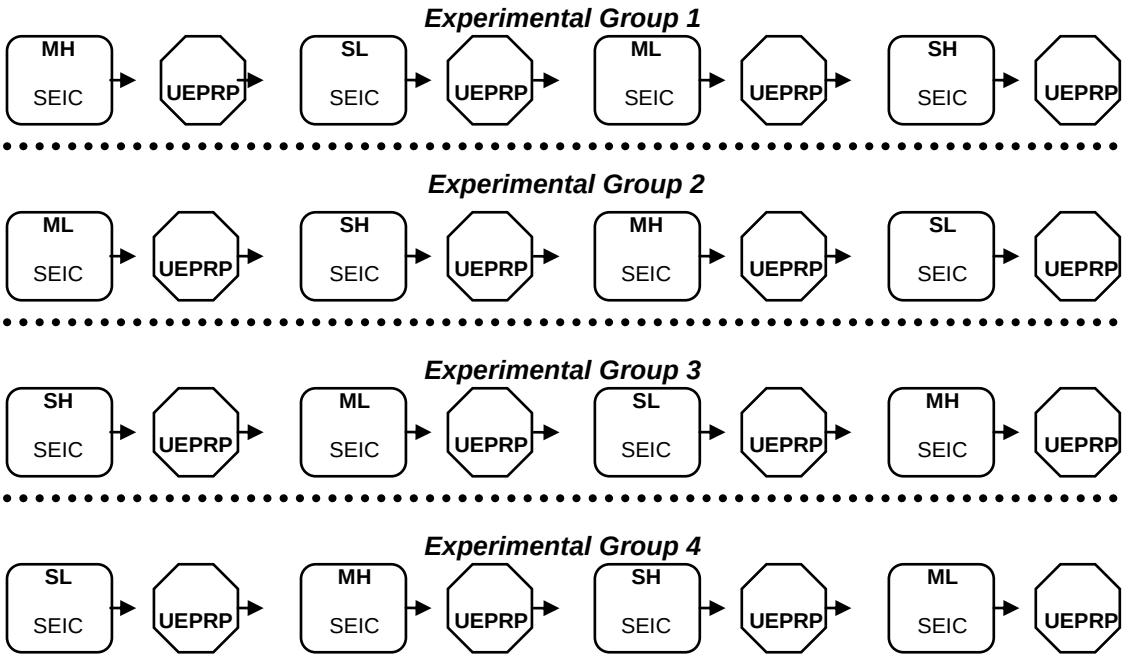
FIGURES

A.1 Structural Design of the Doubly-Multivariate Analysis of Variance

	Session 1			**Session 2**			**Session 3**			**Session 4**		
	M1	**M2**	**M3**	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**
G1	MH	MH	MH	SL	SL	SL	ML	ML	ML	SH	SH	SH
G2	ML	ML	ML	SH	SH	SH	MH	MH	MH	SL	SL	SL
G3	SH	SH	SH	ML	ML	ML	SL	SL	SL	MH	MH	MH
G4	SL	SL	SL	MH	MH	MH	SH	SH	SH	ML	ML	ML

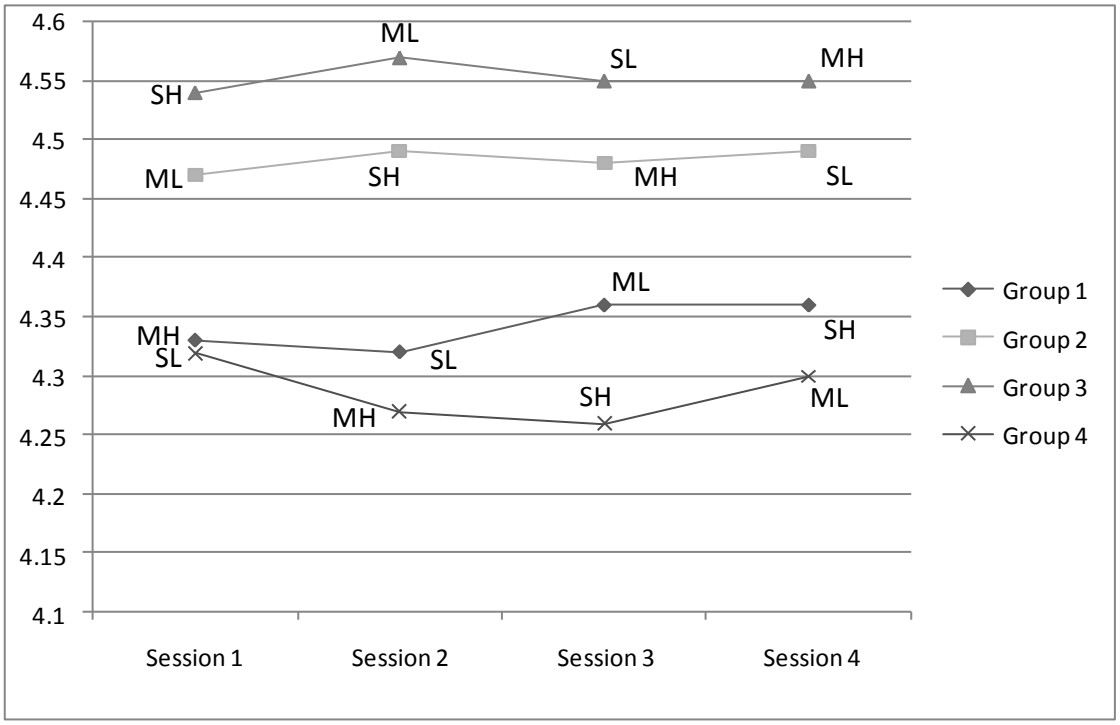
Note: G1 = Experimental Group 1; G2 = Experimental Group 2; G3 = Experimental Group 3; G4 = Experimental Group 4; M1 = Mean Performance Ratings on the SEIC; M2 = Mean Confidence Ratings on the UEPRP; M3 = Mean Satisfaction Ratings on the UEPRP; MH = Multiple-item presentation, high user-flexibility; ML = Multiple-item presentation, low user-flexibility; SH = Single-item presentation, high user-flexibility; SL = Single-item presentation, low user-flexibility.

A.2 Sequence of Rating Formats across the Four Rating Conditions



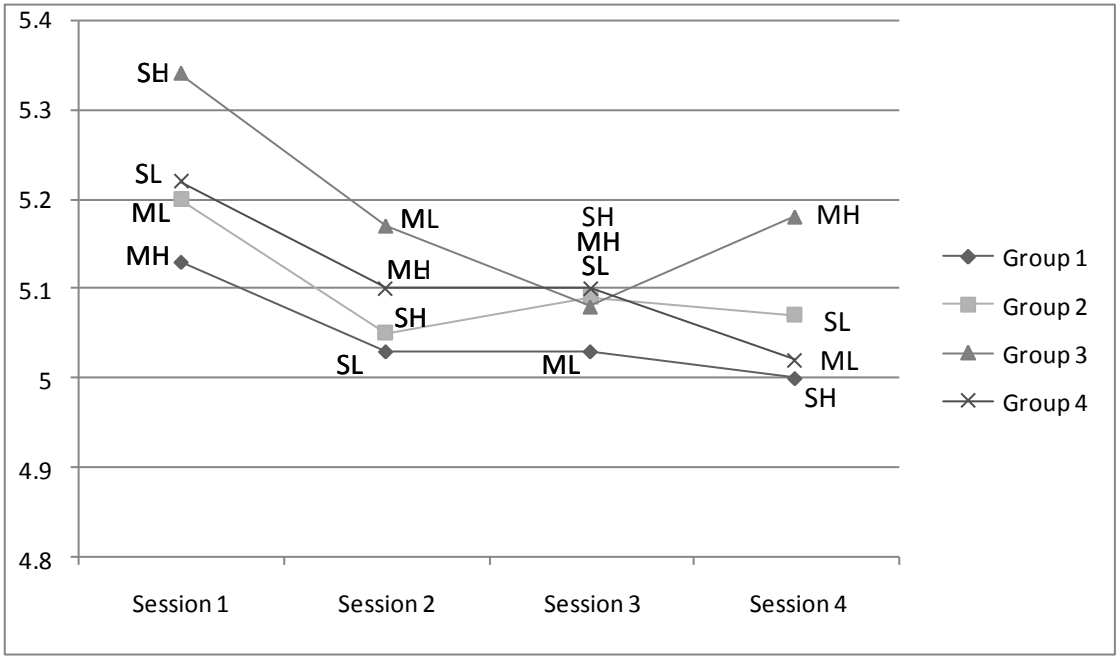
Note: MH = Multiple-item presentation, high user-flexibility; ML = Multiple-item presentation, low user-flexibility; SH = Single-item presentation, high user-flexibility; SL = Single-item presentation, low user-flexibility; SEIC = Student Evaluation of Instructor Competence; UEPRP = User Evaluation of the Performance Rating Process.

A.3 Profiles of Performance Ratings



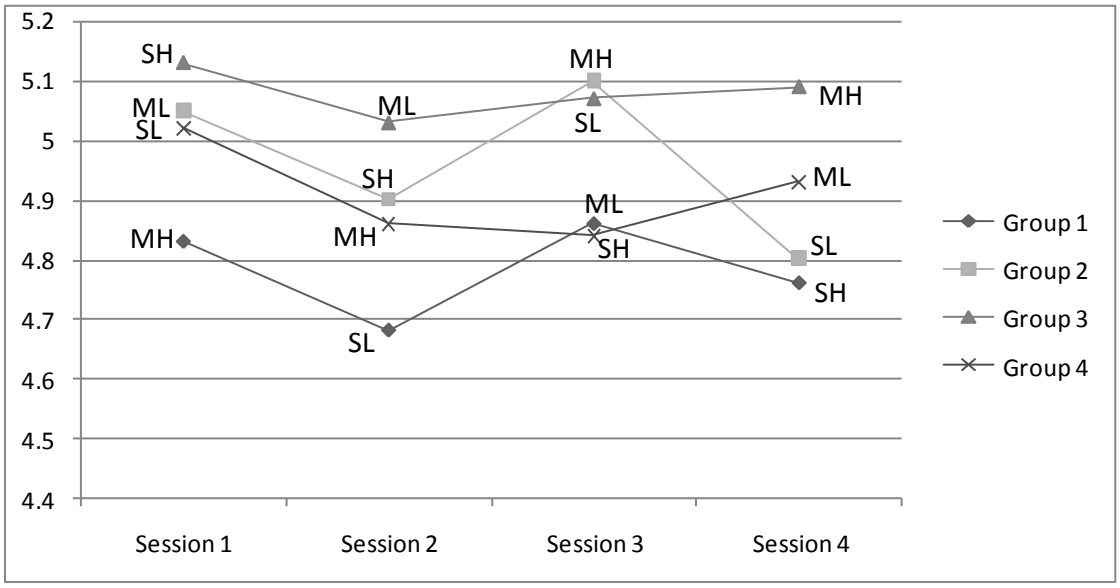
Note: MH = Multiple-item presentation, high user-flexibility; ML = Multiple-item presentation, low user-flexibility; SH = Single-item presentation, high user-flexibility; SL = Single-item presentation, low user-flexibility.

A.4 Profiles of Rater Confidence



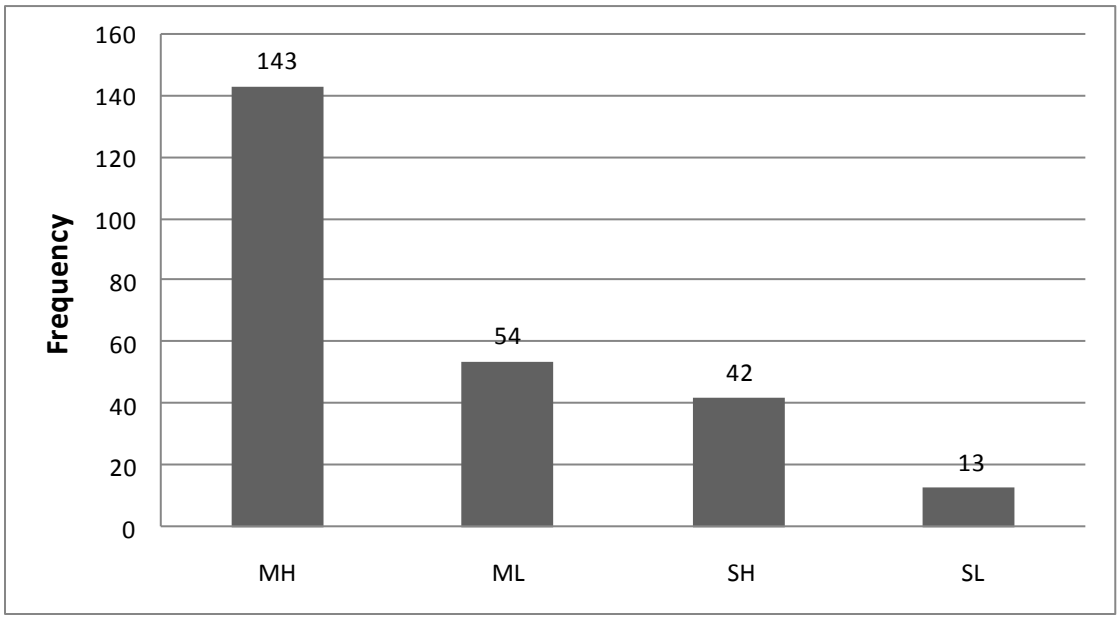
Note: MH = Multiple-item presentation, high user-flexibility; ML = Multiple-item presentation, low user-flexibility; SH = Single-item presentation, high user-flexibility; SL = Single-item presentation, low user-flexibility.

A.5 Profiles of Rater Satisfaction



Note: MH = Multiple-item presentation, high user-flexibility; ML = Multiple-item presentation, low user-flexibility; SH = Single-item presentation, high user-flexibility; SL = Single-item presentation, low user-flexibility.

A.6 Frequency of Preferred Rating Formats



Note: MH = Multiple-item presentation, high user-flexibility; ML = Multiple-item presentation, low user-flexibility; SH = Single-item presentation, high user-flexibility; SL = Single-item presentation, low user-flexibility.

APPENDIX B

TABLES

.

B.1 Summary of Participant Demographics

Variable	Group 1	Group 2	Group 3	Group 4	Total	
					N	%
Age						
18 to 20 years old	35	38	32	36	141	56.0
21 to 23 years old	17	12	21	17	67	26.6
24 to 26 years old	6	10	0	6	22	8.7
27 to 29 years old	1	1	4	1	7	2.8
30 or more years old	2	2	4	4	12	4.8
Missing Data	1	2	0	0	3	1.2
Total:	62	65	61	64	252	100.0
Gender						
Male	16	22	19	25	80	31.7
Female	46	44	42	39	171	67.9
Missing Data	0	1	0	0	1	0.4
Total:	62	65	61	64	252	100.0
Race/Ethnicity						
White/Caucasian American	32	34	23	36	125	49.6
Asian/Asian American	15	11	10	6	42	16.7
Hispanic/Latino American	7	9	13	9	38	15.1
Black/African American	6	9	10	12	37	14.7
Amer. Indian/Native American	0	0	2	0	2	0.8
Other Race/Multiracial	2	2	3	1	8	3.2
Total:	62	65	61	64	252	100.0

B.2 Summary of Statistics of Observed Variables

	Group 1		Group 2		Group 3		Group 4	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Performance Rating Measures:								
SEIC	4.34	.67	4.48	.63	4.55	.55	4.29	.69
User Evaluation Measures:								
UEPRP: Confidence	5.05	.67	5.10	.72	5.19	.76	5.11	.83
UEPRP: Satisfaction	4.78	.77	4.96	.82	5.08	.74	4.91	.80
	Session 1		Session 2		Session 3		Session 4	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Performance Rating Measures:								
SEIC	4.41	.61	4.41	.65	4.41	.67	4.42	.67
User Evaluation Measures:								
UEPRP: Rater Confidence	5.22	.74	5.09	.81	5.08	.79	5.06	.84
UEPRP: Rater Satisfaction	5.01	.81	4.87	.93	4.97	.84	4.89	.95
	MH		ML		SH		SL	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Performance Rating Measures:								
SEIC	4.41	.66	4.42	.65	4.41	.65	4.42	.64
User Evaluation Measures:								
UEPRP: Rater Confidence	5.12	.79	5.10	.79	5.12	.80	5.10	.83
UEPRP: Rater Satisfaction	4.97	.84	4.97	.83	4.91	.94	4.89	.93

Note: MH = Multiple-item presentation, high user-flexibility; ML = Multiple-item presentation, low user-flexibility; SH = Single-item presentation, high user-flexibility; SL = Single-item presentation, low user-flexibility; SEIC = Student Evaluation of Instructor Competence; UEPRP = User Evaluation of the Performance Rating Process.

B.3 Doubly-Multivariate Analysis of Variance

Multivariate Tests	Wilks' Λ	*F*	*df*	Error *df*	*p*	Partial η^2
Between Subjects:						
Group	.954	1.30	9	599	.231	.02
Within Subjects:						
Session	.890	3.31*	9	240	.001	.11
Group*Session	.844	1.56*	27	702	.036	.06

* *Significant at the .05 level.*

B.4 Trend Analysis of the Group by Session Interaction

Variable	Trend	*F*	*df*	Error *df*	*p*	Partial η^2
SEIC:	Linear	0.65	3	248	.585	.01
	Quadratic	2.07	3	248	.105	.02
	Cubic	1.66	3	248	.176	.02
Rater Confidence:	Linear	0.40	3	248	.754	.01
	Quadratic	1.73	3	248	.161	.02
	Cubic	0.79	3	248	.499	.01
Rater Satisfaction:	Linear	0.99	3	248	.396	.01
	Quadratic	2.77	3	248	.042	.03
	Cubic	1.56	3	248	.200	.02

** Significant at the .004 level.*

APPENDIX C

DEPENDENT MEASURES

Student Evaluation of Instructor Competence

DIRECTIONS: This survey is designed to obtain performance ratings from students regarding the level of instruction provided by instructors within the Department of Psychology at the University of Texas at Arlington. These performance ratings will only be used for research purposes and no instructors will have access to the ratings provided their students.

For the purposes of the current study, you are only to provide performance ratings for the primary instructor of the psychology course that you are currently attending. If you are enrolled in more than one psychology course at this time, then please provide performance ratings for the primary instructor of the psychology course where you will apply your research participation credits earned from this study.

Each of the following items contains a behavioral statement that is a characteristic of an effective instructor or ineffective instructor. For each item, we ask that you rate your level of agreement with the behavioral statement where:

1 = STRONGLY DISAGREE
2 = MODERATELY DISAGREE
3 = NEITHER AGREE NOR DISAGREE
4 = MODERATELY AGREE
5 = STRONGLY AGREE
NA = DOES NOT APPLY (not relevant or do not know)

Please rate your level of agreement with the following statements about your instructor:

		SD	MD	N	MA	SA	NA
1.	Communicates ideas and information so that students are able to easily understand.....	1	2	3	4	5	NA
2.	Seems to care less about whether students pass or fail the class.....	1	2	3	4	5	NA
3.	Encourages active participation from students.....	1	2	3	4	5	NA
4.	Demonstrates respect and positive regard towards other instructors.....	1	2	3	4	5	NA
5.	Solicits questions from students to ensure adequate understanding.....	1	2	3	4	5	NA
6.	Responds to questions and comments in a manner that makes others feel understood....	1	2	3	4	5	NA
7.	Recognizes useful relationships between seemingly unrelated information topics.....	1	2	3	4	5	NA
8.	Seeks out opportunities to receive constructive feedback from other instructors..	1	2	3	4	5	NA
9.	Make self available and easily accessible to students.....	1	2	3	4	5	NA

		SD	MD	N	MA	SA	NA
10.	Synthesizes complex information into manageable parts in a systematic way.....	1	2	3	4	5	NA
11.	Rarely makes self accessible to students.....	1	2	3	4	5	NA
12.	Responds to student communications in a timely manner.....	1	2	3	4	5	NA
13.	Communicates information and ideas by appealing to data, facts, and logic.....	1	2	3	4	5	NA
14.	Demonstrates a genuine commitment to helping students succeed.....	1	2	3	4	5	NA
15.	Actively works to keep skills, knowledge, and expertise current.....	1	2	3	4	5	NA
16.	Demonstrates an active curiosity about a broad range of topics.....	1	2	3	4	5	NA
17.	Anticipates emerging problems and/or changing circumstances.....	1	2	3	4	5	NA
18.	Rarely checks to make sure students understand course content.....	1	2	3	4	5	NA
19.	Readily shares relevant or useful information with others.....	1	2	3	4	5	NA
20.	Brings energy and enthusiasm to class discussions.....	1	2	3	4	5	NA
21.	Demonstrates a thorough understanding of course curriculum.....	1	2	3	4	5	NA
22.	Builds positive relationships with other researchers in the field of psychology.....	1	2	3	4	5	NA
23.	Credibly represents knowledge and professional expertise.....	1	2	3	4	5	NA
24.	Makes those who ask questions feel stupid and incapable of understanding.....	1	2	3	4	5	NA
25.	This instructor was an effective teacher.....	1	2	3	4	5	NA
26.	Help was readily available for questions and/or homework outside of class.....	1	2	3	4	5	NA
27.	The instructor was well prepared.....	1	2	3	4	5	NA
28.	The instructor appeared to have thorough knowledge of the subject.....	1	2	3	4	5	NA
29.	The instructor summarized major points.....	1	2	3	4	5	NA
30.	The instructor identified was he/she considered important.....	1	2	3	4	5	NA

		SD	MD	N	MA	SA	NA
31.	The instructor showed interest in, and concern for, the quality of his/her teaching.....	1	2	3	4	5	NA
32.	The instructor kept students informed of their progress.....	1	2	3	4	5	NA
33.	The instructor suggested specific ways students could improve.....	1	2	3	4	5	NA

User Evaluation of the Performance Rating Process

DIRECTIONS: This survey is used to evaluate the performance rating process that you have just completed. The *User Evaluation of the Performance Rating Process* is an 11-item survey divided into two sections: (1) Confidence in Performance Ratings (3 items), and (2) Satisfaction with Performance Rating Process (3 items). To complete the survey, simply follow the instructions for each section.

Please rate your level of agreement with the following statements where:

1 = STRONGLY DISAGREE
2 = DISAGREE
3 = SLIGHTLY DISAGREE
4 = SLIGHTLY AGREE
5 = AGREE
6 = STRONGLY AGREE

Confidence in Performance Ratings:

	SD	D	SD	SA	A	SA
1. I am very confident in the performance ratings that I assigned to my instructor.....	1	2	3	4	5	6
2. It was very difficult for me to decide the performance ratings that I assigned to my instructor.....	1	2	3	4	5	6
3. I would not change any of the performance ratings that I assigned to my instructor.....	1	2	3	4	5	6

Satisfaction with Performance Rating Process:

	SD	D	SD	SA	A	SA
4. I am very satisfied with the performance ratings that I assigned to my instructor.....	1	2	3	4	5	6
5. I would not change anything about this performance rating process.....	1	2	3	4	5	6
6. I would endorse this rating method as an appropriate tool for measuring my own performance.....	1	2	3	4	5	6

Note: An optional comment box is located at the end of the online version of this survey to allow participants to provide additional feedback on the performance rating process.

Supplemental Demographic Questionnaire

1. I entered my first year at the University of Texas at Arlington as a:

- ❑ Freshman
- ❑ Sophomore
- ❑ Junior
- ❑ Senior
- ❑ Other: *Please specify:* _____

2. I have attended _____ of the classes (to date) for the instructor that I evaluated in this study:

- ❑ More than 75%
- ❑ 50% to 75%
- ❑ Less than 50%

3. I expect to receive the following grade in this course:

- ❑ A
- ❑ B
- ❑ C
- ❑ D
- ❑ F
- ❑ Incomplete

4. Please select the range that includes your current GPA at the University of Texas at Arlington:

- ❑ 3.1 to 4.0
- ❑ 2.1 to 3.0
- ❑ 1.1 to 2.0
- ❑ 0.0 to 1.0
- ❑ Other: *Please specify:* _____

5. Please select the range that corresponds to your current age:

- ❑ 18 to 20
- ❑ 21 to 23
- ❑ 24 to 26
- ❑ 27 to 29
- ❑ 30 or older

6. Please select the response option that most clearly identifies your gender:

- ❑ Male
- ❑ Female

7. Please check the response option that most clearly identifies your racial/ethnic background:

- ❑ American Indian / Native American
- ❑ Asian / Asian American
- ❑ Black / African American
- ❑ Hispanic / Latino American
- ❑ White / Caucasian American
- ❑ Other. *Please specify:* _____

Please rate your level of agreement with the following statements pertaining to your level of confidence in computers:

		SD	D	SD	SA	A	SA
1.	I am sure I could learn a computer language.....	1	2	3	4	5	6
2.	Generally, I would feel OK about trying a new problem on a computer.....	1	2	3	4	5	6
3.	I'm not the type to do well with computer.....	1	2	3	4	5	6
4.	I do not think I could handle a computer course.....	1	2	3	4	5	6
5.	I think using a computer would be very hard for me.....	1	2	3	4	5	6
6.	I could get good grades I computer courses.....	1	2	3	4	5	6
7.	I don't think I would do advanced computer work.....	1	2	3	4	5	6
8.	I'm no good with computers.....	1	2	3	4	5	6
9.	I am sure I could do work with computers.....	1	2	3	4	5	6
10.	I have a lot of self-confidence when it comes to working with computers.....	1	2	3	4	5	6

REFERENCES

Bartram, D. (2006). Testing on the Internet: Issues, challenges and opportunities in the field of occupational assessment. In D. Bartram & R. K. Hambleton (Eds.), *Computer-based testing and the Internet: Issues and advances* (pp. 13-37). New York, NY: John Wiley & Son Ltd.

Benjamin, L., Cavell, T. A., & Shallenberger, W. R., III (1984). Staying with initial answer on objective tests: Is it a myth? *Teaching of Psychology*, *11*, 133-141.

Borman, W. C., Buck, D. E., Hanson, M. A., Motowildo, S. J., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology*, *86*, 965-973.

Borman, W. C., Hanson, M. A., Motowildo, Drasgow, F., Foster, L., & Kubisiak, U. C. (1998, April). *Computerized adaptive rating scales that measure contextual performance*. Paper presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Carlson, R. (1971). Where is the person in personality research? *Psychological Bulletin*, *75*, 203- 219.

Carlson, R. (1984). What's social about social psychology? Where's the person in personality research? *Journal of Personality and Social Psychology*, *47*, 1304-1309.

Cascio, W. F. (1995). Wither industrial and organizational psychology in a changing world of work? *American Psychologist*, *50*(11), 928-939.

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance*, *10*, 331-360.

Cozby, P. C. (2000). *Methods in Behavioral Research* (7th ed.). Mountain View, CA: Mayfield, Publishing.

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.

Drasgow, F., & Mattern, K. (2006). New tests and new items: Opportunities and issues. In D. Bartram & R. K. Hambleton (Eds.), *Computer-based testing and the Internet: Issues and advances* (pp. 59-75). New York, NY: John Wiley & Sons Ltd.

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.

Greguras, G. J., & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology*, *83*, 960-968.

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.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, *87*, 72-107.

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

Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing*, *6*, 1-24.

Longenecker, C., & Ludwig, D. (1990). Ethical dilemmas in performance appraisals revisited. *Journal of Business Ethics*, *9*, 961-969.

Loyd, B., & Gressard, C. (1984). Reliability and Factorial Validity of Computer Attitude Scales. *Educational and Psychological Measurement*, *44*, 501-505.

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.

McAdams, D. P. (1997). A conceptual history of personality psychology. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 3-39). San Diego, CA: Academic Press.

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

Moreno, K. E., Lee, J. A., & Sympson, J. B. (1985, July). *Effects of computerized versus paper-and-pencil test administration on examinee performance.* (Available from Navy Personnel Research and Development Center, San Diego, CA 92152-6800).

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Rosenfeld, P., Booth-Kewley, S., & Edwards, J. E. (1993). Computer-administered surveys in organizational settings. *American Behavioral Scientist*, *36*, 485-511.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review*, *16*, 155-180.

Salopek, J. J. (2004). Rethinking likert. *Training + Development 58*(9), 26-29.

Schneider, R. J., Goff, M., Anderson, S., & Borman, W. C. (2003). Computerized adaptive rating scales for measuring managerial performance. *International Journal of Selection and Assessment*, *11*, 237-246.

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Thousand Oaks, CA: Sage.

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

Tabachnick, B., & Fidell, L. (2007). *Using Multivariate Statistics*. (5th ed.). Boston, MA: Pearson Education, Inc.

Terman, L. M. (1917). A trial of mental and pedagogical tests in a civil service examination for policemen and firemen. *Journal of Applied Psychology*, *1*, 17-29.

Toegel, G., & Conger, J. A. (2003). 360-degree assessment: Time for reinvention. *Academy of Management Learning and Education*, *2*, 297-311.

Tziner, A., & Kopelman, R. E. (2002). Is there a preferred performance rating format? A non-psychometric perspective. *Applied Psychology: An International Review*, *51*, 479-503.

Vispoel, W. P. (1998). Reviewing and changing answers on computer-adaptive and self-adaptive vocabulary tests. *Journal of Educational Measurement*, *35*, 328-347.

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

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., et al. (2000). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Erlbaum.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*, 361-375.

BIOGRAPHICAL INFORMATION

Ryan Phillips graduated from Auburn University in 2003 with a Bachelor of Arts degree in Psychology. After working for three years in a range of different career fields, he began his master's degree work in Industrial and Organizational Psychology at the University of Texas at Arlington in the Fall of 2006. Since 2005, he has worked as an Associate Consultant for Leadership Worth Following, LLC. He graduated with his M.S. in August of 2008. Mr. Phillips is an affiliate member of both the Society for Industrial and Organizational Psychology (SIOP) and the Dallas Area Industrial and Organizational Psychologists (DAIOP).