University of Denver

Digital Commons @ DU

Electronic Theses and Dissertations

Graduate Studies

1-1-2010

Item Order Effects on Attitude Measures

Pei-Hua Chen
University of Denver

ITEM ORDER EFFECTS ON ATTITUDE MEASURES

A Dissertation

Presented to

The Morgridge College of Education

University of Denver

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Pei-Hua Chen

November 2010

Advisor: Dr. Kathy E. Green

Author: Pei-Hua Chen
Title: ITEM ORDER EFFECTS ON ATTITUDE MEASURES
Advisors: Dr. Kathy E. Green
Degree Date: November 2010

Abstract

The purpose of this dissertation was to examine the effects of altered item order on attitude measures for both computerized adaptive and conventional survey formats. Based on items modified from a dissertation/thesis completion survey (Green & Kluever, 1997) with three scales, three survey versions were generated with items ordered by difficulty as hard-to-easy (H-E), easy-to-hard (E-H), and five medium trait level items presented first followed by randomly ordered items (M-R) for conventional survey format. Significant differences in item difficulty and item discrimination were found for two of the three scales. Differences in scale reliability were detected for the procrastination and responsibility scales. Also, significant correlations between scale total score and scale attitude strength were discovered with each survey version.

Further, two computerized adaptive survey version were generated. One began with items at medium and the other at extremely high trait levels. Results showed significant differences in number of items administered to achieve a set level of precision for two scales and significant differences in reaction time were found for one scale between the two versions. The version of item starting at the extreme trait level required more items, and took longer to respond to. Further, significant differences in the estimated person parameter were found for one scale between the two survey versions. Based on the results of both survey formats indicating item order effects pose a problem for assessing attitude.

Acknowledgement

I would like to thank Dr. Edith King and Dr. Donald Bacon for their support during the dissertation process. Thank you to Dr. Duan Zhang and Dr. Thomas Obremski for serving on my committee.

I am deeply appreciative of my advisor, Dr. Kathy E. Green, for her intellectual prowess and enthusiasm gave me the motivation and courage to keep on keeping on.

Table of Contents

Chapter One: Introduction 1

Chapter Two: Review of the Literature ……………………..……………………….…..5

Research on Attitude Change.. 5

Theories of Item Order Effects 9

Norm of Evenhandedness ……………………………………………………….…9

Addition Effect.. 10

Subtraction Effect 10

Anchoring and Adjustment.. 11

Attitude Accessibility.. 15

Reaserch on Item Order Effects.. 17

Achievement Test ………………………….……..………………….. 17

Aptitude Test ……………...……………………………………….. 25

Attitude Measures ………………………………………………… 26

Adaptive Testing ……………………………………………………………….. 28

Problems with Conventional Tests 28

Principles of Adaptive Testing.. 29

Adaptive Testing Based on Item Response Theory (IRT).................................... 31

Adaptive Testing and Computers.. 32

Item Order Effects in CAT 37

Research Questions... 42

Chapter Three: Method.. 46

Research Procedure ………………………………………………………….. 46

Phase One.. 47

Participants.. 47

Instrument 48

Procedure 52

Analysis.. 53

Phase Two.. 53

Participants.. 53

Instrument 53

Procedure 54

Analysis.. 54

Phase Three.. 56

Participants 56
Instrument 57
Procedure 57
Analysis 58
Phase Four 58
Participants 58
Instrument 59
Procedure 59
Analysis 62

Chapter Four: Results 63
Phase One 63
Phase Two 64
Phase Three 64
Reliability 65
Test Score 66
Item Difficulty 66
Item Discrimination 69
Correlations Between Perception of Effects of Item Order, Scale Test Score, and Scale Attitude Strength 70
Phase Four 72
Test Length 73
Reaction Time 74
Test Score 74
Person Parameter 74

Chapter Five: Discussion 76
Conventional Survey 77
Reliability 78
Item Difficulty 79
Item Discrimination 81
Test Score 82
Correlations Between Scale Test Score, and Scale Attitude Strength 83
Computerized Adaptive Survey 85
Person Parameter 87
Test Length and Test Score 79
Reaction Time 88
Limitations and Furture Study 89
Conclusion 91

References 95

Appendix A 110

Appendix B 118

List of Tables

Table Page

1 Test Differences, Reliabilities, Discrimination Values, and Significance Test Results 18

2 The Effect of Item Rearrangement on Test Reliability and Test Scores 19

3 Summary of Group Means, Standard Error of the Mean Differences, and T-Test for the Item Rating and Subtest Scores 21

4 Summary of Item Order Effects on Test Scores, Item Statistics, and Reliability..... 23

5 Summary of Group Means and T-Test by Item Order........ 26

6 Summary of T-Tests and Effect Sizes on Two Job Satisfaction Measures 27

7 Summary Statistics of Three Survey Versions 65

8 Item Difficulty and Item Discrimination of Procrastination Scale for Each Survey Version........ 67

9 Item Difficulty and Item Discrimination of Dissertation Barriers Scale for Each Survey Version........ 68

10 Item Difficulty and Item Discrimination of Responsibility Scale for Each Survey Version........ 69

11 Correlations of Item Discrimination Between Two-Half Samples for Each Scale 69

12 Correlations Between Scale Score and Scale Attitude Strength in Hard-to-Easy Version Survey 71

13 Correlations Between Scale Score and Scale Attitude Strength in Easy-to-Hard Version Survey 71

14 Correlations Between Scale Score and Scale Attitude Strength in Medium-then-Random Version Survey 72

15 Summary Statistics of Two Computerized Adaptive Survey Versions 73

16 t-Test Differences Between Two Computerized Adaptive Survey Versions 75

List of Figures

Figure Page

1 Anchor-and-Adjust Judgment Heuristic Flowchart. 14

2 Flowchart of an Adaptive Test. 34

Chapter One: Introduction

Paper-and-pencil tests ruled the measurement field for a long time. However, to standardize administration of paper-and-pencil tests, test takers need to take the same exam at the same day, place, and time no matter the level of ability or position on the trait. In the late 1980s, the personal computer was introduced, and the format of testing shifted to delivery via computer which made tests more flexible. For example, test takers can take exams whenever they are ready. Also, the statistical accuracy of test scores can be enhanced. Computerized adaptive testing (CAT) was developed and resulted in improving method and economy in the field of psychological assessment. During the period of testing, the examinee's ability or trait level is iteratively estimated based on answers to present items (Ortner, 2008; Van der Linden & Glas, 2000).

The idea of computer adaptive testing (CAT) is based on item response theory (IRT) which aims to look at the underlying trait producing the test performance. The key feature of IRT is that the examinee's ability estimate is independent of particular items used, and item values are independent of examinees. Distinct parameter estimates for items and examinees are generated which can easily be used to identify misfitting items and persons. Presently, there are numerous studies emphasizing methodological improvements to CAT. Most of these studies focus on the use of CAT with achievement and personality tests

(Bergstrom, Lunz & Gershon, 1992; Ortner, 2008). The usefulness of CAT for attitude assessment presumes that order effects are trivial or nonexistent.

Attitude measures are used to collect self-report data by using rating scales or selecting one of several alternatives when researchers want to know people’s attitude toward a person, issue, event, or product (Wittenbrink & Schwarz, 2007). For several decades, most research with attitude measures attempted to understand the mental operations leading to responses to attitude items, such as response processes (Nosek & Banaji, 2001), priming procedures (Wittenbrink & Schwarz, 2007), and brain activity (Cacioppo, Gardner & Berntson, 1997). Studies that focused on the relationship between attitude and context are few in number (Schuman & Presser, 1981), especially for CAT.

Investigations of context effect are centered on the areas of anchoring and adjusting and item order. The idea of anchoring and adjusting was first proposed by Tversky and Kahneman (1974) who stated that people tend to use the information of prior items to adjust their responses to subsequent items. Zhao & Linderholm (2008) found that people may provide different estimates based on the information or stimulus of preceding items. That is, they anchor their attitude and adjust their answers based on that anchor. For example, Tversky and Kahneman (1974) asked people to estimate the age at which Gandhi died. People must first decide whether Gandhi died before or after the age of 9 or 140. They found that if people decided Gandhi died before the age of 140, they estimated Gandhi lived roughly 67 years. But if people decided Gandhi died after the age of 9, they estimated Gandhi lived roughly only 50 years. Actually, he died when he was 79 years old. According to this study, item order may be a factor which affects response. This claim was also supported by Hambleton and Traub (1974), and Flaugher, Melton, and Myers (1968) who also discovered that item presentation order has statistically significant effects on test performance.

To date, studies of anchoring-and-adjusting and item order effects are focused on achievement tests. Research on anchoring-and-adjusting effects or item order effects in attitude measurement is restricted. Only a few studies investigated the effects of changed item orders in attitude measurement, especially for CAT. The most essential advantage of CAT is that every examinee can have different orders of items based on their performance on the present item. In this case, if item order is really a factor affecting performance on attitude tests then the merit of CAT turns out to be a defect. If item order affects response to attitude items, it is questionable to apply CAT with attitude tests.

The issue of item order in attitude tests with both conventional and CAT formats has not yet received much attention. This is an important topic for a number of interrelated reasons. First, the question about whether examinees are taking equivalent tests with rearranged orders of items should be taken into consideration. It is essential for test developers to think about the quality and equivalence of the measures. Second, if test performance is affected by the sequence of items, does CAT estimate the identical latent traits or abilities of examinees who take the test with different item orders? If not, this countermands the superiority of CAT. Third, for the conventional attitude measure, long paragraphs of written description are presented sometimes for attitude or judgment measures which put a heavy verbal load on the tests. The benefit of applying CAT is that different kinds of items, such as graphs and video clips, can be presented easily as the stem of an item (Green, 1982). This benefit is null if order effects exist. Further, Fazio (1990) stated that examiners need to spend more time when confronting extreme items. The reaction time for items may be affected by some certain orders of items. Test takers may change their responses based on the level of item difficulties which may also influence the test length. In this case, test length and response time may be affected by item presentation order.

In this study, the effect of item order on attitude measures was explored. Test performance with different sequences of items was compared, as were differences in item discrimination and difficulties, and test reliabilities with testing beginning with easy, medium, or extremely high trait levels. Relationship between test score and scale attitude strength was also assessed. Correlation between participants' perception of whether their answers were influenced by the item order and scale test score was examined. Then, an exploratory study of attitude measure via CAT with items starting with medium or extreme trait levels was also conducted. The differences in test scores, test reliability, item discrimination, test length, and reaction time were assessed.

Chapter Two: Review of the Literature

The research topics of attitude, item order/context effect, and computerized adaptive testing are reviewed. First, research regarding attitude change is summarized. Types of item order effects are addressed next. Then, the discovery of and research on item order effects is summarized. Following a review of adaptive testing, a summary of studies of item order effects with CAT is presented.

Research on Attitude Change

The earliest definition of attitude was proposed by Gordon Allport who defined attitude as “A mental and neural state of readiness, organized through experience, exerting a directive or dynamic influence upon the individual’s response to all objects and situations with which it is related” (Halloran, 1970, p. 14). Early research was focused on investigating processes of attitude formation. At that time, attitude was viewed as an important concept in social psychology as attitudes can be learned and are dynamic. After several decades, different concepts, such as relationships with memory, beliefs, and behavior were introduced in forming new definitions of attitude. One of the more current definitions is that attitude is a “psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor” (p. 1) proposed by Eagly and Chaiken (1993). Studies of attitude shifted to explore factors which affect attitude stability rather than definition. One factor that has been found to affect attitude response is item presentation order.

Extensive research has focused on reasons why attitude changes. Halloran (1970) proposed that attitude change occurs when a new message is related to the individual's needs which can be reinforced by related events. People may change their attitude if they perceive new message to be trustworthy (Cohen, 1964). Thus, attitude is not only affected by the message, but also the way information is presented and its form. For instance, people may have different attitudes toward a topic by reading about or discussing it.

Questionnaires are still an essential method used to assess attitude. Theoretically, when two different questionnaires are used to measure attitude about the same topic, both measures should generate the same outcome. Apart from mental state (e.g., motivation, self-esteem, and confidence), Cohen (1964) stated that if different results appear, it is possible differences are due to context or item order. Item order is the explanation used most often to interpret unexpected test findings.

Attitude measures are used to detect people's dispositions toward the specific topic. Some questions ask people to rate their feelings about an attitude object by retrospective reflection on events or experiences, some ask test takers to make judgments about it. In attitude questionnaires, items are usually similar in content in order to assess varied facets of the disposition. Similar items may interact with each other. Also, the item order can influence the results. This instability in results makes the outcome of attitude trend studies suspect (Schuman & Presser, 1981). Research on item order effects has been conducted in different fields (e.g., marketing, education, and medical science). If item order impacts the results, not only the disposition of attitude but also the accuracy of judgments or diagnoses all face severe challenge. Measurement results are suspect. Crano (1977) found that the history of study of order effects began in 1925 with Lund's study which indicated that the first two opposite pieces of information affected subjects' attitudes more with controversial topics than with non-controversial topics. This was the first study investigating the relationship between presentation order and attitude. In this study, Lund brought up the terms primacy and recency effects. Primacy effects indicate that an individual's opinion is impacted by the message presented first, but recency effect occurs when a person's opinion is impact by the later presented information.

Anderson and Jacobson (1965) concluded that under two conditions the primacy effect may occur; first, inconsistency discounting --- when the later description is inconsistent with the former one, and second, intention decrement --- people decrease attention when processing a series of information. The earlier message influences the result or judgment more than later ones. For example, Bossart and Di Vesta (1966) recruited college students to rate impressions about little-known people who were described by sets of adjectives. The descriptions were presented in two different orders: positive adjectives first then negative ones, and negative adjectives then positive ones. A statistically significant order effect occurred. Impressions tend to be positive when the positive adjectives were presented first. Students had more negative impression when negative adjectives were presented first which indicated that the impression ratings were influenced by primacy. Stewart (1965) observed college student's ratings of personality impression by distributing high- and low-rated likableness adjectives to stimuli. The primacy effect occurred when responses were made only after all adjectives were presented. The recency effect was induced when responses were made after each set of adjectives.

Payne, Bettman, and Johnson (1990) proposed one possible explanation of this phenomenon: when people need to process many pieces of information, responses made only at the end of the presentations increases the task complexity. In this situation, people cannot use comprehensive strategies, and resort to strategies which can ease cognitive strain. Therefore, the primacy effect appears to simplify the choice problem. However, if responses can be made after each piece of information, these short series of messages let people think more deliberately and induce the recency effect. This idea was accepted by later researchers, Hogarth and Einhorn (1992), who indicated that people tire if asked to process several series of information, and people become less sensitive to later messages. They also found that primacy effects occur when subjects report their opinions after all the information has been presented. On the other hand, if subjects express their opinions after each piece of information is presented the recency effect is induced.

To sum up, the concept of primacy and recency was the first idea to explain the item order effect. According to Leary and Dorans (1985), they found that research on effects of item order and context on test performance was presented around the 1950s. Following the years of World War II, the improvement of several important changes were introduced in educational and psychological testing, such as the improvement of the computer, and the development of statistical analyses which changed from abstract theorems of mathematics to more efficient and effective computational techniques. During the 1950s to 1960s, research was focused on investigating the simple main effect of item order on test performance. Researchers were motivated to understand tests using new technology and resources. In the late 1960s, studies focused more on the effects of test taker biological and psychological characteristics on test performance (p. 387-389). Studies emphasized detecting the interaction between factors like anxiety level or time pressure and item order on achievement tests (e.g., Marso, 1970). Subsequently, adaptive testing was introduced, and investigations moved to detect the stability of item parameters by changing item orders (e.g., Whitely & Dawis, 1976).

Reasons for item order effects on test scores are still undetermined. The discussion in the literature gradually moved from the concept of primacy and recency on item order effects to anchoring and adjusting.

Theories of Item Order Effects

One of the most efficient and common forms to understand attitude toward topics is through the use of surveys. Whether item order is a factor affecting responses has been a question for survey researchers for a long time. Dillman (2000) concluded in several situations that responses to subsequent items may be altered depending on which items immediately prior to it. Situations which may evoke the item order effects are addressed by the following.

Norm of Evenhandedness. People tend to adjust answers based on the value of their previous answer. A norm of fairness or evenhandedness makes task taker responses to the following question balance his or her answer to the previous one. For example, Sangster (1993) found that 34% of students agreed that students should be expelled if they plagiarized when the preceding item asked about whether a professor should be fired if he plagiarized. However, only 21% of students proposed that students should be expelled if the question about the student was asked first. This phenomenon of using the value of the former answer to adjust the response to the following answer is also called the value-based effect.

In 1988, Bishop, Hippler, Schwarz, and Strack found that the phenomenon of the norm of evenhandedness appears only with telephone interviews and not with self-administered surveys. In 1995, Schwarz and Hippler examined the norm of evenhandedness effect by administering telephone and mail surveys. The same result was found, with the order effect presented only in telephone but not in mail surveys. They concluded that is because respondents can look ahead to see what is going to be asked in the self-administered questionnaire, and adjust their answers to earlier questions. However, other studies found that this effect is similar in both mail and telephone surveys (e.g., Ayida & McClendon, 1990; Sangster, 1993).

Addition Effect. Schwarz, Strack, and Mai (1991) proposed that "when a specific question precedes a general question, and these two items are not assigned in the same context, respondents use the information primed by the specific question to form the general judgment" (p. 3). For instance, Schuman and Presser (1981) found that when a general question like "How would you say things are these days?" was asked after a specific question like "How would you describe your marriage?", more respondents tended to say very happy to the general question in comparing with the reversed order. They concluded that this phenomenon was because people tend to think about the specific question when answering the general question (Dillman, 2000).

Subtraction Effect. This is the opposite of the addition effect. People may subtract out the reasons that they use to answer the first question to adjust their response to the second item. For example, Mason, Carlson, and Tourangeau (1994) found that when the general question (How do you feel about the economic situation in your state over the next five years?) was presented prior to the specific question (How would you describe the economic situation in your community over the next five years?), there were 7-10% more people who said the state economy is getting better than when questions were presented in the reverse order. According to these results, they concluded that people may subtract the information on which the first item was based. In 1995, Willits and Saltiel concluded that a lower score was found on the summary or general question when it was asked before specific questions. Schwarz (1996) also proposed that people tend to take into account what they have already answered to adjust their following responses.

Anchoring and Adjustment. Research on attitudes could be divided into two positions. One proposed that attitudes are stable. Once stored in memory, they come to mind automatically. The other proposed that attitudes are labile and sensitive to context. In this perspective, researchers found that when people are asked about reasons for their dispositions, their responses are influence by the degree of how easily the information can be accessed. The easier obtained and verbalized information is more likely to be used to construct a new attitude. The resolution of these two contradictory positions is formed in the idea of anchoring-and-adjusting models of attitude change (Wilson, Lindsey, & Shooler, 2000, p. 102).

Anchoring and adjustment was first proposed by Tversky and Kahneman (1974), who asked students what percentages of African countries had joined the United Nations. Before answering the question, students were requested to make a judgment about whether the percentage was greater or less than a number found by spinning a wheel (numbers from 1 to 100). When the number selected by the wheel was 10, subjects gave an average estimate of 25%. The estimate was 45% when the number selected by the wheel was 65. According to these results, they concluded that under conditions of uncertainty, the former messages (10 and 65) served as anchors, even if the information was apparently arbitrary. People tend to anchor based on information first presented then to adjust based on their anchor to generate a plausible final estimate (Zhao & Linderholm, 2008, p. 197). Research in anchoring and adjustment can be found in different fields, such as, lie detection (Zucherman, Koestner, Colella, & Alton, 1984), marketing competition in buying new products (Green, Tull, & Albaum, 1988), and behavior prediction (Davis, 1986). Most of these studies presented a robust impact of anchoring.

In 1988, Tourangeau and Rasinski summarized the processes of answering attitude questions. First, people base responses on an interpretation of what the attitude is about. The semantics of questions is important at this stage. If the question presents precise semantics which matches the anchor, the anchoring effect will emerge (Bishara, 2005). Second, people retrieve relevant memories, beliefs, or feelings toward this attitude. At this stage, people generally recall the overall attitude structure then retrieve details about it. In recalling the overall structure, familiarity with the topic and accessibility of the information are influential factors. Depending on the context of questions, there are three ways to arouse memory: (a) free-recall: the context provides only something that was experienced with particular time or place. (b) cued-recall: more detailed information is provided by context for memory searching. (c) recognition: the item itself provides cues for recall (Tourangeau & Rasinski, 1988, p. 303). Then, judgment is made according to the retrieved information. Finally, people select a response which best fits the judgment.

Context affects the interpretation of attitude measurement because prior items serve as the anchor. According to Strack, Schwarz, and Gschneidinger’s (1985) research which asked participants to rate their current life satisfaction based on their past personal experiences, respondents tended to rate themselves as unhappy if they recalled more positive past events, and those who recalled more negative past experiences tended to rate their current life as more positive. The past life experiences served as anchors used to compare with their current life. However, a follow-up study argued that if respondents can recall their past experiences in detail and vividly, the former events served as carryover, and ratings were influenced by moods.

This result was supported by Harrison and McLaughlin (1993) who concluded that prior responses to items are anchors for subsequent responses. Carryover appears when the interpretation or responses to prior items are embedded in an easily accessible cognition. Respondents can retrieve feelings when encountering relevant attitudinal cues. Response to attitudinal questions is cognitively represented in memory and is activated by appropriate cues of the related feelings or events (Cohen & Reed II, 2006). Therefore, different attitude dispositions will occur if the topic questions are introduced with different passages or items.

Hastie and Dawes (2001) proposed a flowchart which depicts the process of anchoring and adjustment (Figure 1). When faced with an uncertain situation, respondents tend to search their memory or evidence based on prior questions. Then, information is extracted from the most important evidence to determine whether the current message is redundant or not, and according to the anchor information, adjustments are made to subsequent answers.

This phenomenon attracted the attention of researchers interested in attitude accessibility and its effects on attitude change. Recently, investigators have begun to study whether temporarily salient or accessible information affects the retrieval process.

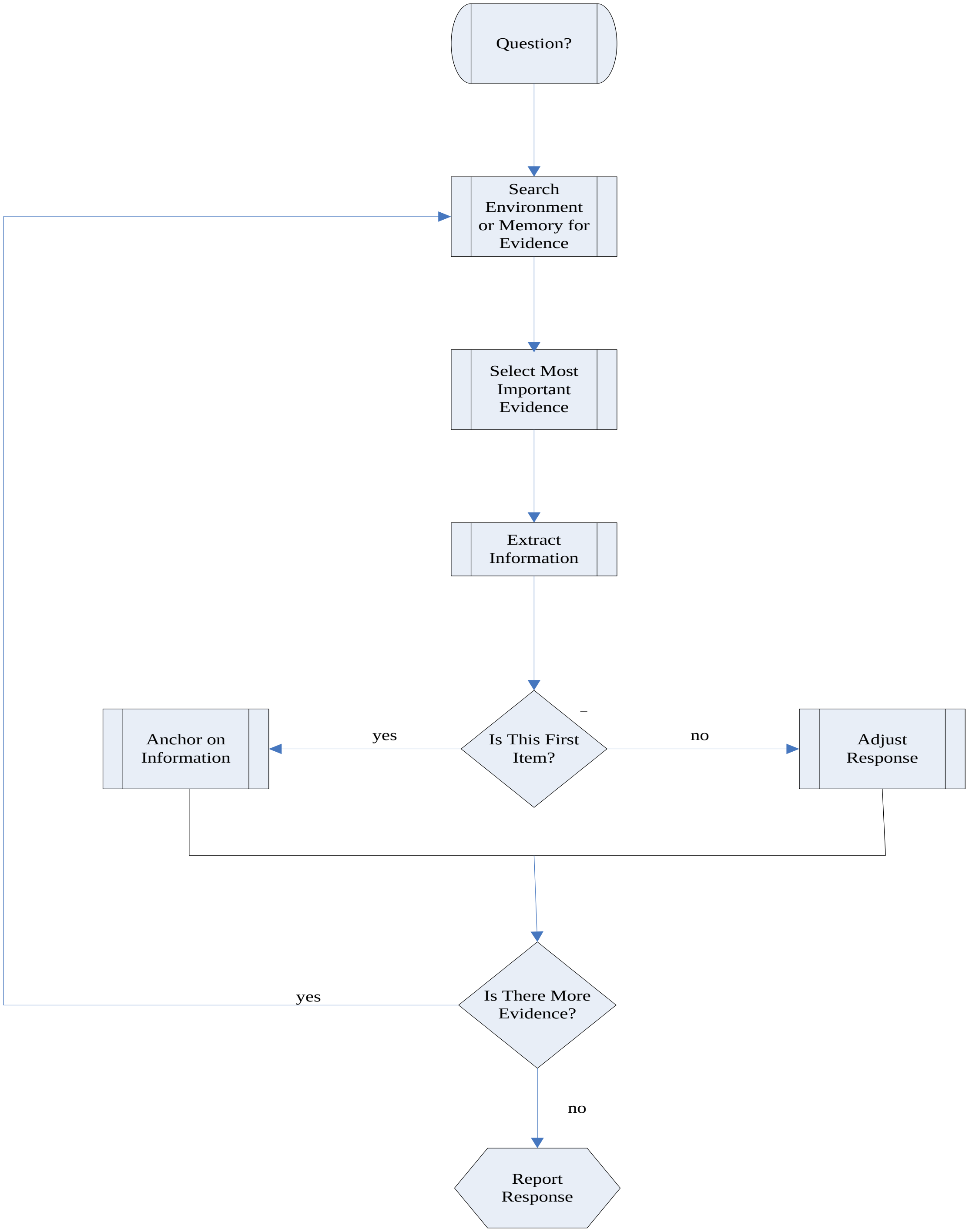


Figure 1. Anchor-and-Adjust Judgment Heuristic Flowchart from Hastie and Dawes (2001).

Attitude Accessibility

Studies have shown that attitude change depends on the cognitive capacity to retrieve information and the accessibility of attitudinal cues (Cohen & Reed II, 2006; Lynch, 2006). Lavine, Huff, Wagner, and Sweeney (1998) stated that “people tend to oversample from whatever information is momentarily salient or accessible” (p. 359). The internal retrospection process and external context are two elements that impact people’s attitudinal responses.

According to an anchoring-and-adjustment approach, when people confront an attitude object, the stored evaluation of this object comes to mind automatically. Then, people might use the currently accessible information, such as the context of questions, to adjust their attitude. For this phenomenon, it was hypothesized that changes in the accessibility of the relevant topic in memory results in survey context effects (Tourangeau, Rasinski, Bradburn, & D’Andrade, 2001, p. 403). People may search their memory for a preexisting evaluation of the attitude issue when they encounter relevant questions. However, this kind of search is not based on a systematic process, but based on a quick sampling of the relevant beliefs.

People have a large and complex belief structure on several issues, but only a small part of their beliefs about a topic is sampled when they have time pressure in answering a survey (Tourangeau et al., 2001, p. 403). Therefore, when the strength of these pieces of stored information is different, varied responses occur based on the weight of the information received (Tourangeau, Rasinski, & Bradburn,1989). Fazio (1990) found that the more accessible the information, the more likely the previous attitude is to be activated. On the other hand, if the attitude is inaccessible, people tend to consider current feelings or thoughts deliberatively which affects their responses (Park, Levine, Westerman, Orfgen, & Foregger, 2007).

However, sometimes people have no past experience or only very weak attitudes. Converse (1970) and Hovland (1959) observed these extreme conditions and found that under this circumstance, people's responses were constructed completely based on currently accessible thoughts. But, if people hold very strong attitudes, the current information might receive no weight, and the stored attitude might dominate. This is because the strong attitude is well-rehearsed and highly accessible from memory (Lavine, et al., 1998, p. 360).

The idea of attitude accessibility is included in many theories, such as Petty and Cacioppo's (1986) elaboration likelihood model (ELM) and Chaiken's (1987) heuristic-systematic model (HSM). And, both theories also proposed that when the initial attitude is strong, people tend to maintain it and are biased in processing new information. Attitude is easily biased in the direction of how the new stimulus (e.g., item order) is introduced.

To sum up, item order is a factor which may influence judgment. Research in this field began with finding a main effect of item order on tests which were divided into four kinds: random arrangement (items are assigned randomly to examinees to examine the effection of test scores), section arrangement (the entire section of items is moved instead of moving individual item), easy-to-hard and hard-to-easy (the sequence of items depends on the item difficulty), and altering context (changing the difficulty or content of preceding items). Then, research shifted to probe other biological and psychological factors (e.g., gender or anxiety levels) which likely affect test results. More recently, research has integrated the idea of adaptive testing to estimate item parameters with altered question orders (Leary & Dorans, 1985, p. 389-393).

In attitude measures, people are asked to express their opinions about the specific event or issue. According to the item context, people might first try to recall relevant information about it. This process might result in a successful recollection if they had experience of it before. However, if this process fails, people might rely on whatever is accessible currently to construct an attitude toward the topic (Gregoire, 2003). In this way, former items serve as anchors and arouse respondents' memory. Different item contexts lead to different answers. When the stored memory or attitude is strong, it can be retrieved easily and the current messages will receive little weight in forming final responses. However, if only weak attitudes exist, people's responses might be based on the current information (e.g., test questions or item context) or statements (e.g., moods or thoughts). Item order is factor which influences people's responses, particularly when attitudes are weak (Fazio, 1990).

Research on Item Order Effects

Achievement Tests. One of the reasons for studying item order effects was that some researchers queried whether two tests still measured the same thing if item sequences were changed. Results on this topic are conflicting. Some studies fail to show the effects of item order on test performance. In 1964, Brenner conducted four experiments to examine test reliability, difficulty, and discrimination by altering the item order. Three measures were estimated by the following experimental forms: difficulty (average numbers of item correct), reliability (evaluated by Kuder-Richardson Formula 8), and discrimination (average point-biserial correlation between item and total test score). The results indicated that with items arranged from easy to hard, hard to easy, or randomly, no statistically significant difference appeared in test difficulty, discrimination and test reliability at alpha equal to .01 level (Table 1). Munz and Smouse (1968) observed students' achievement test scores with three forms of item difficulty orders (easy to hard, hard to easy, and random) which revealed that item difficulty order failed to show an effect on total test score (*F* = 1.05, *p* > .05).

Table 1.

Achievement Test Differences, Reliabilities, Discrimination Values, and Significance Test Results

Form		Difficulty		Reliability		Discrimination	
First							
	Easy to Hard	21.18	*p* > .50	.578	*p* > .75	.220	*p* > .40
	Random	21.04		.553		.232	
	Hard to Easy	20.90		.598		.218	
Second							
	Easy to Hard (first 10 Items) and Random for rest	24.04	*p* > .70	.753	*p* > .40	.283	*p* > .02
	Hard to Easy (first 10 items) and Random for rest	23.93		.674		.250	
Third							
	Easy to Hard	26.14	*p* > .60	.778	*p* > .70	.309	*p* > .20
	Hard to Easy	26.33		.805		.326	
Fourth							
	Easy to Hard	23.69	*p* > .30	.736	*p* > .90	.284	*p* > .70
	Hard to Easy	24.17		.747		.289	

Note. From Brenner (1964).

Later, Monk and Stallings (1970) generated 22 forms of a test by using random ordering of items. The result indicated that there was no statistically significant difference in test scores for the 22 forms of the test (Table 2). In this study, the reliabilities of each form were also calculated which presented that only slight variations between the arrangements of items.

Table 2.

The Effect of Item Rearrangement on Achievement Test Reliability and Test Scores

Test Form	Mean	Standard Deviation	Reliability (KR-20)
1	69.66	13.44	.892
2	72.10	11.10	.845
3	73.11	16.56	.938
4	75.21	11.14	.858
5	68.46	18.15	.944
6	73.49	10.74	.840
7	69.97	10.92	.826
8	62.62	9.14	.727
9	60.28	12.46	.854
10	62.21	11.74	.838
11	58.75	8.67	.802
12	58.61	8.21	.707
13	52.36	10.12	.834
14	50.55	9.68	.811
15	134.43	25.07	.935
16	129.13	23.21	.920
17	49.70	8.19	.728
18	48.53	8.29	.731
19	48.74	9.86	.814
20	48.29	9.86	.813
21	130.02	23.36	.921
22	129.33	22.26	.916

Note. From Monk and Stallings (1970).

In 1973, Klosner and Gellman examined 54 students' test performance with three forms of tests (ordered by subjects, easy-to hard within subjects, easy-to-hard across subjects). No statistically significant difference in test performance was found for different orders of items, $F = 1.104$, $p > .01$. Kleinke (1980) observed 484 students' performances in two forms of test ordered from easy-to-hard and uniform. There was no significant difference in test score between these two forms of the test, $F = 2.92$, $p > .05$. Plake, Melican, Carter, and Shaughnessy (1983) also examined test performance by administering three forms of tests (Easy to Hard, Spiral Cyclical, and Random). No significant difference in test performance was presented ($F = .28$, $p > .10$). Klimko (1984) administered three different difficulty orders (easy-to-hard, hard-to-easy, random) of tests to 111 college students. Two hierarchical multiple regression analyses were applied, and the result shown that item arrangement based on item difficulty did not affect test performance, F (5, 105) = 1.04, $p < .24$, and F (5. 105) = 1.68, $p < .19$.

Similar results were also presented in Laffitte's (1984) study with four versions of achievement tests: (1) items arranged from easy to hard within each chapter, (2) items arranged from easy to hard across chapters, (3) items arranged randomly within each chapter, and (4) items arranged randomly across each chapter, for college students to observe differences in their total test scores. The results indicated no significant differences among test scores on these four versions of the test. Furthermore, students' perception of test difficulty was not influenced by test item order. Plake, Patience, and Whitney (1988) applied three forms of test (Easy-to-Hard, Easy-to-Hard within content, Spiral Cyclical) and no significant order effect was found either.

Some investigators did find evidence that test performance is affected by the order of items. For instance, Flaugher, Melton, and Myers (1968) applied four patterns of item order (standard arrangement, reordering within blocks, reordering between blocks, and reordering between and within blocks) to examine the verbal and math test scores of over 10,000 students. Results indicated that some arrangements were more difficult than others. Hambleton and Traub (1974) observed the performance on mathematics test of 11th graders with two different patterns of item order (easy-to-hard and hard-to-easy). The results showed that students obtained higher scores on items arranged from easy-to-hard than hard-to-easy with *F* (1, 102) = 4.06, *p* < .05. Barcikowski and Olsen (1975) administered two types of reading test (multiple-choice and true-false) in two different orders (hard-to-easy, easy-to-hard) for 85 students to examine the difference of test scores and perception of item difficulty. The results found that test scores were influenced by the order of items, *F* (12, 72) = 7.58, *p* < .05. In multiple-choice items, students perceived items were easier when presented by difficulty ordered from hard to easy. In true-false items, only the difficult items were viewed as significantly easier when items were presented in the order of hard to easy (Table 3).

Table 3.

Summary of Group Means, Standard Error of the Mean Differences, and t-test for Item Rating and Subtest Scores on a Reading Test

	Subtest Mean			
Question Type	Hard to Easy	Easy to Hard	Standard Error	t-value
	Item Rating			
Multiple Choice				
Easy	2.81	3.02	.10	2.18*
Medium	3.03	3.29	.10	2.59*
Hard	3.07	3.70	.10	6.51*
True-False				
Easy	2.28	2.46	.11	1.42
Medium	3.04	3.19	.10	1.61
Hard	2.91	3.39	.11	4.49*
	Subtest Scores			
Multiple Choice				
Easy	7.72	6.19	.32	4.85*
Medium	4.93	5.36	.35	1.20
Hard	2.13	1.93	.30	.79
True-False				
Easy	9.63	9.43	.14	1.39
Medium	6.91	7.12	.30	.71
Hard	3.70	3.74	.35	.12

Note. 1 = very easy; 5 = very difficult. From Barcikowski and Olsen (1975).
* *p* < .05

Plake, Ansorge, Parker, and Lowry (1982) also found that males scored significantly better than females for items arranged from easy-to-hard and randomly. In addition, significant order effects were also found in both perceived performance and perceived difficulty.

Table 4

Summary of Item Order Effects on Achievement Test Scores, Item Statistics, and Reliability

		Ordering	Statistics				
Study	N	(Forms)	Mean	SD	t-value	F-value	Reliability
1. Munz, Smouse	102	E-H, H-E, R				*F* = 1.05,	
(1968)						*p* > .01	
2. Klosner,	54	By chapter	60.06	5.98		*F* = 1.104,	.675
Gellman		E-H within chapter	61.67	5.10		*p* > .01	.586
(1973)		E-H across chapter	59.72	6.08			.680
3. Hambleton,	106	E-H	11.41	3.4		*F* = 4.06,	
Traub (1974)		H-E	9.96	3.71		*df* = 1/102	
						p < .05	
4. Kleinke	484	E-H	22.5	5.4		*F* = 2.92	
(1980)		Uniform	21.8	5.6		*p* < .089	
5. Plake,	170		Male Female		t = 9.04, p < .001	F = 3.41	
Ansorge,		E-H	33.40 20.56		t = 1.08, p > .15	df = 4/316	
Parker,		SC	25.12 20.32		t = 3.26, p < .01	p < .009	
Lowry (1982)		R	25.75 23.62				
6. Plake,	167	E-H	Male Female	Male Female		*F* = .28	
Melican,		SC	43.74 49.11	7.3 5.99		*P* > .01	
Carter,		R	45.15 49.31	5.80 5.46			
Shaughnessy			44.9 49.07	5.82 5.86			
(1983)							

Table 4 (continued)

Study	N	Ordering (Forms)	Statistics								
			Mean		SD		t-value		F-value	Reliability	
1. Klimko (1984)	111	E-H	35.92		6.78				*F* = 1.04, df = 5/105 *p* < .24; *F* = 1.68, *p* < .19		
		H-E	36.96		6.23						
		R	35.08		7.19						
2. Plake, Patience, Witney (1988)	779		Male	Female	Male	Female	Male	Female		Male	Female
		E-H	9.53	11.07	4.05	4.36	.78	.81		.78	.81
		E-H within content	9.18	11.17	4.79	4.40	.85	.81		.85	.81
		SC	9.64	10.85	4.27	4.36	.81	.80		.81	.80

Note. N, numbers of participants; SD, standard deviation; E, easy; H, Hard; R, random; SC, spiral cyclical

Some researchers also investigated item order effects based on both statistical and cognitive difficulties. Newman, Kundert, Lane, and Bull (1988) observed undergraduate students' performance on both statistical and cognitive item difficulty. Four forms of an educational psychology test were created: (1) items were presented by ascending statistical difficulty (easy, medium, hard), (2) items were presented by increasing cognitive difficulty (knowledge, comprehension, application), (3) items were presented by descending cognitive difficulty (application, comprehension, knowledge), (4) items were presented by decreasing statistical difficulty (hard, medium, easy). Results showed that there were no statistically significant differences in statistical difficulty, F (1, 116) = .19, p > .05, or cognitive difficulty, F (1, 116) = .30, p > .05, but subscores were affected by the item order. Examinees scored higher on hard items when they confronted items presented by increasing cognitive difficulty. Test takers who received forms ascending with cognitive difficulty ordering obtained higher subscores for hard comprehension items. These studies indicated that item order can have a significant effect on results. Test takers might have different perceptions with different orders of items. Research on effects of context and item order on achievement tests, is, thus, mixed.

Aptitude Tests. Item order effects were also investigated in some aptitude tests. Gershon (1989) administered three forms of an aptitude test (easy-to-hard, hard-to-easy, random) to 1233 students. A statistically significant order effect was found (F = 3.08, p < .05), and students performed better on the items arranged from easy to hard than hard to easy or random condition. However, the results of item order effects in aptitude tests were also mixed. Different results can also be found in this field. Vega and O'Leary (2006) applied two versions (hierarchical: least severe to most severe, interspersed: mixed order of severity) of the test to examine 641 students' intimate partner aggression. The test outcomes of two subscales of the test were not affected by the order of items, $t_1 = .59$, *p* > .05 and $t_2 = .29$, *p* > .05 (Table 5).

Table 5.

Summary of partner aggression group means and t-test by item order

	Item Order			
Scale	Hierarchical (N = 323)	Interspersed (N = 318)	t-value	*p*-value
Subscale 1				
Mean	3.43	3.34	0.59	.56
SD	2.05	1.98		
Subscale 2				
Mean	1.50	1.45	0.29	.77
SD	2.15	2.12		

Note. From Vega and O'Leary (2006).

Attitude Measures. Tourangeau, Rasinski, and Bradburn (1991) stated that different responses appeared when the order of items about general happiness or marital happiness was varied. Respondents may use prior items to produce different interpretations of later items. Similar results can also be found for attitude measures.

Frantom, Green, and Lam (2002) applied two forms (grouped and randomly ordered items) on an attitude test. Statistically significant differences in item local independence and invariance were presented in both forms of the tests. The correlation between logit item position for both forms of the test for the first student-oriented attitude scale was .95. In the first subscale, four items showed statistically significant differences in logit item position. Three items presented significant differences in logit item position in the second subscale, and the correlation between logit item position for both forms of the test for the second subscale was .51. In addition, for grouped items, 6 of 7 items with significantly different logit positions occurred when the wording was in the same direction as the preceding items.

Bowling, Boss, Hammond, and Dorsey (2009) investigated the susceptibility of job attitudes to context effects for college students. Two job satisfaction scales were administered in three experimental conditions (positive, negative, control). Participants in the positive condition were asked questions in a positive way, and negative questions were asked in the negative condition. Questions that did not contain a positive or negative tendency were asked in the control condition. The evidence suggested that responses to job attitude items were influenced by context. The responses to job attitude depended on whether participants were asked to think about positive or negative aspects of their jobs (Table 6).

Table 6.

Summary of t-tests and effect sizes on two job satisfaction measures

	Measure 1			Measure 2		
Attitude	t-value	Effect Size	*p*-value	t-value	Effect Size	*p*-value
PN	3.75	.65	*p* < .01	4.45	.79	*p* < .01
PC	2.59	.44	*p* < .05	2.64	.46	*p* < .01
NC	.95	.16	*p* > .05	1.37	.24	*p* > .05

Note. PN, Positive versus Negative; PC, Positive versus Control; NC, Negative versus Control. From Bowling, Boss, Hammond, and Dorsey (2009).

According to these studies, results of item order are inconsistent, especially for achievement tests. Some studies found that item and section orders may influence item and section characteristics, such as difficulty and inter-item correlation, which may result in effects on test performance (Moses, Yang & Wilson, 2007; Schurr & Henrisken, 1980; Zwick, 1991). However, more consistently results showed significant effects of item order on responses for most studies involving attitude tests. This result suggests researchers may need to be more critical when attitudes are assessed via CAT.

Adaptive Testing

An adaptive test is one in which items for each examinee are selected during the process of administering the test, with items selected at an appropriate difficulty level for each participant's current trait level (Weiss, 1983). In contrast, fixed length and fixed sets of items (e.g., paper-and-pencil tests) administered to every examinee are called conventional tests. There are other terms which also refer to adaptive tests, such as tailored, sequential testing, programmed, individualized, branched, and response-contingent (Weiss, 1985).

Problems with Conventional Tests

A feature of conventional tests is that every examinee is administered a fixed number of items which evokes some problems when the test purpose is to measure a wide range of trait levels. Based on measurement precision, test constructors can develop a peaked conventional test or a rectangular conventional test. In a peaked conventional test, items are selected centered around a level of difficulty. Generally, items of difficulty of .50 are chosen to maximize the variance of test scores and internal consistency reliability. However, this kind of test provides only a little information for individuals with relatively high or low trait levels. It measures well only for people whose trait levels are close to the difficulty level at which the test peaked (Weiss, 1985).

In the rectangular conventional test, equal numbers of items are selected for a useful range of each difficulty level, which can provide information for people even with very high or low trait levels. The rectangular conventional test can provide equal precision at different trait levels, but based on the feature of fixed length, only a few items will be suitable for people at any trait level which affects the quality and precision of the assessment (Weiss, 1985).

Further, in the classical measurement model, the characteristics of examinee and test cannot be separated and are based on the particular test that was administered. The item difficulty and item discrimination both depend on the particular samples of participants, which also holds for the score reliability and validity. Generally speaking, the classical measurement model is test-driven not item-driven. There is no clear basis for predicting examinee performance on an item, and the standard error of measurement is the same for every test taker which is clearly not the case in practice. This makes it difficult to compare examinees who take different tests because there is no relationship between the tests (Bond & Fox, 2001; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991).

Principles of Adaptive Testing

To remedy the shortcomings of the conventional test, adaptive testing was created which provides a set of items for each examinee with appropriate levels of difficulty to measure different trait levels with equal precision (Weiss, 1985). Adaptive testing was first applied by Alfred Binet and his colleagues in 1905 on a measure known as the Binet intelligence test. In that test, the trained examiner needed to determine the starting age level by estimating an individual’s ability level. After finishing one block of items, the examiner has to decide whether a more difficult or easier block of questions should be administered next. This process is repeated until all questions of a block are answered correctly, which can be identified as the basal age. Therefore, the next higher level of difficulty (age level) items would be administered. During this process, the items are administered and scored immediately until all items of a block are answered incorrectly. This level is defined as the ceiling age of this test taker (Weiss, 1983; Weiss & Vale, 1987).

Since the publication of Binet's intelligence test, the idea of adaptive testing attracted much attention. Several methods were based on Binet's IQ test, such as Lord's (1980) Flexilevel testing, and Sheehan and Lewis' (1992) Testlets. These procedures were intended to determine a student's general ability level within the first several test items (Georgiadou, Triantafillou, & Economides, 2006). According to the adaptive testing research literature, McBride (1997) concluded that the data sources of adaptive testing can be divided into four different kinds. First, with live testing data, a sample of examinees are administered both adaptive and conventional tests, then the test scores and item response level data on these two forms are compared. Second, real data simulation simulates adaptive testing by collecting response data from the conventional test. Both live testing data and real data simulation are expensive and time-consuming. Third is theoretical analysis which is usually based on item response theory (IRT). This method deduces test information, measurement error, or item means analytically to specify item parameters and levels of ability of the test. Fourth are the computer simulation studies which specify item parameters, ability levels, and item response models to produce data by using random number generators.

With these four kinds of data sources, correlations between adaptive and conventional test scores were compared before IRT was introduced. After IRT was applied to test design, the comparison of measurement precision by varied ability levels was assessed (McBride, 1997).

Adaptive Testing Based on Item Response Theory (IRT)

Because of the shortcomings of classical test theory, it is not suited to adaptive tests. In the classical measurement model, validity, reliability, and item quality are inter-correlated when test takers take the same set of test; but this is not the case for adaptive tests. The appropriate theory for adaptive tests was proposed by Birnbaum in 1958, called latent trait theory. Lord and Novick (1968) also discussed this theory in their treatises. In 1980, Lord gave a complete account of latent trait theory, now called item response theory (Green, Bock, Humphreys, Linn & Reckase, 1984, p. 348).

In item response theory (IRT), three parameters can be used to characterize test items. Item difficulty refers to the position at which the examinee has a .5 probability of answering a question correctly. "Item discrimination represents the slope of the item characteristic curve (the probability of a correct response as a function of trait level) at the difficulty level for the item (Weiss, 1985, p. 781)". The third parameter is the pseudoguessing parameter which refers to the probability of the test taker correctly answering the item with an extremely low trait level. These parameters are independent for each test taker (Simms & Clark, 2005; Weiss, 1985).

An item response model can be used to specify the relationship between the test performance of examinees and traits or abilities. Hambleton and Swaminathan (1985) proposed three characteristics of item response models. First, item parameter estimates are independent of the particular group of test takers. Second, examinee ability estimates are also independent of the specific sample of test items. Third, the precision of ability estimates for each examinee is available. Adaptive testing is designed based on these features of IRT.

The strategy of adaptive testing is to select the item with maximum information at an individual's current estimated trait level. IRT-based methods are based on the responses of several administered items to calculate the current trait level, and a new item is administered which provides the maximum information according to the prior responses. This item selection process indicates that the item with the highest value of information at the current point of trait level is selected to be administered. This process is repeated until there are no items left at the examinee's trait level or sufficient precision is achieved, and the test will be terminated at that time. The IRT-based estimation also provides the standard error of measurement at any given trait level. Thus, when a given level of standard error of measurement is reached, the test can be stopped (Green, 1982; Weiss, 1985; Weiss & Vale, 1987).

Adaptive Testing and Computers

Adaptive testing based on IRT became feasible since the advent of computers. The power of the computer to store test information and to administer and score items, make adaptive testing wide spread (Bunderson, Inouye, & Olsen, 1989). In the late 1960s, research was supported by the U.S. Armed Services and other federal agencies; many related conferences were also held to discuss applications of adaptive testing (Hambleton et al., 1991) .

The major idea of CAT is to administer test questions appropriate for the test taker's current trait or ability level. Generally speaking, the CAT starts with items randomly selected from an average level of difficulty. If the examinee answers the question correctly, the ability or trait level of the examinee will be recalculated and a more difficult question will be administered. In contrast, when the examinee provides an incorrect response, the following sequence of items will become easier. CAT is based on the performance on a prior item to select items with maximum information at that current trait or ability level (Lilley, Barker & Britton, 2004, p. 110). Figure 2 illustrates the components and processes of the typical CAT (Waller & Reise, 1989).

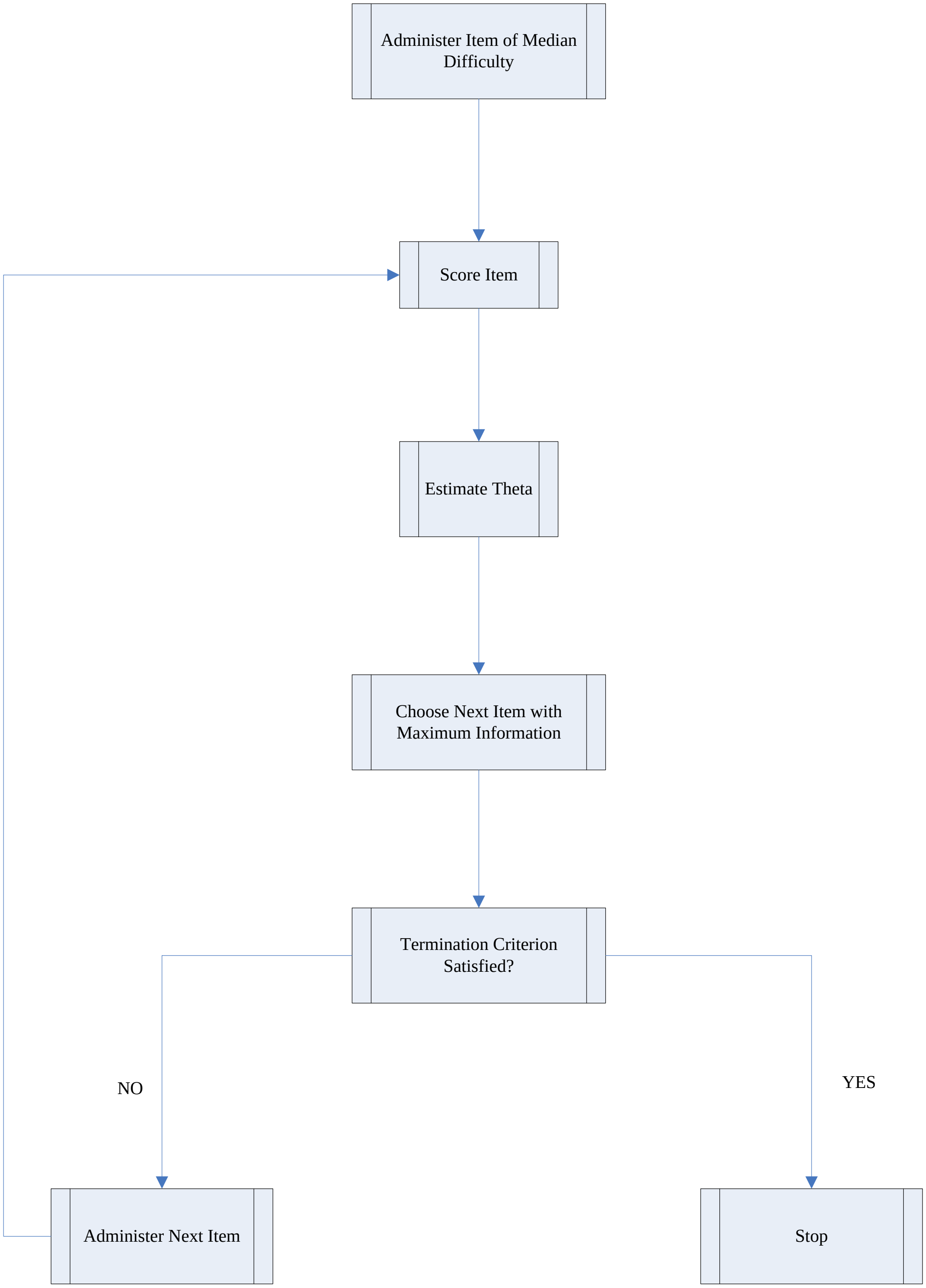

Administer Item of Median Difficulty
Score Item
Estimate Theta
Choose Next Item with Maximum Information
Termination Criterion Satisfied?
NO
YES
Administer Next Item
Stop

Figure 2. Flowchart of an adaptive test from Waller and Reise (1989).

In 1984, Weiss and Kingsbury concluded that the structure of CAT comprises the following components: (a) an item response model, (b) a calibrated item pool for each trait or ability level, (c) an entry level, (d) an item selection rule, (e) a scoring method, and (f) a termination rule to stop administering the test. For each component, there are numbers of options to implement. Six of these options are summarized below.

1. Item response model. Based on the test response formats (free-response or multiple choice), there are three different models can be selected: one-, two-, or three-parameter logistic model.

2. Item pool. The parameters of each item should be calculated following an appropriate procedure. No specific guideline is provided as an appropriate numbers of items. Weiss and Kingsbury (1985) proposed that a pool of satisfactory quality for CAT is 100 items, and a pool of 150 to 200 items is preferred. Further, items must have high discrimination and span the full range of trait or difficulty levels in order to match the level of the population (Urry, 1977).

3. Entry level. In adaptive testing, a test can be started at different levels of difficulty for different examinees. If a test taker is known to have high ability, the test can be administered with more difficult items. Adaptive testing assumes that different initial entry levels do not severely influence the precision of the test, but longer test length will be required if the test does not begin with an accurate entry level.

4. Item selection. There are two procedures used currently in selecting items. The first method is maximum information. Items that provide the maximum information (e.g., minimize standard error) of the test taker’s current trait or ability level are selected (Weiss, 1982). The second method is Bayesian item selection. Items that minimize the variance of

the posterior distribution of the examinee's ability are selected. The posterior distribution is concentrated as more items are administered, and the precision also gets higher. In this procedure, the item exposure problem needs to be taken into consideration because informative items tend to be administered time and time again. Therefore, Green and his colleagues (1984) suggested that slightly less than optimal items can also be administered to avoid an item over exposure problem.

5. Scoring method. A major advantage of CAT is that test score or ability can be obtained during the testing process. Examinees can receive feedback immediately after finishing the test. There are two estimation procedures applied in CAT. One is maximum likelihood estimation. This method is implemented when the number of questions is small. To overcome this limitation, Bayesian estimation was introduced. But when an inappropriate prior distribution is chosen, the result may be biased.

6. Termination rule. The essential feature of CAT is that individuals based on their ability or trait levels obtain different sets of items. Each examinee has a different length of test. The test is stopped when the prespecified standard error is reached, which means that the necessary information has been obtained (Green, 1982).

In applying CAT, despite the merit of shortening the test length without sacrifice of measurement precision there are still several advantages which the classical measurement model cannot achieve (Green, 1982; Hambleton et. al., 1991; McBride & Martin, 1983; Wainer, 2000). These advantages include:

1. Test security is enhanced. A test is more secure in the computer than on the desk. Further, it is very difficult for examinees to obtain higher scores by memorizing only a few items from the item pool.

2. Individuals can have their own pace in testing. The test is on demand. The time limit is the additional information to estimate the test taker's proficiency.

3. The frustration for examinees is minimized. Items are administered based on the examinee's current trait or ability level. Individuals stay busy and challenged during the test but are not discouraged.

4. There is no need for an answer sheet. The response to items is by clicking the answer on the computer. The problem of alternatives for erased answers is solved.

5. The test score can be reported immediately. Test takers can receive feedback right after finishing the test.

6. Faulty items can be removed easily. Once a defective item is identified, the computer can expunge it from item pool, which is easier than deletion from a conventional test.

7. Item formats can be flexible. With a voice synthesizer, not only multiple-choice questions, but spelling or conversation tests can be included in CAT.

8. Test standardization is greater, and the test supervision time is less.

Item Order Effects in CAT

Since the computer was invented and applied to the measurement area, a new era of psychological assessment has been presented. Nowadays, computerized tests are applied in many different fields, such as academic achievement (Mills, 1999), intellectual ability (Weiss et. al., 1987), vocational interests (Hansen, Neuman, Haverkamp & Lubinski, 1997), neuropsychology (Russell, 2000), and personality testing (Butcher, 1987).

However, even though the use of computerized adaptive testing is increasing, CAT research is still focused on the domain of achievement assessment (Wainer, 2000). Most studies emphasized methodological issues, such as item selection procedures (Dodd, 1990), scoring method (Weiss & McBride, 1984), validation studies (Simms et al., 2005), and quality of the item pool (Belov & Armstrong, 2009). The use of CAT for attitude and personality assessment is limited.

There are two reasons that may explain the absence of personality assessment with CAT. First, the idea of CAT is more complex than classical test theory since it's based on IRT. Therefore, the application of CAT is based on psychometrics which traditionally focus on ability and achievement tests (Waller et. al., 1989, p. 1051). Second, the theoretical framework of IRT only works for unidimensional tests which is difficult to achieve in personality testing (Ortner, 2008). These two reasons can also be inferred to apply to attitude tests.

Although CAT has many advantages which cannot be achieved with the classical measurement model, there are still other shortcomings to which attention should be paid: (a) the range of test items seems restricted because only suitable items are administered to the examinee; (b) the test taker may find the principles of a questionnaire who is possible to hypothesize the target of measured trait and change behavior in the test and affects the quality of measurement; (c) the item pool is an important issue for CAT, the test result is easily to be affected by lacking of items toward the end of the test section, the test result may be influenced by extreme items and other unusual behaviors; (d) Every test taker will receive different order of items. An examinee's responses may be affected by the preceding items that he or she had confronted which makes context effect occurs (Ortner, 2008).

Prior studies focused on item order effects for conventional tests. Different explanations were also proposed by researchers to interpret this phenomenon, such as anchoring-and-adjustment, primacy and recency, and attention decrement. The results of studies were inconsistent. There are still no specific reasons that can be used to describe order effects. It is possible to start a test at different levels of difficulty for examinees in CAT, but shorter test length and greater precision is what most CAT application pursues. One feature of CAT is that item selection will move to the examinee's trait level as the test progresses. Therefore, investigators believe that the test results will not be seriously influenced with different entry levels (Weiss et al., 1984).

Recently, this issue of item order effects on adaptive testing attracted the attention of some investigators. Ortner (2004) investigated the effect of changing item positions in the Eysenck Personality Profiler. Two versions of tests were administered. One consisted of Rasch-homogenous items in its conventional order, and the other was distributed in the exact reverse order. Results presented that there was no statistically significant difference in mean scores on these two versions. However, in applying IRT to analyze the data, different item difficulties were found in three of seven scales, and the model fit of these three scales also failed. According to these outcomes it can be concluded that the item parameters of the personality test were unstable, and altering the item order led to changes in difficulty.

In 2008, Ortner investigated the effects of item order in CAT on the domain of personality assessment in the Eysenck Personality Profiler. One conventional and three adaptive versions were administered: (a) the conventional version items in the original order; (b) an adaptive version beginning at a medium trait level; (c) an adaptive version beginning at a high trait level; and (d) an adaptive version starting at a low trait level. Significant differences in mean person parameters were found in three of seven scales of the adaptive versions. Furthermore, the average reaction time in answering the item also varied. Ortner found a similar conclusion as in his prior study that item presentation order is a problem in applying CAT in personality assessment.

However, different outcomes appeared in Bergstrom, Lunz, and Gershon's (1992) study. They observed an effect of different test difficulty on examinee ability measures and test length in a CAT. A total of 225 examinees were randomly assigned to hard, medium, and easy test difficulty conditions (50%, 60%, and 70% probability of correct response). Results indicated that there was no statistically significant difference in estimation of examinee ability with administering different difficulty levels of tests. But more items are required when the probability of correct response increases. When the test is easier, the number of items increases slightly.

In comparing these studies, the major difference is the domain of assessment. The statistically significant differences in item difficulties, person parameters, and reaction time all presented only with the personality measure; the ability test differences appeared only as a slight increase in number of test items. The test stability in applying CAT to domains besides ability assessment needs to be taken into serious consideration.

The merit of CAT is that items are selected and administered tailored to the individual trait level, and each examinee is confronted with different items. This cannot be achieved with conventional tests. In CAT, test length and time are saved. However, these advantages of CAT basically involve the measurement process rather than the underlying attitude structures. If the attitude structures are to be measured, context effects may be a problem.

To sum up, research on item order effects mostly focused on achievement tests. The results of differences on test scores, item difficulty, and perception of item difficulty on these studies are mixed. On the other hand, research on this topic on attitude measures administered via both conventional and adaptive testing is limited. For studies of conventionally administered attitude measures, results were consistent in that responses were influenced by the preceding items. People tend to adjust answers based on the prior item, and their perceptions may be affected by the order of items presented. However, the arrangements of items in the prior studies mostly ordered the items from either easy to hard, hard to easy, or in random order. There was no research specifically examining effects of initial item order. This topic of item order effect has not received much attention on CAT attitude measures.

In this study, a conventional attitude measure was administered to detect item order effects. Effects due to altered item order were hypothesized if items with particular item parameters (high, medium, or easy) precede other items. Further, an exploratory study of item order effect on CAT was conducted. The reaction time to answer an item via CAT was observed. Fazio (1990) hypothesized that the time it takes to finish a questionnaire may be affected by certain orders of items. If examinees confront extreme items, the response time may be longer. Test takers may change responses to more careful answers, and responses of unacceptable categories will be avoided. On the other hand, if a neutral question is administered first in the questionnaire, the feeling of being examined is less. It was hypothesized that reaction time is longer when a test begins with extreme level versus easy level. For this reason, the mean time to answer items can provide insight into the examinee's cognitive process (Ortner, 2008).

There are several contributions of this study examining item order effects on attitude measures. First, if the item order is a factor influencing responses, the equivalence of tests with rearranged items should be seriously questioned. Second, if the performance is affected by item order, whether the identical latent trait is estimated is in question. Further, the exploratory study of CAT provides an indication of whether it is feasible to administer attitude measures via CAT.

Research Questions

Above all, the research on item order with attitude tests administered via both conventional and computerized adaptive testing is limited. Research on the topic was most focused on achievement tests, and results indicated that items presented first usually serve as the anchor for test takers. Examinees' tend to adjust their answers to subsequent items based on this anchoring item. Further, the strength of test takers' attitude toward a specific topic or event might be different when different items are presented in different orders, and it is likely that this effect might reflect on a survey test score. In this case, it is highly possible that different item arrangements result in different response patterns. Therefore, the aim of the present dissertation was to investigate the effects of item order when an attitude measure with different versions of the conventional and computerized adaptive formats. For the conventional format of surveys, it was hypothesized that different test score, item difficulty, item discrimination, and test reliability would be found with items ordered in different difficulty orders (easy-to-hard, hard-to-easy, and medium-then-random). For computerized adaptive surveys, it was hypothesized that different test score, reaction time, and test length would be detected when the survey began with items representing different trait levels (medium or extremely high). Based on the reviewed studies and hypotheses, the following research hypotheses were addressed in this dissertation:

1) Significantly different test scores are obtained on the scales listed when a measure begins with items ordered from different trait levels (easy-to-hard, hard-to-easy, or medium-then-random).

   Responsibility Scale

   Dissertation Barriers Scale

   Procrastination Scale

2) Significantly different item difficulties (parameters) are obtained on the scales listed for an attitude test that contains items ordered from different trait levels (easy-to-hard, hard-to-easy, or medium-then-random).

   Responsibility Scale

   Dissertation Barriers Scale

   Procrastination Scale

3) Significantly different test reliabilities are obtained on the scales listed for an attitude measure with items ordered from easy-to-hard, hard-to-easy, or medium-then-random trait levels.

   Responsibility Scale

   Dissertation Barriers Scale

   Procrastination Scale

4) Significantly different item discriminations are obtained for an attitude measure with items ordered from easy-to-hard, hard-to-easy, or medium-then-random trait levels.

Responsibility Scale

Dissertation Barriers Scale

Procrastination Scale

5) A statistically significant relationship is found between test scores on the scales listed and people's attitude strength toward their dissertation/thesis process.

   Responsibility Scale

   Dissertation Barriers Scale

   Procrastination Scale

6) Significantly different test scores are obtained in computerized adaptive testing starting with items representing either medium or extreme trait levels.

   Responsibility Scale

   Dissertation Barriers Scale

   Procrastination Scale

7) Significantly different reaction times are obtained on the scales listed for computerized adaptive attitude tests starting with items representing either medium or extreme trait levels.

   Responsibility Scale

   Dissertation Barriers Scale

   Procrastination Scale

8) Significantly different test lengths (number of items administered) are detected in computerized adaptive tests starting with items representing either medium or extreme trait levels.

   Responsibility Scale

   Dissertation Barriers Scale

   Procrastination Scale

9) Significantly different mean person parameters are detected in computerized adaptive tests starting with items representing either medium or extreme trait levels.

   Responsibility Scale

   Dissertation Barriers Scale

   Procrastination Scale

Chapter Three: Method

In this dissertation, both conventional and computerized adaptive surveys were conducted. For this conventional survey formats, three versions of a survey with items ordered by different difficulty sequences was administered. Due to the limitations of simulation studies of context effects, this study employed a cross-sectional survey design where “data on a sample of respondents chosen to represent a particular target population are gathered at essentially one point in time” (Singleton & Straits, 2005, p.228). Further, an exploratory study of item order effects on CAT testing was also conducted. In both studies, participants were recruited with the assistance of university professors and graduate students.

All measures in this study were self-report. Test scores, item difficulties, test reliability, item discriminations, length of test, reaction time, and mean person parameters were outcome variables in this study. The independent variable is item order. For the conventional survey format, items were ordered by difficulty from hard-to-easy (H-E), easy-to-hard (E-H), or medium-then-random (M-R) trait levels. In CAT, items were started with either medium or extreme difficult trait levels.

Research Procedure

This study was conducted in four phases. In the first phase, extant data from administration of a dissertation/thesis completion survey (Green & Kluever, 1997) were used. Due to measurement requirements and to the multi-faceted nature of the scales on the dissertation completion survey, the data collected by Green and Kluever (1997) were recoded into dichotomous responses. Then a principal components analysis (PCA) was conducted to reduce the multiple facets of measures in that survey into a single dimension for each scale. Therefore, a modified version of the dissertation/thesis completion survey was generated and applied in the following three phases.

The second phase involved analyzing the modified dissertation completion survey items using the data collected by Green and Kluever (1997) in order to estimate item parameters of the responses. Based on the estimated parameters, three forms of the modified dissertation/thesis completion survey in a conventional format were generated with items ordered from hard-to-easy (H-E), easy-to-hard (E-H), or medium-then-random (M-R) trait levels.

In the third phase, comparison tests were applied. The purpose of this phase was to examine whether test scores, item logit position (difficulty), item discriminations, and test reliabilities were influenced by item order in the conventional format.

In the fourth phase, CAT versions of the dissertation completion survey were created and administered. An exploratory study of item order effects for CAT was conducted with a starting item representing medium or extreme trait levels. Differences in test scores, test reliabilities, test length, and reaction time in CAT formats of the survey were assessed.

Phase One

Participants.

Subjects were drawn from an urban private college of education in a western state. Respondents were doctoral students (ABD) and a smaller number of doctoral graduates. The ABD students refer to those who had finished coursework and had passed comprehensive exams, but had not finished the dissertation yet. Data were collected from a total of 239 respondents to a paper-and-pencil version of a dissertation completion survey created by Green and Kluever (1997). The sample comprised 142 graduates and 97 doctoral candidates (ABDs), 65.3% females and 34.7% males. The age of participants ranged from 28 to 70 years old with mean of 44.4 years old. There were 77.8% of participants who reported full time employment, 19.0% indicated part-time employment, and 3.2% of participants reported being unemployed. About half of both graduates and students reported they had experience with data analysis and conducting research, but only 10% to 23% of participants had published research.

Instrument.

Attrition from doctoral programs in education while completing a dissertation was estimated at approximately 50% (Johnson, Green & Kluever, 2000). Failure at this point is discouraging and frustrating for both students and faculty involved. Twenty percent of students give up at the dissertation stage and 50% drop out of educational doctoral programs (Bowen & Rudenstine, 1992). In education at doctoral level, students are trained to have the ability to understand and execute research. Many students give up after struggling for several years. Hence, studies have focused on identifying variables related to noncompletion of the dissertation, such as situational, program-specific, cognitive, and affective or personality factors (Germeroth, 1991; Jacks, Chubin, Porter & Connolly, 1983; Wagner, 1986).

For a number of investigations, reasons for failure to complete dissertations can be generalized into three aspects: responsibility, dissertation barriers, and procrastination. Green and Kluever (1997) designed their dissertation completion study around these three aspects. In this study, a modified version of the dissertation completion survey was used which comprised measures from those three domains.

(1) *Responsibility* was assessed by the Responsibility Scale (RS) which was developed by Green and Kluever in 1996. The 16-item measure includes two scales assessing students' and graduates' concepts of responsibility related to completion of the doctoral dissertation. This scale was generated based on Brickman and his colleagues' (1982) work.

In this measure, a seven-point continuum was applied, and the choices of student/university were at opposite ends. One end of the continuum (point 1) indicates total students' responsibility, and the opposite end (point 7) represents total university responsibility, so lower scores indicate stronger perception of student responsibility. The two scales measure the dissertation preparation and evaluation tasks. Sample items are "responsibility for progressing through the dissertation rests with …" and "responsibility for evaluating the content of the dissertation rests with …" Each item of the RS is answered twice. The first response is for the "IS" scale which assesses the current state of responsibility for tasks. The other response is for the "Should Be" scale which measures the subjects' opinion about who should be responsible for tasks for an ideal program. There are 32 choices for the 16 items of the RS. The reliability of this scale was .89 (Johnson et al., 2000).

(2) *Dissertation barriers* have been identified which relate to people's cognitive and affective characteristics (Green & Kluever, 1997). Many investigators had studied these characteristics in different ways, such as history of separation and loss in childhood (Stern, 1985), perfectionism (Germeroth, 1991), and persistence as a coping style (Weiss, 1987).

From these studies, self-discipline and self-motivation were identified as two major personal factors necessary for students to complete degree programs.

Grives and Wemmerus (1988) proposed a model of graduate student persistence which comprises three factors: program involvement (e.g., financial support and perceptions of relationships with the faculty), the actual student/faculty relationship, and department characteristics. On the other hand, Tinto (1993) suggested a model which posited stages of completing doctoral degree and factors within these stages which can be distinguished by the major tasks or relationships achieved. The first and second stages are achieved when students obtain content and research abilities, and also build both of academic and social relationships with faculty. The third stage is about the function of external commitments, such as family and job to doctoral candidates. To sum up, Tinto's model involved student attributes, program entry goals and orientation, institutional and program experiences, academic and social integration into a program, and research experiences. Financial aid, opportunities to work with faculty, and relationships between faculty and advisor are factors that impact the research experiences of students (Green et al., 1997).

The Dissertation Barriers Scale was developed by Green and Kluever in 1997 and was designed to identify the specific factors suggested by Tinto of doctoral students' and graduates' conception of barriers to dissertation completion. This scale comprised a total of 45 items with response on a -3 (major hindrance) to +3 (major help) scale. A midpoint (NC: not a concern to you) and a not applicable option were also provided. Sample items are "schedule meetings with a advisor(s) …"and "lack of structure of dissertation process …"In this scale, nine concerns were addressed: financial concerns (2 items), family/relationship concerns (4 items), relationship with advisor/committee (8items), dissertation topic concerns (4 items), structure/time concerns (6 items), working with committee (5 items), institutional resources (2 items), affective concerns (7 items), and perception of skills (7 items). For use in the current study, the responses were rescaled into a 1-7 point scale with 1 as major hindrance and 7 as major help.“NC” responses were treated as missing. Lower scores indicated greater perception of hindrances. The reliability coefficient of the total scale was .91.

(3) *Procrastination* was assessed using the revised Procrastination Inventory (Green, 1997). The definition of procrastination is “the tendency to put off doing something until a future date” (Johnson et al., 2000, p. 270). Studies showed that nearly one fourth of college students have a problem with procrastination, which is usually associated with negative academic performance (Ellis & Knaus, 1977; Johnson et al.). For instance, Semb, Glick, and Spencer (1979) found that students with procrastination problems tended to have poorer grades and course withdrawals. Research indicated that perfectionism, frustration tolerance, a high need for autonomy and approval, and fears of failure, success, and separation are related to procrastination (Burka & Yuen, 1983). In addition, cognition (e.g., self-efficacy and self-esteem), affection (e.g., depression and anxiety), and behavior (e.g., punctuality and organization) all correlated with procrastination (Johnson et al.).

The Procrastination Inventory was originally developed by Muszynski and Akanatsu (1991) to assess the cognitive and affective traits of scientist-practitioners. According to Muszynski and Akanatsu’s study, results indicated that the completion of dissertations of clinical psychology students can be predicted by the total procrastination scores and subscale scores. In 1997, Green and Kluever revised this inventory to measure procrastination of doctoral students in education. This inventory was formed by 43 items which were grouped into 11 subscales: low frustration tolerance, perfectionism, rebellion, difficulty of making decisions, need for approval, inability to take help, procrastination as work style, fear of finishing school, self-denigration, insufficient reinforcement/lack of structure, and task aversiveness. A 5-point scale was applied with 1 as “Not At All True of Me” and 5 as “Definitely True of Me.” Higher scores indicate higher levels of procrastination. The reliability of the total scale was .86.

Other items about subject’s experiences with dissertation preparation, strategies they employed in the process when working on the dissertation, and attitudes related to events of doing dissertation work were also included in this survey. Demographics and background information such as employment status while doing the dissertation, previous research experience, distance of residence from campus, financial support, and the amount of emotional support while doing a dissertation were also covered in this survey, but were not used in the current study.

Procedure.

The dissertation/thesis completion survey comprises three scales. Each scale includes several subscales. Due to the multi-faceted scales and measurement problems in the original survey format, a modified version of the dissertation completion survey was developed. In this phase, answers to the original survey were recoded into dichotomous responses. Then, a principal components analysis (PCA) was applied to reduce survey dimensions in order to generate the unidimensional dissertation completion survey with dichotomous responses. Three unidimensional scales were used reflecting each of the original three domains.

Analysis.

First, data collected by Green and Kluever (1997) were recoded into dichotomous responses due to the measurement problems introduced in the original format. For example, if the middle option of “not a concern for you” in the dissertation barriers scale is retained, people may choose this option for various reasons, such as the test taker might not understand the question, might not know him or herself well, or might not be interested in answering this item. Therefore, the retention of the middle option makes the appropriate model difficult to find (Ortner, 2008). For this reason, response categories were merged into a dichotomous format. Second, in order to reduce the survey dimensions, a principal components analysis (PCA) was conducted. Items loaded on the first component were selected to identify a unidimensional scale.

Phase Two

Participants.

In this phase, the participants were the same as those in phase one. There were 142 doctoral graduates and 97 doctoral candidates (ABDs) who responded to the original survey for a total of 239 respondents.

Instrument.

The modified dissertation/thesis completion survey was applied. This survey comprises three scales (Responsibility, Dissertation Barriers, and Procrastination), background, and demographic information (see Appendix A). In this phase, three versions of a conventional survey (items ordered with trait levels from hard-to-easy, easy-to-hard, and medium-then-random) were developed.

Procedure.

Three conventional versions of survey were generated. The item parameters were estimated by using the recoded responses to the dissertation completion survey collected by Green and Kluever (1997). In this stage, a graded response model (Samejima, 1969) was applied to estimate item parameters. Three new forms of a modified dissertation/thesis completion survey with items ordered from easy-to-hard (E-H), hard-to-easy (H-E), and medium-then-random (M-R) trait levels were developed based on the estimated item parameters.

Analysis.

The recoded data from the modified dissertation/thesis completion survey collected by Green and Kluever (1997) were analyzed to estimate the item parameters. First, the item parameters were calculated using PARSCALE 4 (Muraki & Bock, 2003). PARSCALE 4 is based on item response theory (IRT). In this software, Samejima's (1969) graded response model generalizes to the rating scale or partial credit model.

The graded response model is an extension of dichotomous IRT which can be applied to deal with ordered polytomous responses. In the general graded response model (Samejima, 2008), let $X_g(= 0, 1, \ldots, m_g)$ refer to a graded item score to item g and $\mathrm{X_g}$ be its realization, and the values of $\mathrm{m_g}$'s can be different for separate items.

The operating characteristic, $\mathrm{P_{x_g}}$, of the graded item score $\mathrm{X_g}$ is defined by

$$P_{x_g}(\theta) \equiv \mathrm{prob.}[\mathrm{X}_g = x_g \mid \theta]$$

The general graded response model is defined by

$$P_{x_g}(\theta) = [\prod_{u \le X_g} M_u(\theta)][1 - M_{(X_g+1)}(\theta)]$$

,

and $M_{X_g}(\theta)$ refers to the processing function. It increases strictly in θ except

$$M_{X_g}(\theta)\begin{cases} = 1 \text{ for } & X_g = 0 \\ = 0 \text{ for } X_g = & m_g + 1 \end{cases}.$$

Let $P^*_{X_g}(\theta)$ be the cumulative operating characteristic of the graded item score, $X_g = X_g$, then

$$P^*_{X_g}(\theta) \equiv \text{prob.}[X_g \ge x_g \mid \theta] = \prod_{u \le x_g} M_u(\theta)$$

According to the processing function and cumulative operating characteristic of the graded item score,

$$P^*_{X_g}(\theta)\begin{cases} = 1 \text{ for } & X_g = 0 \\ = 0 \text{ for } X_g = & m_g + 1 \end{cases}$$

From all of these equations, the operating characteristic $P_{x_g}(\theta)$ can be written as

$$P_{x_g}(\theta) = P^*_{X_g}(\theta) - P^*_{(X_g+1)}(\theta),$$

and $X_g = 0, 1, \ldots, m_g$.

For PARSCALE 4, parameters are estimated based on the graded response model, and the prerequisite condition of this software is that data are needed on all test items, and each item requires at least 200 or more responses. This requirement is met by Green and Kluever's (1997) data. After the item parameters were estimated, items were arranged based by item position estimates. For the easy-to-hard (E-H) version of the survey, items were ordered based on trait levels from easy to hard in each scale. The same method was applied for hard-to-easy (H-E) version of the survey. For the medium-then-random (M-R) version of the survey, five medium trait level items were placed at the beginning of each scale. Aside from the five initial medium trait level items, the remaining items of each scale were ordered randomly.

Phase Three

Participants.

In this phase, snowball sampling was employed. E-mail lists of the target population were accessed, relying on university professors and graduate students to identify participants. Further, surveys were also delivered via listserves to the target population. The main participants in this research study were primarily doctoral students and doctoral graduates. Doctoral students who had finished most of their coursework and doctoral graduates comprised the sample along with master's students and graduates who have experience in doing a thesis.

Prior to analysis, cases were deleted in which the number of missing responses was greater than 15 items in order to ensure that individuals responded to at least 70% of the items. Therefore, a total of 132 participants were included in the first version (H-E), 124 people responded to the second version (E-H) survey, and 118 people answered the third version (M-R) survey. The dropout rate for each survey version was: H-E, 64%, E-H, 59%, and M-R, 32%, indicating substantially more dropouts when items were ordered from hard to easy.

Instrument.

The instrument applied in this phase was the modified dissertation/thesis completion survey with items ordered by different trait levels. Three versions of the survey ordered by different difficult trait levels (hard-to-easy, easy-to-hard, and medium-then-random) were administered. For the hard-to-easy (H-E) version of the survey, items were ordered based on trait levels from hard to easy in each scale. The same method was employed for easy-to-hard (E-H) survey version. For the medium-then-random (M-R) version of the survey, five medium difficulty trait levels of items were placed at the beginning of each scale. Aside from the five initial medium difficulty trait level items, the remaining items of each scale were ordered randomly. Further, questions about participants' perceptions of whether their answers were affected by item order and their self-report of attitude strength toward each scale were also included.

Procedure.

For administering the conventional format survey, the potential participants were invited to participate in the survey via an e-mail. The purpose of the study, response method, and a link to the survey were included in the e-mail. Those who decided to participate in the study accessed the survey by clicking the link in the e-mail which took them to a SurveyMonkey (http://www.surveymonkey.com) site. SurveyMonkey is an online survey tool. People can create their own survey with any level of experience. The website employs a third-party to audit the security and privacy by keeping the data and account behind up-to-date firewall and intrusion prevention technology. In SurveyMonkey, potential participants first need to complete a consent form. Once they complete the consent form, the potential participants were forced to choose to continue or quit the survey.

Participants who were willing to take this survey affirmed their choice by clicking the "continue" button. Those who chose to quit the survey were thanked and exited the site automatically.

The forms of the survey were delivered randomly. It is estimated that it took from seven to ten minutes to complete each form of the survey. Responses to the surveys were confidential and not available to college faculty, but were available to participants if they sent a separate email to the researcher requesting their score.

Analysis.

In this phase, descriptive statistics were provided. Information about frequency, mean, standard deviation, effect size, kurtosis, and skewness were provided. The item difficulty and item discrimination of items on each survey version were calculated using PARSCAL 4. Pairwise difference tests were employed to examine differences in item logit position (difficulty) and item discrimination for each item among scale orders for the three versions of the survey. Reliabilities were also calculated for each scale of every survey version, and Feldt's (1969) test was conducted to detect the difference in reliability among scales for the three survey versions. Further, analysis of variance (ANOVA) was employed to assess differences in test scores among three scales for the three survey versions. Finally, correlations between total score and self-reported attitude strength of each scale for every version were also calculated. An alpha (α) level of .05 was applied for all statistical tests.

Phase Four

Participants.

In this phase, snowball sampling was also employed. Respondents were accessed, assisted by university professors and graduate students to identify other participants. The study was also addressed in some graduate courses by the researcher to invite members of the target population to participate in this survey. The main participants in this phase were similar to those in phase three in that they were doctoral students who have finished most of their coursework, doctoral graduates, and master's students and graduates who have experience in doing a thesis.

A total of 30 participants (7 male, 23 female) were included in this study. Equal numbers of volunteers were recruited for CAT surveys with items beginning with medium or extreme trait levels (15 taking the medium version and 15 the extreme version). The participants in this phase were 14 doctoral students who had finished most of their course-work and were working on dissertations, three master's graduates who had experience in doing a thesis, and 13 doctoral graduates who had experience doing a dissertation.

Instrument.

The instrument applied in this phase was the computerized version of the modified dissertation/thesis completion survey with items beginning with different trait levels. Two versions of the CAT survey with items beginning with either medium or extremely difficult trait levels were administered.

Procedure.

For the exploratory study of CAT, two versions of the survey were developed. The estimated item parameters calculated by PARSCALE 4 (Muraki & Bock, 2003) in phase two were applied in this phase. And, a post-hoc simulation study was conducted following the calculation of item parameters in order to generate the best set of options for survey items to transform into CAT versions. The estimated item parameters were then input into the program POSTSIM 2.0 (Weiss, 2005) which is useful when a calibrated item bank is available. Responses of a group of examinees on a survey administered as a conventional test are needed. POSTSIM 2.0 implements post-hoc real data simulation to evaluate various combinations of CAT parameters prior to live testing, including identifying entry points of CAT, the item selection rule, scoring method, and termination criteria.

In this software, an ASCII/text file is required. The implementation of POSTSIM2.0 to CAT applies only for dichotomously scored items. POSTSIM 2.0 assumes a 3-parameter logistic IRT model with D = 1.7.

$$P_{ij}(\theta) = C_i + (1 - C_i)\frac{\exp\,[D_{a_i}(\theta_j - b_i)]}{1 + \exp[D_{a_i}(\theta_j - b_i)]}$$

where

Pij is the probability of a correct response to item *i* by person *j*

θ*j* is the achievement level for person *j,*

ai is the discrimination parameter for item *I*,

bi is the difficulty or location parameter for item *i*,

ci is the lower asymptote or “pseudo-guessing” parameter for item *i*, and

D=1.7 to approximate the cumulative normal ogive

After calculating the item parameters, a random number seed file is implemented by using a random number routine. Three integer numbers are placed in a single line, and separated by spaces. For instance,

15424   1113   21032

After each run, the random number seed file is updated which ensures a different random sequence for each subsequent run. In CAT, IRT-calibrated items are included which comprise the CAT item bank. The post-hoc simulation is then applied to “re-administer” the examinees those items by using the responses they have already provided “as if” the item bank and various CAT procedures are administered (Weiss, 2005).

After the appropriate options for CAT were calculated, items were input into the program FastTEST Professional Testing System 2.0 (Weiss, 2008) to develop the CAT version of the dissertation completion survey. Surveys with a starting item representing medium and extreme trait levels were generated. First, the structured item bank was created, and items were imported into the software from an ASCII file which is generated in POSTSIM 2.0. After items are created and edited, the spelling of items was checked. Then, item statistics which were calculated from POSTSIM 2.0 were imported, and the test was assembled by applying IRT criteria with a desired test information function which provides the precision/information for a test as a function of the IRT θ (trait) variable. Further, the test standard error of measurement function, and the test response function as both expected number correct and expected proportion correct are also presented. Then, the CAT survey versions with different initial item trait levels were distributed to the target population. Reaction time for answering the items was recorded automatically by the computer. Time was recorded from when the item appeared on the screen until a response was confirmed by the participant. Two versions of the CAT survey were administered to the target population.

For administering the CAT versions of the survey, measures were administered using the researcher’s laptop. The survey was administered in a classroom or at the researcher’s office. The purpose of this study and response method was addressed before the survey begins. The forms of the survey were administered randomly. It took from five to seven minutes for participants to complete each form of the survey. Responses were confidential and were not available to college faculty, but were available to participants if they sent a separate email to the researcher requesting their score.

Analysis.

Descriptive statistics were provided. Information about frequency, mean, standard deviation, effect size, kurtosis, and skewness were examined. An independent-samples t-test was used to calculate the difference in test scores, length of tests, reaction time, and mean person parameters to the two CAT versions.

Chapter Four: Results

In this chapter, the research questions described in chapter 2 are addressed. Results are organized by phase.

Phase One

In phase one, due to the multi-faceted scales and measurement problems in the original survey format, a modified version of the dissertation/thesis completion survey was developed. Data collected by Green and Kuever (1997) were first recoded into dichotomous responses. Then, a principal components analysis (PCA) was applied to reduce survey dimensions in order to generate the unidimensional dissertation completion survey with dichotomous responses.

In the procrastination scale, the answer categories 1 and 2 were recoded into 0, and answer categories 3 to 5 were recoded into 1. Based on a PCA, 24 items were selected and reliability was .92. Higher scores indicated that participants agreed more frequently about the circumstances which items describe representing higher levels of procrastination. In the dissertation barriers scale, the answer categories -3 to -1 were recoded into 0 as hindrance, and answer categories “not a concern for you” and 1 to 3 were recoded into 1 as help. The category “not applicable to you” was recoded into missing. Twenty-two items were chosen based on PCA and reliability was .83. Higher scores indicated that test takers confronted these difficulties less representing a lower level of hindrance. In the responsibility scale, only the “IS” subscale was chosen. The 1, 2, and 3 responses were all recoded into 0 as student’s responsibility, and the 5, 6, and 7 responses were record into 1 as university’s responsibility. The response of fourth (middle) “×” was recoded into missing. Based on PCA, a total of 9 items were chosen, and reliability was .66 (Appendix A). Higher scores indicated a lower perception that the student should take responsibility for these tasks during the process of doing a dissertation/thesis.

Phase Two

In this phase, three conventional versions of the survey were generated. The item parameters were estimated by using the recoded responses to the dissertation/thesis completion survey collected by Green and Kluever (1997). The item parameters were calculated using PARSCALE 4 (Muraki & Bock, 2003). The two-parameter IRT was applied. In the procrastination scale, the item difficulty ranged from -1.815 to 1.457, and item discrimination ranged from .405 to 1.640. The item difficulty ranged from -2.705 to 1.378, and item discrimination ranged from .234 to 1.481 in the dissertation barriers scale. Finally, in the responsibility scale, the item difficulty ranged from .888 to 2.046, and item discrimination ranged from 1.168 to 2.807. Three versions of surveys with items ordered by trait levels (H-E, E-H, and M-R) were created.

Phase Three

In this phase, differences in scale reliabilities, test score, item difficulty, and item discrimination in three conventional survey versions were assessed. And, correlations were estimated between total score and attitude strength for each scale for each survey version and between total score and perception of effect of item order for each scale for each survey version.

Reliability.

For the three versions of the survey, the reliability estimates ranged from moderate to high. The reliability for the three survey versions for each scale ranged from: Procrastination scale, .83 to .90, Dissertation Barriers scale, .82 to .84, and Responsibility scale, .54 to .74. For each survey version, the lowest reliabilities were found for the responsibility scale. The lowest reliability of each scale appeared in the easy-to-hard survey version (Table 7).

Table 7.

Summary Statistics of Three Scales by Three Survey Versions

Scale	Procrastination			Dissertation Barrier			Responsibility		
Version	H-E	E-H	M-R	H-E	E-H	M-R	H-E	E-H	M-R
Mean	7.33	7.30	6.29	12.97	13.77	14.32	.54	.66	.83
SD.	5.88	4.69	5.05	4.78	4.70	4.78	1.05	1.07	1.45
Skewness	.76	.73	.82	-.23	-.36	-.33	2.86	1.86	3.11
Kurtosis	-.04	.36	-.25	-.79	-.40	-.69	9.57	3.34	12.40
Reliability	.90	.83	.87	.84	.82	.84	.63	.54	.74

Note. SD., standard deviation; H-E, hard-to-easy; E-H, easy-to-hard; M-R, medium-then-random.

Feldt's (1969) test for detecting the difference in reliability between the three survey versions was employed. Results indicated a statistically significant different in reliability in the procrastination scale in comparing H-E to E-H survey versions, *F* (131, 123) = 1.70, *p* < .05. In the responsibility scale, a statistically significant difference in reliability was also detected in comparing the H-E and M-R survey versions with *F* (117, 131) = 1.42, *p* < .05. Further, a statistically significant difference in reliability was also found in comparing the E-H and M-R survey versions, *F* (117, 123) = 1.77, *p* < .05. No other statistically significant difference in reliability was discovered for the dissertation barriers and responsibility scales among the three survey versions.

Test Score.

Scale test scores were calculated by summing the score for each item per scale. Although a higher average test score was found for the M-R survey version of each scale, no difference in test score between versions was statistically significant for any of the three scales: Procrastination, *F* (2, 328) = 1.34, *p* = .263, η^2 = .008, Dissertation Barrier, *F* (2, 336) = 2.30, *p* = .102, η^2 = .013, and Responsibility, *F* (2, 339) = 1.74, *p* = .177, η^2 = .010.

Item Difficulty.

The pairwise correlations between item difficulties for the three survey versions for each scale ranged from: Procrastination, .67 to .79, Dissertation Barriers, .45 to .71, and Responsibility, -.47 to .20. These pairwise correlations indicate that items are somewhat consistently ordered by difficulty across forms, indicating some level of invariance, but that the order is far from exactly the same by form.

Pairwise difference tests were employed to examine the difference in item difficulty for each item among scales for three survey versions. In comparing the item difficulty of each item in H-E to E-H survey versions, 5 (item 5, 8, 9, 11, 22) out of 24 items were found that differed statistically significantly at $p < .05$ in the procrastination scale. All items were easier to agree with in the H-E survey version. Five (item 5, 9, 11, 21, 23) items were discovered that differed statistically significantly in comparing in H-E to M-R survey versions, and all of these items were easier to agree with in the H-E survey versions. Finally, 3 items (item 2, 5, 8) were found differed significantly in comparing in E-H to M-R survey versions, with 2 of the 3 being easier to agree with in the E-H version (Table 8).

Table 8.

Item Difficulty and Item Discrimination of Procrastination Scale for each Survey Version

	Item Difficulty			Item Discrimination		
	H-E	E-H	M-R	H-E	E-H	M-R
Item 1	0.18	0.50	0.12	1.35	*0.39	*0.90
Item 2	0.41	*0.28	*0.64	1.24	0.67	1.12
Item 3	0.88	1.01	1.24	1.81	2.15	1.63
Item 4	1.49	2.53	1.25	0.81	0.43	0.82
Item 5	*-0.10*	*0.27	_*0.81_	1.25	0.94	0.67
Item 6	-0.29	-0.07	-0.18	1.42	0.94	1.08
Item 7	0.60	1.32	1.12	1.03	*0.81	*1.45
Item 8	0.81	*1.96	*0.66	1.91	*0.73	*2.91
Item 9	*0.73*	1.83	*1.37*	2.86	1.54	1.39
Item 10	-0.23	-0.10	0.43	1.03	0.76	0.40
Item 11	*-0.09*	0.32	*0.40*	1.78	*0.79	*1.33
Item 12	0.85	1.27	1.47	*2.32*	1.24	*0.97*
Item 13	0.31	0.64	0.37	1.67	0.73	1.04
Item 14	0.45	0.57	0.40	1.17	0.92	1.32
Item 15	0.12	-0.48	-0.80	0.46	0.34	0.17
Item 16	1.25	6.27	1.35	1.19	*0.36	*1.70
Item 17	1.04	1.80	1.35	1.53	1.05	1.48
Item 18	-0.18	-0.17	-0.10	1.44	0.67	1.36
Item 19	0.48	0.40	0.39	*1.53*	*1.08	_*2.71_
Item 20	-0.12	-0.50	-0.15	0.68	0.70	0.64
Item 21	*0.06*	0.39	*0.58*	1.32	0.90	0.82
Item 22	0.18	0.51	0.44	1.50	1.50	1.51
Item 23	*0.01*	0.38	*0.46*	1.09	0.91	0.98
Item 24	-0.10	0.30	0.13	0.82	0.40	0.70

Note. H-E, hard-to-easy; E-H, easy-to-hard; M-R, medium-then-random; Underscore, significant difference in comparing H-E and E-H; Italic, significant difference in comparing H-E and M-R; *, significant difference in comparing E-H and M-R.

For the dissertation barriers scale, 5 (item 7, 12, 15, 16, 19) of 22 items were discovered that differed statistically significantly in item difficulty in comparing H-E to E-H survey versions, with 4 of the 5 being easier to agree with in the E-H version. Four (item 12, 18, 19, 22) items were discovered that differed statistically significantly in comparing H-E to M-R survey versions. Three of the 4 items were easier to agree with in the M-R version. Statistically significantly different item difficulties were found for 6 (item

2, 3, 10, 16, 18, 20) items in comparing E-H to M-R survey versions, and 5 of the 6 items were easier to be agreed with in the M-R version (Table 9).

Table 9.

Item Difficulty and Item Discrimination of Dissertation Barriers Scale for each Survey Version

	Item Difficulty			Item Discrimination		
	H-E	E-H	M-R	H-E	E-H	M-R
Item 1	5.02	1.27	0.45	_0.10_	0.24	_0.59_
Item 2	-2.51	*-1.81	*-2.80	0.69	0.96	0.54
Item 3	0.10	*-0.74	*1.02	0.37	0.24	0.25
Item 4	-0.26	-0.50	-0.52	1.29	0.71	1.27
Item 5	-0.02	0.01	-0.02	2.47	1.61	2.55
Item 6	-1.04	-1.32	-1.19	0.53	0.62	0.47
Item 7	-0.12	-1.20	-0.48	0.61	0.41	0.53
Item 8	-0.83	-0.40	-0.81	0.56	0.98	0.66
Item 9	-1.22	0.00	-1.58	0.54	0.01	0.27
Item 10	0.00	*2.09	*0.00	0.04	*0.47	*0.05
Item 11	1.39	0.26	0.26	0.41	0.25	0.61
Item 12	_1.00_	0.13	_0.22_	1.11	0.95	1.30
Item 13	-0.49	-0.43	-0.82	0.83	0.61	0.36
Item 14	-0.39	-4.87	-1.58	0.30	0.08	0.29
Item 15	-0.57	-1.16	-1.04	_2.38_	0.85	_1.13_
Item 16	-0.57	*-0.11	*-0.44	1.32	4.33	2.26
Item 17	-1.11	-0.84	-1.87	0.41	0.38	0.54
Item 18	_-0.56_	*-0.51	*_-0.93_	1.36	1.22	1.43
Item 19	_1.53_	0.22	_0.08_	0.41	0.48	0.88
Item 20	-1.26	*-0.89	*-1.95	0.65	0.69	0.52
Item 21	-0.25	-0.13	-0.13	1.14	1.19	1.92
Item 22	_0.19_	0.09	_-0.05_	_3.42_	1.89	_1.90_

Note. H-E, hard-to-easy; E-H, easy-to-hard; M-R, medium-then-random; Underscore, significant difference in comparing H-E and E-H; Italic, significant difference in comparing H-E and M-R; *, significant difference in comparing E-H and M-R.

In the responsibility scale, no items were discovered with statistically significant different item difficulties (Table 10).

Table 10.

Item Difficulty and Item Discrimination of Responsibility Scale for each Survey Version

	Item Difficulty			Item Discrimination		
	H-E	E-H	M-R	H-E	E-H	M-R
Item 1	0.86	3.68	1.49	1.23	0.31	0.78
Item 2	1.50	1.24	1.07	2.16	2.12	1.72
Item 3	1.16	1.71	1.33	2.27	1.38	1.18
Item 4	2.35	2.10	1.41	0.65	0.70	1.18
Item 5	1.72	1.43	1.60	1.35	1.37	1.12
Item 6	1.56	2.37	2.07	1.77	1.02	1.26
Item 7	1.30	5.56	1.27	2.39	0.38	2.12
Item 8	1.49	4.78	1.48	1.69	0.00	2.33
Item 9	2.40	1.12	0.78	0.43	1.02	0.91

Note. H-E, hard-to-easy; E-H, easy-to-hard; M-R, medium-then-random.

Item Discrimination.

The correlations between item discrimination for the three survey versions for each scale ranged from: Procrastination, .31 to .62, Dissertation Barriers, .52 to .76, and Responsibility, -.26 to .58. This means that item discrimination is somewhat invariant when items were presented in different orders. Furthermore, by splitting the sample for each survey version, correlations of item discrimination between the two split-samples for each scale of every survey version all presented significant relationships, except for the procrastination and responsibility scales in the M-R survey version (Table 11) which indicated that item discrimination difficulty is also invariant within most but not all survey versions.

Table 11

Correlations of Item Discrimination between Two Split-Samples for Each Scale

	Procrastination	Dissertation Barriers	Responsibility
H-E	.70**	.99**	1.00**
E-H	.97**	.59**	.78*
M-R	.87**	.30	.40

Note. * *p* < .05; ** *p* < .01; H-E, hard-to-easy survey version; E-H, easy-to-hard survey version; M-R, medium-then-random survey version.

Pairwise difference tests were then employed to examine the difference in item discrimination for each item among the three survey versions. In the procrastination scale, 6 (item 1, 6, 8, 11, 13, 18) out of 24 items differed significantly in item discrimination in comparing H-E to E-H survey versions. Lower item discriminations were discovered in all of these items for the E-H version. Two (item 12, 19) items were found that differed statistically significantly in comparing H-E to M-R survey versions; 1 of the 2 had a lower item discrimination in the M-R version. Six (item 1, 7, 8, 11, 16, 19) items differed in item discrimination in comparing E-H to M-R survey versions. All of these items had lower item discrimination in the E-H version (see Table 8).

For the dissertation barriers scale, 6 (item 4, 9, 10, 15, 16, 22) out of 22 items differed significantly in comparing the H-E to E-H survey versions, with 4 of the 6 having lower item discrimination in the E-H version. Three (item 1, 15, 22) items showed significantly different item discriminations in comparing H-E to M-R survey versions, with 2 of the 3 having lower item discrimination in M-R version. In comparing with E-H and M-R versions, 1 (item 10) item differed statistically significantly with lower item discrimination in the M-R version (see Table 9). Results showed no items differed statistically significant in the responsibility scale (see Table 10).

Correlations between Perception of Effects of Item Order, Scale Test Score, and Scale Attitude Strength.

The test score was calculated by summing the score of each item per scale. In the H-E version, participants' perception of whether their answers were influenced by the item order was significantly correlated with total score for the procrastination (r = -.22, p < .05) and dissertation barriers (r = .28, p < .05) scales. Statistically significant correlations were found between reported attitude strength for each pair of scales. However, no statistically significant relationships were found between reported attitude strength and scale total score for any scale (Table12).

Table 12.

Correlations between Scale Score and Scale Attitude Strength in Hard-to-Easy Version Survey

	T-Proc	T-Bar	T-Resp	OE	A-Proc	A-Bar
OE	-.22*	.28**	.07			
A-Pro	-.01	.12	.06	-.11		
A-Bar	.13	-.12	.14	.02	.53**	
A-Resp	-.08	.16	.00	.06	.54**	.28**

Note. * *p* < .05; ** *p* < .01; T-Pro, total score on procrastination scale; T-Bar, total score on dissertation barriers scale; T-Resp, total score on responsibility scale; OE, perception of order effect; A-Proc, attitude strength toward procrastination; A-Bar, attitude strength toward dissertation barrier; A-Resp, attitude strength toward responsibility.

In E-H version, results indicated a statistically significant relationship between total score and attitude strength for the procrastination scale (*r* = -.23). Participants' perception of order effect was also found to be statistically significantly correlated with attitude strength for the responsibility scale (*r* = .22). And, attitude strength for each scale was significantly correlated with attitude strength for the others (Table 13).

Table 13.

Correlations between Scale Score and Scale Attitude Strength in Easy-to-Hard Version Survey

	T-Proc	T-Bar	T-Resp	OE	A-Proc	A-Bar
OE	-.16	-.05	-.16			
A-Pro	-.23*	.13	-.08	.17		
A-Bar	.10	-.19	-.04	.03	.44**	
A-Resp	.01	.00	-.08	.22*	.60**	.40*

Note. * *p* < .05; ** *p* < .01; T-Pro, total score on procrastination scale; T-Bar, total score on dissertation barriers scale; T-Resp, total score on responsibility scale; OE, perception of order effect; A-Proc, attitude strength toward procrastination; A-Bar, attitude strength toward dissertation barrier; A-Resp, attitude strength toward responsibility.

In the medium-then-random version, participants' perception of order effects was statistically significantly correlated with total score on the procrastination (*r* = -.36) and responsibility (*r* = -.25) scales. Also, statistically significant correlations were found for attitude strength between each pair of scales. However, no statistically significant relationship was found between scale attitude strength and total score (Table 14).

Table 14.

Correlations between Scale Score and Scale Attitude Strength in Medium-then-Random Version Survey

	T-Proc	T-Bar	T-Resp	OE	A-Proc	A-Bar
OE	-.35**	.17	-.25**			
A-Pro	.11	-.03	-.14	-12		
A-Bar	.22*	-.13	-.13	.14	.55**	
A-Resp	.04	-.08	-.15	.13	.59**	.51*

Note. * *p* < .05; ** *p* < .01; T-Pro, total score on procrastination scale; T-Bar, total score on dissertation barriers scale; T-Resp, total score on responsibility scale; OE, perception of order effect; A-Proc, attitude strength toward procrastination; A-Bar, attitude strength toward dissertation barrier; A-Resp, attitude strength toward responsibility.

Phase Four

In this phase, CAT versions of the survey with items beginning at medium or extreme trait levels were generated and administered. First, the post-hoc simulation was conducted by using POSTSIM 2.0 (Weiss, 2005). The phase one sample was used to calculate the minimum standard error of each scale. The calculated standard errors of the three scales ranged from .19 to .31. Therefore, the termination rule for all scales was fixed with standard error ≤ .5. Then, CAT surveys were designed using the program FastTEST Professional Testing System 2.0 (Weiss, 2008). Based on the item difficulty range of the three scales, the initial starting value was set at θ = 0.0 for the medium version and θ = 1.3 for the extreme version where θ is the person ability level.

In this exploratory study, descriptive statistics were calculated for both CAT versions (Table 15). A higher average value for test length (number of items administered), test score, and reaction time were found in the survey beginning with an extreme trait level, except for the responsibility scale. A higher average person parameter was found for all scales beginning with a medium trait level.

Table 15.

Summary Statistics of Two Computerized Adaptive Survey Versions

		Mean		SD		Skewness		Kurtosis	
Scale	Var	Me	Ex	Me	Ex	Me	Ex	Me	Ex
Proc	TL	5.73	12.53	2.37	8.63	1.22	.62	-.10	-1.67
	TS	2.33	4.07	.62	5.01	-.31	3.75	-.40	14.34
	RT	70.13	121.20	35.46	82.30	1.24	.86	1.07	-.31
	PA	.15	-.04	1.63	1.28	-.27	.23	-1.80	-.91
Bar	TL	6.47	11.60	1.25	7.90	-.30	.56	.47	-1.65
	TS	3.27	7.60	1.10	8.39	-13	1.17	-1.34	-.71
	RT	46.27	86.80	12.79	83.46	.55	2.22	.10	5.29
	PA	-.87	.422	-.95	2.06	1.33	.21	1.60	-1.11
Resp	TL	5.87	7.67	2.75	2.19	.26	-1.81	-2.04	2.35
	TS	.93	.60	2.50	.63	2.50	.55	7.67	-.39
	RT	53.13	56.93	28.31	22.99	.94	.600	-.07	-.68
	PA	-.92	-1.41	2.62	2.51	-.40	-.14	-2.04	-2.29

Note. Proc, procrastination; Bar, dissertation barrier; Resp, responsibility; Var, Variable; SD., standard deviation; TL, test length; TS, test score; RT, reaction time; PA, person parameter; Me, medium; Ex, extreme.

Test Length.

Comparisons of test length per scale for the two CAT versions indicated statistically significant differences for the procrastination (t = -2.94, p = .009, η^2 = .236) and dissertation barriers (t = -2.49, p = .025, η^2 = .181) scales. For these two scales, in order to achieve a set level of precision, more items were required for the version that began at the extreme difficult trait level than the version at the medium difficult trait level

($M_{\text{extreme-procrastination}}$ = 12.53; $M_{\text{medium-procrastination}}$ = 5.73; $M_{\text{extreme-dissbarriers}}$ = 11.60; $M_{\text{medium-dissbarriers}}$ = 6.47).

Reaction Time.

A statistically significant difference in reaction time was found for the procrastination scale (t = -2.21, p = .040, η^2 = .149). For this scale, reaction time was significantly shorter in the version starting with items representing a medium trait level than for the version with items beginning at an extreme trait level. The same direction of effect was also found for the other two scales but differences were not statistically significant.

Test Score.

As for the conventional survey format, scale test scores were also calculated by summing the raw score for each item per scale. While a lower test score for the version starting with a medium trait item was found for the procrastination and dissertation barriers scale, it was not for the responsibility scale. No difference between versions was statistically significant for any of the three scales: Procrastination, t (28)= -1.33, p = .194, η^2 = .060, Dissertation Barriers, t (28)= -1.98, p = .057, η^2 = .123, and Responsibility, t (28) = .90, p = .374, η^2 = .028.

Person Parameter.

Person parameter refers to the person's latent trait as calculated based on Item Response Theory (IRT) and differs from total score. In the CAT survey, test takers received different items based on their response to the current question, so it is possible that two test takers answered different numbers of items but obtained the same total score (raw score). However, based on the varied difficulty of items they answered and different response patterns, the calculated person parameter diverges from the total score. In this dissertation, a significant effect on person parameter was found for the dissertation barriers scale (t = -2.21, p = .039, η^2 = .149). For this scale, the mean person parameter was significantly lower for the version starting with items representing an extreme trait level than for the version with items beginning at a medium trait level. No significant effects were found for the procrastination and responsibility scales, but the average person parameter was also lower for these two scales.

Table 16.

t-Test Differences between Two Computerized Adaptive Survey Versions

Scale	Variable	*t*	*p*
Procrastination	TL	-2.94	.009
	TS	-1.33	.194
	RT	-2.21	.04
	PAR	.528	.603
Dissertation Barriers	TL	-2.49	.025
	TS	-1.98	.057
	RT	-1.86	.083
	PAR	-2.21	.039
Responsibility	TL	-1.98	.058
	TS	.90	.374
	RT	-.40	.690
	PAR	.52	.607

Note. TL, test length (in number of items administered); TS, test score (in number of items scored 1); RT, reaction time (in seconds); PAR, person parameter (in logits).

Chapter Five: Discussion

Attitude measures are effective tools to collect self-report data. In attitude measures, items are generated reflecting similar content in order to assess attitude of a person about an issue, event, or product. In measurement, test takers' consistency and appropriate response is the essential element to determine whether the answers are valid. It is important for researchers to investigate factors which may influence response patterns, and whether this kind of impact is ignorable or not. If item order effects exist, different response patterns may occur by changing the presentation order of items. Several theories were proposed to explain this phenomenon, such as anchoring-and-adjustment, attitude accessibility, primacy and recency, and attention decrement.

For a long time, researchers tried to understand what makes people's attitude change. Studies of the relationship between attitude change and item order were few in numbers. Results of investigating item order effects were diverse, especially for achievement tests. But more consistently significant effects of item order were discovered for studies involving attitude measures. In this dissertation, the effects of item order of attitude measures with different item arrangements in both conventional and computerized adaptive forms were assessed.

The results of this dissertation suggest that item order is a factor influencing survey responses. In this chapter, the results of this dissertation are discussed from several aspects. Discussion is based on the survey format (conventional and computerized adaptive). For each format of the survey, the most important findings and the possible causes are summarized. Limitations and other potential moderators are proposed regarding the direction for future studies.

Conventional Survey

The purpose of the first part of this dissertation was to investigate the effects of item order in attitude measures administered via a conventional survey format. It was hypothesized that differences in item difficulty, item discrimination, and test scores would be discovered in a survey with items ordered by different (hard-to-easy, easy-to-hard, and medium-then-random) trait levels. Based on the extant data collected by Green and Kluever (1997), items were calibrated and three versions of a modified dissertation/thesis completion survey were created and administered.

In the three conventional survey versions, the dropout rate was diverse. The lowest dropout rate was discovered in the M-R (32%) survey version, and the highest dropout rate was found in H-E (64%) survey version. This suggests that the item presentation order may influence the survey completion rate. People may be more willing to answer the survey when it starts at the medium trait level as compare to begin at extreme (hardest or easiest) trait levels. On the other hand, the mean score for each scale suggest that respondents on average reported less procrastination, perceive more help than hindrance, and indicated that they should take more responsibility than advisor/university while doing a dissertation/thesis. Analysis of data from the three survey versions showed effects due to changes in item order. Moreover, significant differences were found for item difficulty and item discrimination in two (procrastination and dissertation barriers scales) of three scales between three survey versions. In addition, statistically significant differences in scale reliabilities and correlations between scale test score and scale attitude strength were also detected.

Reliability.

In this research study, scale reliabilities were statistically significantly different for the same survey with different item orders. Two scales were found with significantly different reliabilities in pairwise comparisons between three survey versions. The highest reliability for the procrastination scale was found in the H-E version. For the dissertation barriers scale, both H-E and M-R versions had higher reliabilities. The M-R version had the highest reliability for the responsibility scale. The lowest scale reliabilities were all found in the E-H version. In 1988, Knowles conducted a similar study and concluded that item position is statistically significantly related to item reliability. For both studies, the test reliability was affected by the item presentation orders.

In the present study, a lower scale reliability was discovered for the responsibility scale in every survey version. This scale was placed at the end of each survey version. This outcome may be explained in several ways. In order to fulfill the assumption of unidimensionality, principal components analysis (PCA) was conducted, and only nine items were selected for the responsibility scale. The length of the responsibility scale was shorter compared to other two scales which is one reason the scale had lower scale reliability. Second, the effect of attention decrement may have influenced this outcome. In taking a survey, it is likely that people may feel tired by the end of the survey. Test takers' attention may have decreased when processing information toward the end of the survey compared to the beginning. In this case, responses to items appearing earlier in a survey may be more consistent than later ones. Therefore, higher reliability might be obtained for scales or items when they are placed at the beginning of a survey. Discussions of attention decrement can be found in Anderson and Jacobson’s (1965) study. Results here suggest that altering item presentation orders may generate problems with test reliability.

Item Difficulty.

For item difficulty, the results of this dissertation support the hypothesis that item difficulty is different when items are presented in different orders. Results here are consistent with those of Frantom et al. (2002), Ortner (2004), and Pomplun and Ritchie (2004) who found that item difficulty changed when the item presentation orders were altered. Though not all items or every scale version comparison resulted in the hypothesized effects, some similar patterns were discovered. In the procrastination and dissertation barriers scales, both positive and negative item wording were applied. In this dissertation, results showed an association of the disparities in item difficulty and the direction of item wording.

In the procrastination scale, in comparing H-E to E-H and H-E to M-R survey versions, 4 of the 5 items identified as being significantly different in logit position (difficulty) were worded in a negative direction while the items preceding them were also negatively worded. In comparing E-H and M-R survey versions, 3 items identified with significantly different difficulty were all also worded in a negative direction which was the same as the wording of the preceding items. The opposite direction was detected in the dissertation barriers scale. When comparing H-E and E-H survey versions, 3 of the 5 items identified as being statistically significantly different in difficulty were worded in a positive direction. Further, when comparing H-E and M-R survey versions, all items identified with significantly different difficulty were all positively worded. Finally, 5 of the

6 items with significantly different difficulty were also worded positively when comparing E-H and M-R survey versions. Research about use of positive or negative item wording can be found in Benson and Hocevar's (1985) and Deemer and Minke's (1999) studies. Results here suggest that using different item wording directions may generate problems in terms of item functioning with respect to invariance, and that the impact of wording is complex.

In the procrastination scale, when comparing H-E to E-H and H-E to M-R versions, all items identified as being significantly different in item difficulty were easier to agree with on the H-E version. Two of the 3 items identified as statistically significant different in item difficulty were easier to agree with on M-R version when comparing E-H and M-R survey versions. In the dissertation barriers scale, four of 5 items identified as being significantly different in item difficulty were easier to agree with in E-H version when comparing H-E to E-H survey versions. Further, when comparing H-E to M-R and E-H to M-R versions, most items identified as significantly different in item difficulty were easier to agree with in M-R survey version.

Results here suggest that item order is an issue influencing participants' responses. When a survey begins with an extremely difficult trait level of items (very difficult or very easy items), survey takers may establish a boundary or an anchor based on these extreme items. Survey takers may then calibrate the following items in accordance with this anchor, which is extreme. For instance, when items were ordered from hard to easy, the item listed first is the most difficult one to agree with, and this item may serve as the anchor for subsequent items. People may either assimilate the following responses in accordance with preceding items or suppress the anchor idea to select a contrast category for the later items. Further, the item input order also influences people's recall strategies. Feelings or attitude toward the specific topic tend to be activated by the first item presented (Siminski, 2008; Tan & Ward, 2007). Different emotions will then be aroused based on the anchor item for different people. In this case, the responses to following items might be adjusted. Similar response processes can also be applied to the E-H and M-R survey versions. The current study provides added evidence suggesting item order may be a factor limiting use of attitude assessments with varied item orders.

Item Discrimination.

It was hypothesized that item discriminations would be different when items were arranged in different orders. Based on the analyses of data from three groups of participants, results showed effects due to changes in item order: statistically significant differences were found for item discrimination in two (procrastination and dissertation barriers scales) of the three scales between the three survey versions. This result is consistent with one found in one of Brenner's (1964) four experiments. Results here support the hypothesis that item discrimination would be influenced by changing item presentation orders. Although significantly different item discrimination was not found for every item of each scale for all version comparisons, a pattern was found in the results.

Except for the comparison between the hard-to-easy (H-E) and easy-to-hard (E-H) survey versions, the items with the most extreme discrimination (highest and lowest) of each survey version were all identified as being statistically significant different in discrimination for all version comparisons in both scales. In each pairwise comparison, items with the most extreme discriminations differed statistically significantly in discrimination. This phenomenon might suggest that the extreme discriminations become more unstable when altering the item presentation orders. Results here indicated that item order may be a factor influencing item discrimination and so limiting use of attitude assessments with varied item orders.

Test Score.

Based on the idea of anchoring-and-adjusting, it was hypothesized that test scores would be different when items are ordered in different difficulty sequences. In this current study, analyses of data from three groups of real persons presented no significant effect on scale test score by changing the orders based on item difficulty. No significant difference in test score of each scale for three survey versions was discovered. Klimko (1984), Plake (2002), and Monk and Stallings (1970) found similar results in their research on paper-and-pencil tests, but contrasting results were detected by Marso (1970) and Newman (1988) and his colleagues with the same test format (paper-and-pencil). Results here did not support the hypothesis of anchoring-and-adjusting effects associated with item order on test scores. This outcome may potentially be due to no item order effects overall or to a lack of dispersion in the response scale, which was dichotomous.

In this dissertation, the response category of the extant data was recoded into dichotomies in order to avoid the problem of finding an appropriate model or clustering information in the suitable category. However, this may also bias responses in that sometimes people might not have a strong perception about items. The variability of small differences in perceptions toward these items was lost by using the dichotomous response categorization, and it is difficult to reflect these variances on test scores. This is a potential reason that significant differences in item difficulty and item discrimination were discovered but significant differences in test scores were not.

Based on the outcomes of present study, it indicated that if the result of a measure is only explained by the survey total score, then the practitioner may not need to consider issues of item order effects. Therefore, when the purpose of a survey is, for example, only to understand the customers' overall satisfaction toward the shopping experiences and the researcher does not care about item statistics, and the total score is the only index used to explain the result, item order may not be pertinent. Item order effects would not be a problem. Different versions of a survey with different item arrangements can be used. However, if the purpose of a survey is to comprehend customers' evaluations toward each item not only overall satisfaction, the issues about effects of different item arrangements should be considered, and different forms of a survey with different item presentation orders are not recommended.

Correlations between Scale Test Score and Scale Attitude Strength.

Results of the current research showed that participants' perceptions of whether their answers were influenced by item order correlated with the scale total score for two out of three scales for the H-E (procrastination and dissertation barrier scales) and M-R (procrastination and responsibility scales) survey versions. Results suggest that participants can reflect about the impact of changing item order on the way they respond and it correlates with the scale total score. For this reason, it is reasonable to hypothesize that participants might alter their response strategies if they perceive the issue of item order effects before answering the survey. It is possible that different performance on a survey might be found based on whether test takers pay attention to this issue or not. The idea of informing respondents about a specific issue to make participants notice or expect it before taking a survey had been studied by Ofir and Simonson (2007). They found statistically significantly lower satisfaction if customers were asked to pay attention to their purchase experiences before taking a survey.

A statistically significant *negative* correlation between the scale total score and attitude strength was detected for the procrastination scale E-H survey version. This result indicates that when test takers' attitude toward the issue of time management (procrastination) was strong, they tended to disagree with the concerns listed in the survey, i.e., expressed lower levels of procrastination. However, even though no other significant relationship between these two variables was found, the correlations between these two variables were negative in most scales for the three survey versions. For the dissertation barriers scale, participants tended to report that they confronted more hindrances when their attitude toward the listed difficulties was strong. For the responsibility scale, respondents felt that they should take more responsibility during the process of doing a dissertation/thesis when they had a stronger attitude toward the issues about where these responsibilities rest. Results suggest that scale total score might relate to people's attitude strength toward that specific topic. Higher scale total scores tended to be reported when participants had weaker attitudes toward the topic.

For this survey, higher or lower scores have different meanings for each scale. When participants had higher scores for every scale which indicated that they were more procrastination, more help, and took less personal responsibility, so scales were oriented in different directions. That is when test takers have strong attitude for each topic of this survey, they tend to report less procrastination, perceived more hindrance than help, and took more personal responsibility than others during the processes of doing a dissertation/thesis.

Computerized Adaptive Survey

The second part of the dissertation was an exploratory study investigating whether different performance occurs by altering initial item entry levels. It was hypothesized that different mean person parameters, test length, test score, and reaction time would be discovered in surveys with item starting at different difficult trait levels. Based on the extant data collected by Green and Kluever (1997), items were calibrated and two versions of a modified computerized adaptive dissertation/thesis completion survey were created and administered. For both survey versions, the mean person parameter of each scale suggests that respondents agreed more with the listed issues with respect to time management (procrastination), and perceived more difficulties when surveys start at medium trait levels. However, respondents all indicated that they should take more responsibility than advisor/ university during the process of doing a dissertation/thesis for both survey versions. Significant differences were discovered for mean person parameters, test length, and reaction time between two survey versions. But, no statistically significant difference in scale test score was found.

Person Parameter.

The aim of this exploratory study was to investigate the effects of changing initial item difficulty trait levels in attitude measures of computerized adaptive testing. It was hypothesized that different mean person parameters would be found with items starting at different trait difficulty levels. People may tend to agree more when a survey begins at a medium difficulty level as compared with items starting at an extremely high difficulty level. In this dissertation, the analysis of data from two groups presented significant effects of changing initial item difficulty levels. This result is consistent with that of Ortner (2008).

For one (dissertation barriers scale) of the three scales, a statistically significantly different person parameter was found between two survey versions. Person parameters were statistically significantly lower in the version with items beginning at the medium level. In this scale, items were asked about what difficulties people had confronted during the processes of doing a dissertation/thesis. Based on the coding scheme, a lower score in this scale indicated they confronted more difficulties. Therefore, the lower person parameters in this scale represented that people agreed with or confronted more difficulties. Results of this study also supported the hypothesis that people tend to agree less when items started at an extremely high trait level. Although this hypothesized effect was not found for all scales for both survey versions, the results were uniformly in the same direction: the performance of answering a survey items was different when altering item difficult entry levels in CAT. People tended to agree less in surveys starting with items representing an extremely difficult trait level.

The results can be predicted from an anchor-and-adjust perspective. The initial item seems to provide the mental boundary for test takers. The item listed first served as the anchor for participants. The anchor item is then applied as a standard against which participants evaluate the following items. Further, the first item also provides the starting point for participants to recall their feelings or experiences toward the specific topic. Different emotions or attitude are aroused by the first item for different participants. Adjusting then applied; the responses of subsequent items would be shifted based on test takers' attitude toward this anchor item. Different response patterns emerged as a function of which items were presented first. Results here are consistent with those of Ortner (2008) and Siminski (2008), and discussions of this idea can also be found in their research.

Test Length and Test Score.

In this exploratory study, differences in test length (numbers of item administered) were also assessed. The result indicated a statistically significant difference in test length to achieve the same precision on two (procrastination and dissertation barriers scales) out of three scales between the two survey versions. Although no significant result was found for the responsibility scale, results were in the same direction as for the two other scales with more items required for the version with items starting at an extreme trait level. This outcome may have occurred because the extreme test conditions target the examinee's attitude inappropriately. It is possible that the test taker's attitude is far from the administered item difficulty. Therefore, more items are required to achieve the specific level of precision under extreme (very easy or very difficult) test conditions. Results here support Wright and Stone's (1979) idea that the greater the distance between item difficulty and examinee's ability, the more items are needed to achieve comparable precision.

In the current study, according to descriptive analysis, two (procrastination and dissertation barriers scales) out of three scales were detected with the average higher scale scores, but none of the scale scores were statistically significantly different for these two survey versions. Even though a significantly higher numbers of items needed to be administered to achieve comparable precision for the survey starting at an extremely difficult trait level, the scale test score did not differ. Results of the given study indicated that scale test score is not influenced by item entry levels in CAT which is consistent with Knowles' (1988) finding that item positions had no significant effect on the mean answers. Therefore, in order to administer CAT more efficiency, it is important to providing a suitable starting level for test takers.

Reaction Time.

Based on Fazio's (1990) hypothesis, when an extremely difficult trait level item is presented, the test taker might respond to the item more cautiously in order to avoid inappropriate or unacceptable answers. Therefore, more time might be spent on evaluating the response categories, and a longer reaction time is needed. In this study, the average and difference in reaction time per scale in two survey versions were estimated. The descriptive analysis indicated that a longer reaction time was required in the version with items starting at the extremely high difficult trait level for each scale. Although a statistically significant difference in reaction time was only found for the procrastination scale, the current result confirmed the hypothesis that longer reaction time is required when a survey starts with items representing an extremely difficult trait level. People need to spend more time evaluating the response categories in order to avoid a socially undesirable response when confronting the extreme item. Further, according to results of test length in this dissertation, a longer test is necessary in order to achieve the specific precision level for a survey with item starts at extreme condition. More time will be spent to administer the longer test. This outcome is consistent with that of Vega and O'Leary (2006) and Ortner (2008). Results suggest that this effect might be caused by effects of changing item order.

Based on the results of both survey formats, survey test score seems not to be impacted by different item arrangements. Practitioners don't need to be concerned about the item order effects if the measure is explained by the overall test score, and surveys with different item arrangements can be recommended. However, if the purpose of the survey is to realize the attitude or evaluation toward each specific item, the item order effect should be taken into consideration. The present dissertation indicated that changing item presentation orders results in different item statistics (item difficulty and discrimination). Results here found that the item difficulty and discrimination were altered if the sequence of item presentation was changed. Further, test takers might have different perceptions toward the same item when it is presented in different orders. It is possible that the results of psychometric indices are different in an assessment with different item arrangements. The use of only one form attitude instruments is suggested under this kind of condition.

However, only one form of a survey is applied in the most situations. In this case, if the total score is the only index to explain the survey result, both conventional and computerized adaptive testing format surveys are suggested. But, in order to administer the survey more efficiently and obtain a higher response rate, a survey beginnings at the medium difficult trait level is recommended. On the other hand, if the purpose of the survey is to evaluate respondents' perceptions or attitude toward each item, then the influence of item parameters should be taken into consideration, and only a conventional survey format with items start at the medium difficult trait level is recommended.

Limitations and Future Study

In this dissertation, some limitations exist which can be addressed in future studies. The purpose of the current research was to investigate the effects of item order on attitude measures. By examining the effects of item arrangements, it is assumed that all participants answered the survey questions following the sequences of the order in which items were presented, and this condition is very difficult to control in conventional surveys. In this case, ensuring that test takers answer the questions in the proper order would be an issue for future research with a conventional survey format. In order to rule out this limitation, two computerized adaptive versions of survey were also conducted in this dissertation which added evidence that item presentation order may be a factor affecting use of attitude assessments.

Although the main findings in the CAT study presented significant evidence of item order effects, there are some limitations which can be addressed in future research. First, the item pool used in this dissertation did not contain a sufficient number of items. A larger numbers of items and sample size are desirable which would result in a smaller standard error in future research. Second, current results indicated that reaction time is different when the level of initial item difficulty is altered, but the relationship between attitude strength and reaction time remains unknown. Third, the effects of changing item entry levels on the degree of attitude accessibility in attitude measures for CAT can also be examined. Finally, the investigation about whether controversial performances appear when item content is either more or less salient to the respondent can be conducted.

For either conventional or computerized adaptive surveys, the impacts of potential moderators on the described effects should be examined in the future research. For instance, do personological variables relate to the strength of item order effects, are the described effects influenced by the passage of time (e.g., a longer or shorter time interval) since a specific event or by the complexity of survey context. Relationships between these variables and the described effects can be probed in future investigations.

Also, it is possible that participants in different stages (e.g., doing or finished with the dissertation/thesis) might have different perceptions toward the specific topics. Thus, use of focus groups might provide an in-depth investigation to see whether stage differences would be a factor impact the responses. Moreover, the distribution of item variability can also be examined to see whether the variance of each item is altered when items are presented in different orders. This dissertation maintained a focus on effects on item and overall means so the study could be replicated in part with item and overall variance as the focus. Further, results of the present dissertation presented that item statistics are altered when survey items are arranged in different orders which indicates that item order effects constitute a violation of local independence and so violate a basic assumption of IRT. The phenomenon about violation of local independence by changing item presentation order in other domains of assessments (e.g., achievement, aptitude, or personality test) should be examined in the future. Finally, rating scales without a middle option is also suggested for the future research. Although the yes-no response category can differentiate people's tendency or attitude very clearly, the degree of variability obtained by using multiple categories is lost. It is difficult to understand the degree of the respondent's perception toward an item. In this case, use of a rating scale without an ambiguous middle option, such as "I don't know", is a way to avoid the measurement problem and help to obtain more variance. However, as software POSTSOM 2.0 and FastTEST Professional Testing system 2.0 can only be applied to dichotomous responses, the extension of the software for polytomous response categories would also be necessary.

Conclusion

Attitude measures are still the most effective tools to collect self-report data of people's feeling and attitude toward things. Studies exploring the mechanism of how attitude changes have been conducted for several decades, but research that focused on the relationship between attitude and item presentation order is deficient. Diagnosis of effects of item presentation order on attitude assessments is important for many reasons. Most importantly, if different item arrangements really impact people's responses to a survey or test, the measurement efficiency of CAT would be doubted and the merits of parallel tests vanish.

Effects of item order on attitude measures were examined on several variables in this dissertation: test reliability, item difficulty, item discrimination, test score, test length, reaction time, and person parameters. Analysis of real data from both conventional and computerized adaptive surveys supported the hypothesis that item order is a factor limiting the use of attitude measures. Items presented first might serve as an anchor for the subsequent questions. People tend to adjust their responses to the following questions based on preceding items.

According to the results of this dissertation, evidence of order effects on attitude measures was provided. For conventional surveys, results showed that different sequences of item difficulty orders influenced the test reliability, item difficulty, and item discrimination, but not test score. The highest reliability of procrastination scale was found in the hard-to-easy version. For dissertation barriers scale, the highest reliabilities were found in both hard-to-easy and medium-then-random versions. The medium-then-random version was also detected with highest reliability on responsibility scale. But, the lowest scale reliabilities were all discovered in surveys with items ordered from easy to hard. Some items were found differing statistically significantly in logit position (difficulty) and discrimination in procrastination and dissertation barriers scales for all version comparisons. However, no significantly different test score was detected for scales between all survey versions which indicated that only test score was not influenced by item presentation orders in the conventional survey format. Similar results were detected in CAT format surveys, but mean person parameter (position) differed.

For the CAT format surveys, the influences of altering initial item difficulty levels were explored. Statistically significant different test lengths (numbers of item to be administered), reaction time, and mean person parameters were discovered in comparing versions starting at medium or extreme trait levels. Longer test lengths and reaction times were required, but lower mean person parameters were detected in the survey version with items beginning at an extreme trait level. However, similar to the result for the conventional survey formats, no significantly different test score was found between two test versions.

In this research, even though the test score of both survey formats was not significantly impacted by item order, results still indicated that survey reliability, survey performance (mean person parameters), item difficulty, and item discrimination on attitude measure are influenced by the item presentation order. Therefore, if one is concerned with test takers' attitude or evaluation of each item, then attitude measures should all have the same order of items in order to ensure that item difficulty and discrimination are invariance for every participant. However, if the purpose of the survey is to understand the overall attitude toward the specific topics, the total score is the only index for interpreting the survey results, the issue of item order effects may not need to be taken into consideration. In this case, when doing a survey research, the usage of interpretation index, such as overall survey score or each item score, is the most essential issue should be considered first.

As discussed previously in this chapter, there were limitations in this dissertation. Although results provided significant evidence to propose that item order effects did exist in attitude measures, several improvements in research methods can be made in the future studies. Directions for future research include: (1) using focus groups to compare performance in surveys with different item arrangements; (2) discovering phenomenon about violation of local independence on other domains of assessments; (3) applying rating scales without an ambiguous middle option in both conventional and computerized adaptive testing for generating more variance; and (4) investigating the impacts of possible moderators on the described effects in attitude measures. Through such research, it may help researchers and test developers have a better understanding about the item order effects and should ultimately finding alternative methods to deal with them. The essential issue for future investigation is to probe a way to maximize the efficiency of survey measures and this dissertation provides a starting point.

References

Anderson, N. H., & Jacobson, A. (1965). Effect of stimulus inconsistency and discounting instructions in personality impression formation. *Journal of Social Psychology, 2*, 531-539.

Ayida, S. A., & McClendon, M. J. (1990). Response effects in mail surveys. *Public Opinion Quarterly, 54*, 229-247.

Barcikowski, R. S., & Olsen, H. (1975). Test item arrangement and adaptation level. *Journal of Psychology, 90,* 87-93.

Belov, D. I., & Armstrong, R. D. (2009). Direct and inverse problems of item pool design for computerized adaptive testing. *Educational and Psychological Measurement, 69,* 533-547.

Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary children. *Journal of Educational Measurement, 22*, 231-240.

Bergstrom, B., A., Lunz, M.,E., & Gershon, R., C. (1992). Altering the level of difficulty in computer adaptive testing. *Applied Measurement in Education, 5,* 137-149.

Bishara, A. J. (2005). Control and accessibility bias in single and multiple anchoring effects. Unpublished doctoral dissertation, Washington University, Missouri.

Bishop, G., Hippler, H. J., Schwarz, N., & Strack, F. (1988). A comparison of response effects in self-administered and telephone surveys. In R. M. Groves, P. P. Biemer, L. E. Lysberg, J. T. Massey, W. L. Nicholls, II, & J. Wakesberg (Eds.), *Telephone survey methodology*. New York, NY: Wiley.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Association.

Bossart, P., & Di Vesta, F. J. (1966). Effects of context, frequency, and order of presentation of evaluative assetions on impression formation. *Journal of Personality and Social Psychology, 4*, 538-544.

Bowen, W. G., & Rudenstine, N. L. (1992). *In pursuit of the PhD. Princeton*. NJ: Princeton University Press.

Bowling, N. A., Boss, J., Hammond, G. D., & Dorsey, B. (2009). Susceptibility of job attitudes to context effects. *Journal of Career Assessment, 17*, 298-311.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational environment. In R. L., Linn (Ed.), *Educational measurement* (3ed ed.) (pp.364-407). New York: Macmillan.

Burka, J. B., & Yuen, L. M. (1983). *Procrastination: Why you do it, what to do about it*. Reading, MA: Addison-Wesley.

Butcher, J. N. (1987). *Computerized psychological assessment: A practitioner's guide*. New York, NY: Basic Books.

Brenner, M., H. (1964). Test difficulty, reliability, and discrimination as functions of item difficulty order. *Journal of Applied Psychology, 48*, 98-100.

Brickman, P., Rabinowitz, V. C., Karuza, J., Coates, D., Cohn, E., & Kidder, L. (1982). Models of helping and coping. *American Psychologist, 37*, 368-384.

Cacioppo, J. T., Gardner, W. L., & Berntson, G. B. (1997). Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. *Personality and Social Psychology Review, 1*, 3-25.

Chaiken, S. (1987). The heuristic model of persuasion. In M. P., Zanna, J. M. Olson, & C. P. Herman (Eds.), *Social influence: The Ontario Symposium* (Vol. 5, pp. 3-39).

Hillsdale, NJ: Erlbaum.

Converse, P. E. (1970). Attitudes and non-attitudes: Continuation of a dialogue. In Tufte, E. R. (Ed.), *The quantitative analysis of social problems* (pp. 168-189). Reading, MA: Addison-Wesley.

Crano, W. D. (1977). Primacy verusus recency in retention of information and opinion change. *Journal of Social Psychology, 101*, 87-96.

Cohen, A., R. (1964). *Attitude change and social influence.* New York, NY: London.

Cohen, J. B., & Reed ll, A. (2006). A multiple pathway anchoring and adjustment (MPAA) model of attitude generation and recruitment. *Journal of Consumer Research, 33*, 1-15.

Davis, H. L. (1986). Anchoring and adjusting model of spousal prediction. *Journal of Consumer Research, 13*, 25-27.

Deemer, S. A., & Minke, K. M. (1999). An investigation of the factor structure of the Teacher Efficacy Scale. *Journal of Educational Research, 93*, 3-10.

Dillman, D. A. (2000). *Mail and internet surveys: The tailored design method*. New York: John Wiley & Sons, Inc.

Dodd, B. G. (1990). The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement, 14,* 355-366.

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitude*. Fort Worth, TX: Harcourt Brace Jovanovich College.

Ellis, A., & Knaus, W. J. (1977). *Overcoming procrastination*. New York: New American Library.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustment are insufficient. *Psychological Science*, *17*, 311-318.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50,* 229-238.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In Zanna, M. P. (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 75-109). San Diego, CA: Academic Press.

Flaugher, R., L., Melton, R., S., & Myers, C., T. (1968). Item rearrangement under typical test conditions. *Educational and Psychological Measurement, 28*, 813-824.

Frantom, C., Green, K. E., & Lam, T. C.M. (2002). Item grouping effects on invariance of attitude items. *Journal of Applied Measurement, 3,* 38-49.

Germeroth, D. (1991). Lonely days and lonely nights: Completing the doctoral dissertation. *ACA Bulletin, 76,* 60-89.

Gershon, R. C. (1989, March). *Test anxiety and item order: New parameters for item response theory*. Paper presented at the meeting of the American Educational Research Association.

Girves, J. E., & Wemmerus, V. (1988). Developing models of graduate student degree progress. *Journal of Higher Education*, *59*, 163-189.

Green, B. F. (1982). Adaptive testing by computer. In R. B. Ekstrom (Eds.), *Measurement, technology, and individuality in education*. San Francisco,CA: Jossey-Bass.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational*

Measurement, 21, 347-360.

Green, K. E., & Kluever, R. C. (1997, March). *The dissertation barriers scale*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Green, K. E. (1997). Psychological factors affecting dissertation completion. In Goodchild, L. F., Green, K. E., Katz, E. L., and Kluever, R. C., *Rethinking the Dissertation Process*, pp. 57-64. San Francisco: Jossey-Bass.

Green, P., Tull, D., & Albaum, G. (1988). *Researcher for marketing decisions* (5th ed.). Englewood Cliffs, NJ: Prentice Hall.

Georgiadou, E., Triantafillou, E., & Economides, A. A. (2006). Evaluation parameters for computer-adaptive testing. *British Journal of Educational Technology, 37*, 261-278.

Gregoire, M. (2003). Is it a challenge or a threat? A dual-process model of teachers' cognition and appraisal processes during conceptual change. *Educational Psychology Review, 15*, 147-179.

Halloran, J. D. (1970). *Attitude formation and change*. New York, NY: Leocester University Press.

Hambleton, R. K., & Traub, R. E. (1974). The effects of item order on test performance and stress. *Journal of Experimental Education, 43*, 40-46.

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

Hambleton, R. K., & Traub, R. E. (1974). The effects of item order on test performance and stress. *Journal of Experimental Education, 43*, 40-46.

Hansen, J. C., Neuman, J. L., Haverkamp, B. E., & Lubinski, B. R. (1997).

Comparison of user reaction to two methods of Strong Interest Inventory administration and report feedback. *Measurement and Evaluation in Counseling and Development, 30*, 115-127.

Harrison, D. A., & McLaughlin, M. E. (1993). Cognitive processes inself-report responses: Tests of item context effects in work attitude measures. *Journal of Applied Psychology, 78,* 129-140.

Hastie, R., & Dawes, R., M. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making*. Thousand Oaks, CA:Sage.

Hogarth, R., M., & Einhorn, H., J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24,* 1-55.

Hovland, C. I. (1959). Reconciling conflicting results derived from experimental and survey studies of attitude change. *American Psychologist, 14*, 8-17.

Jacks, P., Chubin, D. E., Porter, A. L., & Connolly, T. (1983). The ABCs of ABDs: A study of incomplete doctorates. *Improving College and University Teaching, 31*, 74-81.

Johnson, E. M. (1998). *A taxonomy of person misfit on affective measures*. Unpublished doctoral dissertation, University of Denver, Colorado.

Johnson, E. M., Green, K. E., & Kluever, R. C. (2000). Psychometric characteristics of the revised procrastination inventory. *Research in Higher Education, 41*, 269-279.

Kleinke, D. J. (1980). Item order, response location and examinee sex and handedness and performance on a multiple-choice test. *Journal of Educational Research, 73*, 225-228.

Klimko, I. P. (1984). Item arrangement, cognitive entry characteristics, sex, and test anxiety as predictors of achievement examination performance. *Journal of Experimental Education, 52*, 214-219.

Klosner, N. C., & Gellman, E. K. (1973). The effect of item arrangement on classroom test performance: Implications for content validity. *Educational and Psychological Measurement, 33*, 413-418.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55*, 312-320.

Kruglanski, A. W., & Freund, T. (1983). The freezing and unfreezing of lay-inferences: Effects on impressional primacy, ethnic stereotyping, and numerical anchoring. *Journal of Experimental Social Psychology, 19*, 448-468.

Laffitte Jr., R. G. (1984). Effects on item order on achievement test scores and students' perception of test difficulty. *Teaching of Psychology, 11*, 212-214.

Lavine, H., Huff, J. W., Wagner, S. H., & Sweeney, D. (1998). The moderating influence of attitude strength on susceptibility to context effects in attitude surveys. *Journal of Personality and Social Psychology, 75*, 359-373.

Leary, L. F., & Dorans, N., J. (1985). *Implications for altering the context in which test items appear: A historical perspective on an immediate concern*. *Review of Educational Research, 55*, 387-413.

Lilley, M., Barker, T., & Britton, C. (2004). The development and evaluation of a software prototype for computer-adaptive testing. *Computers and Education, 43*, 109-123.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lynch Jr., J. G. (2006). Accessibility-diagnosticity and the multiple pathway anchoring and adjustment model. *Journal of Consumer Research, 33*, 25-27.

Marso, R. N. (1970). Test arrangement, testing time, and performance. *Journal of*

Educational Measurement, 7, 113-118.

Mason, R., Carlson, J. E., & Tourangeau, R. (1994). Contrast effects and subtraction in part-whole questions. *Public Opinion Quarterly, 58*, 569-578.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J., Weiss (Eds.), *New horizons in testing: Latent trait test theory and computerized adaptive tetsing*. New York, NY: Academic Press.

McBride, J. R. (1997). Research antecedents of applied adaptive testing. In W. A., Sands, B. K., Waters, & McBride, J. R. (Eds.), *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Mills, C. N. (1999). Development and introduction of a computer adaptive Graduate Record Examinations General Test. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment*. Mahwah, NJ: Erlbaum.

Monk, J. J., & Stallings, W. M. (1970). Effects of item order on test scores. *Journal of Educational Research, 63*, 463-465.

Moses, T., Yang, W., & Wilson, C. (2007). Using Kernel Equating to assess item order effects on test scores. *Journal of Educational Measurement, 44*, 157-178.

Munz, D. C., & Smouse, A. D. (1968). Interaction effects of item-difficulty sequence and achievement-anxiety reaction on academic performance. *Journal of Educational Psychology, 59*, 370-374.

Muraki, E., & Bock, D. (2003). PARSCALE 4. Lincolnwood, IL: Scientific Software International.

Muszynski, S. Y., & Akamatsu, T. J. (1991). Delay in completion of doctoral dissertations in clinical psychology. *Professional Psychology: Research and Practice, 22,* 119-123.

Neely, D. L., Springston, F. J., & McCann, Stewart J. H. (1994). Does item order affect performance on multiple-choice exams? *Teaching of Psychology, 21,* 44-45.

Newman, D. L., Kundert, D. K.,Lane, D. S., & Bull, K. S. (1988). Effect of varying item order on multiple-choice test scores: Importance of statistical and cognitive difficulty. *Applied Measurement in Education, 1,* 89-97.

Nosek, B. A., & Banaji, M. R. (2001). The Go/No-go association task. *Social Cognition, 19(6),* 625-666.

Ofir, C., & Simonson, I. (2007). The effect of stating expectations on customer satisfaction and shopping experience. *Journal of Marketing Research, 66*, 164-174.

Ortner, T. M. (2004). On changing the position of items in personality questionnaires: analyzing effects of item sequence using IRT. *Psychology Science, 46,* 466-476.

Ortner, T. M. (2008). Effects of changed item order: A cautionary note to practitioners on jumping to computerized adaptive testing for personality assessment. *International Journal of Selection and Assessment, 16,* 249-257.

Park, H. S., Levine, T. R., Westerman, C. K., Orfgen, T., & Foregger, S. (2007). The effects of argument quality and involvement type on attitude formation and attitude change: A test of dual-process and social judgment predictions. *Human Communication Research, 33,* 81-102.

Payne, J. W., Bettman, J. R., & Joknson, E. J. (1990). The adaptive decision-maker: Effort and accuracy in choice. In R. M. Hogarth (Ed.), *Insights in decision making* (pp. 129-153). Chicago. IL: Unic. Of Chicago Press.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and* *peripheral routes to attitude change.* New York, NY: Spring-Verlag.

Plake, B., S. (2002). Item arrangement and knowledge of arrangement on test scores. *Journal of Experimantal Education*, *49*, 56-58.

Plake, B. S., Ansorge, C. J., Parker, C. S., & Lowry, S. R. (1982). Effects of item arrangement, knowledge of arrangement test anxiety and sex on test performance. *Journal of Educational Measurement, 19*, 49-57.

Plake, B. S., Melican, G. J., Carter, L., & Shaughnessy, M. (1983). Differential performance of males and females on easy to hard item arrangements: Influence of feedback at the item level. *Educational and Psychological Measurement*, *43*, 1067-1075.

Plake, B. S., Patience, W. M., & Whitney, D. R. (1988). Differential item performance in mathematics achievement test items: Effect of item arrangement. *Educational and Psychological Measurement, 48*, 885-894.

Pomplun, M., & Ritchie, T. (2004). An investigation of context effects for item randomization within testlets. *Journal of Educational Computing Research, 30*, 243-254.

Russell, E. W. (2000). The application of computerized scoring programs to neuropsychological assessment. In R. D. Vanderploeg (Ed.), *Clinician's guide to neuropsychological assessment*. Mahwah, NJ: Erlbaum.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph, No. 17*.

Samejima, F. (2008). Graded response model based on the logistic positive exponent family of models for dichotomous responses. *Psychometrika Monograph, 73*, 561-578.

Sangster, R. L. (1993). *Question order effects: Are they really less prevalent in mail surveys?* Unpublished doctoral dissertation, Department of Sociology, Washington State

University, Pullman, WA.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and content*. New York, NY: Academic Press.

Schurr, K. T., & Henriksen, L. W. (1980, April). *Can questionnaire item ordering affect the inter-item relationship? An empirical example*. Paper presented at the meeting of the National Council on Measurement in Education, Boston, MA.

Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation*. Mahwah, NJ: Erlbaum.

Schwarz, N., & Hippler, H. (1995). Subsequent questions may influence answers to preceding questions in mail surveys. *Public Opinion Quarterly, 59,* 93-97.

Schwarz, N., Strack, F., & Mai, N. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly, 55,* 3-23.

Schwarz, N., & Wyer, R., S. (1985). Effects of rank ordering stimuli on magnitude rating of these and other stimuli. *Journal of Experimental Psychology, 21,* 30-46.

Semb, G., Glick, D. M., & Spencer, R. E. (1979). Students withdrawals and delayed work patterns in self-paced psychology courses. *Teaching of Psychology, 6,* 23-25.

Sheehan, K., & Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets. *Applied Psychological Measurement, 16*, 65-76.

Siminski, P. (2008). Order effects in batteries of questions. Quality and Quantity, 42, 477-490.

Simms, L. J., & Clark, L. A. (2005). Validation of a computerized adaptive version of the schedule for nonadaptive and adaptive personality (SNAP). *Psychological Assessment,* *17*, 28-43.

Singleton, R. A., & Straits, B. C. (2005). *Approaches to social research* (4th ed.). New York: Oxford University Press.

Stern, F. P. (1985). The effects of separation-individuation conflicts on length of time to complete the dissertation. Unpublished doctoral dissertation, City University of New York. *Dissertation Abstracts International, 46*, 11B, 4030.

Stewart, R. H. (1965). Effect of continuous responding in the order effect in personality impression formation. *Journal of Personality and Social Psychology, 1*, 161-165.

Strack, F., Schwarz, N., & Gschneidinger, E. (1985). Happiness and reminiscing: The role of time perspective, affect, and mode of thinking. *Journal of Personality and Social Psychology, 47*, 1460-1469.

SurveyMonkey. (n.d.). Retrived Feburary 1, 2010, from http://www.surveymonkey .com

Tan, L., & Ward, G. (2007). Output order in immediate serial recall. Memory and Cognition, 35, 1093-1106.

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition* (2en ed.). Chicago: University of Chicago Press.

Tourangeau, R., & Rasinski, K., A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 3*, 299-314.

Tourangeau, R., Rasinski, K. A., Bradburn, N. & D'Andrade, R. (1989). Belief accessibility and context effects in attitude measurement. *Journal of Experimental Social Psychology, 25*, 401-421.

Tourangeau, R., Rasinski, K., A., & Bradburn, N. (1991). Measuring happiness in surveys: A test of the substraction hypothesis. *Public Opinion Quarterly, 55*, 255-266.

Tourangeau, R., Rasinski, K. A., Bradburn, N. & D'Andrade, R. (2001). Carryover effects in attitude surveys. *Public Opinion Quarterly, 53,* 495-524.

Tversky, A., & Kajneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124-1130.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14,* 181-196.

Vega, E. M., & O'Leary, K. D. (2006). Reaction time and item presentation factors in the self-report of partner aggression. *Violence and Victims, 21*, 519-532.

Wagner, D. V. (1986). Selected personality characteristics and situational factors as correlates of completion and non-completion of the doctoral dissertation. *Dissertation Abstractions International, 47*, 3377A.

Wainer, H. (2000). *Computerized adaptive testing: A primer*. Mahwah, NJ: Lawrence Erlbaum Associates.

Waller, N. G., & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the absorption scale. *Journal of Personality and Social Psychology, 57*, 1051-1058.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.

Weiss, D. J. (1983). *New horizons in testing: Latent trait test theory and computerized adaptive tetsing*. New York, NY: Academic Press.

Weiss, D. J., & McBride, J. R. (1984). Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement, 8*, 273-285.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53*, 774-789.

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

Weiss, L. (1987). The relationship between personality variables and completion of a doctoral dissertation. Unpublished doctoral dissertation, California School of Professional Psychology. *Dissertation Abstracts International, 48,* 3814-3815A.

Weiss, D. J., & Vale, D. (1987). Computerized adaptive testing for measuring abilities and other psychological variables. In J. N., Butcher (Eds.), *Computerized psychological assessment: A practitioner's guide*. (pp. 325-343). New York, NY: Basic Books.

Weiss, D. J. (2005). *Manual for POSTSIM: Post-hoc simulation of computerized adaptive testing. Version 2.0.* St. Paul MN: Assessment Systems Corporation.

Weiss, D. J. (2008). *Manual for the FastTEST Professional Testing System, Version 2.* St. Paul MN: Assessment Systems Corporation.

Whitely, S., & Dawis, R. (1976). The influence of test context on item difficulty. *Educational and Psychological Measurement, 36*, 329-337.

Willits, F. K., & Saltiel, J. (1995). Question order effects on subjective measures of quality of life. *Rural Sociology, 60*, 654-665.

Wilson, T. D., & Hodges, S. D. (1992). Attitudes as temporary constructions. In. Tesser, A., & Martin, L. (Eds.), *The construction of social judgment* (pp. 37-65). Hillsdale, NJ: Erlbaum.

Wilson, T., D., Lindsey, S., & Schooler, T., Y. (2000). A model of dual attitudes.

Psychological Review, 107, 101-126.

Wittenbrink, B., & Schwarz, N. (2007). *Implicit measures of attitudes*. New York, NY: Guilford Press.

Wright, B. D. (1977). Misunderstanding the rasch model. *Journal of Educational Measurement, 14*, 219-225.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA.

Zhao, Q., & Linderholm, T. (2008). Adult metacomprehension: Judgment processes and accuracy constraints. *Educational Psychological Review, 20*, 191-206.

Zucherman, M., Koestner, R., Colella, M. J, & Alton, A. O. (1984). Anchoring in the detection of deception and leakage. *Journal of Personality and Social Psychology, 47*, 301-311.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice, 10*, 10-16.

Appendix A

DISSERTATION (THESIS) COMPLETION SURVEY

generated by Green and Kluever (1997)

Please circle the appropriate response to each of the following questions. Remember, your individual responses will not be seen by any faculty member; responses will be aggregated before examination.

1. Gender: Female Male
2. Degree: Master's Student Master Graduate Doctoral Student Doctoral Graduate

3. Age: 20-30 31-40 41-50 51-60 over 60
4. Programs: Education Business Social Science Natural Science
   Engineering/Computer Science Others_________

Prior to your dissertation, had you:

5.	Performed any data analyses for research projects?	Yes	No
6.	Conducted any research projects?	Yes	No
7.	Presented any research results, e.g., at a conference?	Yes	No
8.	Published any research papers?	Yes	No

For the following items think back to the time when you were working on your dissertation (thesis). Answer each of the following items on a yes-no scale according to your thoughts, feelings, or behavior at the time of working on your dissertation (thesis).

1.	I enjoy practical/clinical work more than I enjoy research.	Yes	No
2.	I would have finished my dissertation/thesis quicker if I had more incentives to work for (e.g., going on to a job/internship).	Yes	No

3. I couldn’t bear working on my dissertation Yes No

4. I wrote an acceptable but mediocre Yes No
dissertation so as to finish quickly.

5. I was afraid that I wouldn’t be able to Yes No
reach the academic goals I set for myself

6. I felt rotten about avoiding doing my Yes No
Dissertation/thesis.

7. I disliked the fact that after completing Yes No
coursework, I was entirely responsible
for planning and structuring my time.

8. I worked on a dissertation so long that Yes No
I lost all desire to do it.

9. I felt that writing a dissertation/thesis was Yes No
a waste of time, and I didn’t feel like doing it.

10. The thought of my advisor (or others) Yes No
finding out that I was not as bright as
s/he thought was unsettling to me.

11. I wish the college had set up small goals Yes No
For me and rewarded my progress on my
Dissertation/thesis

12. The dissertation/thesis was so difficult Yes No
I often felt why bother.

13. Any delay on my dissertation/thesis made me question my ability to handle such a project. Yes No

14. The thought of working on a major project that would take a long time to complete was overwhelming to me. Yes No

15. Choosing a dissertation/thesis topic was difficult since there were so many different things in which I was interested. Yes No

16. I felt that the College shouldn’t require students to do a dissertation. Yes No

17. I would have been better suited to a more structured program than a Ph.D. Yes No

18. I found that the obstacles I encountered in doing my dissertation/thesis resulted in my avoiding the task for a while. Yes No

19. In doing my dissertation/thesis, when I found I must do things which I did not enjoy, I started to view the entire dissertation/thesis as not enjoyable. Yes No

20. When a problem came up with my Dissertation/thesis, I tended to get anxious and worried about whether I would be able to handle it. Yes No

21. If I had been required to complete my dissertation in a reasonable, specified amount of time, I could have done it quicker. Yes No

22. I found that I could not devote enough time to my dissertation because there were so many more interesting things I would rather be doing. Yes No

23. I was too exhausted with all the other things in my life to finish a dissertation quickly. Yes No

24. The College provided little support for students once coursework was finished. Yes No

25. How strongly do you remember these components listed above when you were going through the dissertation (thesis) processes (please rank 1-6):

Weak 1 2 3 4 5 6 Strong

Were each of the following *concerns* to you or *difficulties* you encountered in completing your dissertation (thesis)? Answer each of the following items on a yes-no scale.

1. my own perfectionism Hindrance Help

2. my lack of interest in Dissertation/thesis topic Hindrance Help

3. narrowing the dissertation/ thesis topic Hindrance Help

4.	lack of structure of Dissertation/thesis process	Hindrance	Help
5.	difficulty with time Management	Hindrance	Help
6.	inadequate prior exposure to research	Hindrance	Help
7.	inadequate prior exp. with data analysis	Hindrance	Help
8.	doing the literature review	Hindrance	Help
9.	collecting the data	Hindrance	Help
10.	typing/word processing	Hindrance	Help
11.	job related pressures/demands	Hindrance	Help
12.	setting aside time for the dissertation (thesis)	Hindrance	Help
13.	setting aside a space/room for dissertation/thesis	Hindrance	Help
14.	conflict with role as home/family head	Hindrance	Help
15.	inability to plan ahead	Hindrance	Help
16.	self direction	Hindrance	Help

17. support of family, friends Hindrance Help

18. organizational skills Hindrance Help

19. time pressures Hindrance Help

20. love of the dissertation (thesis) topic Hindrance Help

21. persistence Hindrance Help

22. sticking to a schedule Hindrance Help

23. How strongly do you remember these concerns or difficulties listed above when you were going through the dissertation (thesis) processes (please rank 1-6):

Weak 1 2 3 4 5 6 Strong

Completion of the dissertation (thesis) involves the cooperation and effort of a number of people and resources. Some people have major *responsibility* for certain components of this process and others have less responsibility for it. Below are some of the major components that relate to completion of a dissertation. The term **"University"** is intended to include **all resources of the University including advisor(s), faculty, courses, seminars, independent study, library, computing services, and administrative functions**. The term "Student" relates to yourself.

Please go through the scale for each item, select the answer which represents your impression of the current state of where responsibility rests with.

1. Responsibility for progressing through the dissertation/thesis rests with: Student Advisor/University

2. Responsibility for locating and acquiring relevant research materials relating to the dissertation/thesis topic rests with: Student Advisor/University

3. Responsibility for selecting a Dissertation/thesis topic rests with: Student Advisor/University

4. Responsibility for preparing a human subjects application rests with: Student Advisor/University

5. Responsibility for locating subjects (or sources) to provide data for the study rests with: Student Advisor/University

6. Responsibility for collecting the Dissertation/thesis data rests with: Student Advisor/University

7. Responsibility for analyzing the Dissertation/thesis data rests with: Student Advisor/University

8. Responsibility for interpreting the data rests with: Student Advisor/University

9. Responsibility for developing research tool skills (computer, Library, etc.) rests with: Student Advisor/University

10. How strongly do you remember these components listed above when you were going through the dissertation/thesis processes (please rank 1-6):

Weak 1 2 3 4 5 6 Strong

Do you think your responses were affected by the order of the items?

□ No. □ Yes.

Please indicate the strength of your overall attitudes toward the issues listed below in regard to the dissertation/thesis processes: (please rank 1-6)

1. Procrastination/Time Management:

Weak 1 2 3 4 5 6 Strong

2. Dissertation Barriers/Difficulties:

Weak 1 2 3 4 5 6 Strong

3. Responsibility:

Weak 1 2 3 4 5 6 Strong

Would you willing to participate the interviews to talk about your opinions of doing a dissertation/thesis in the future?

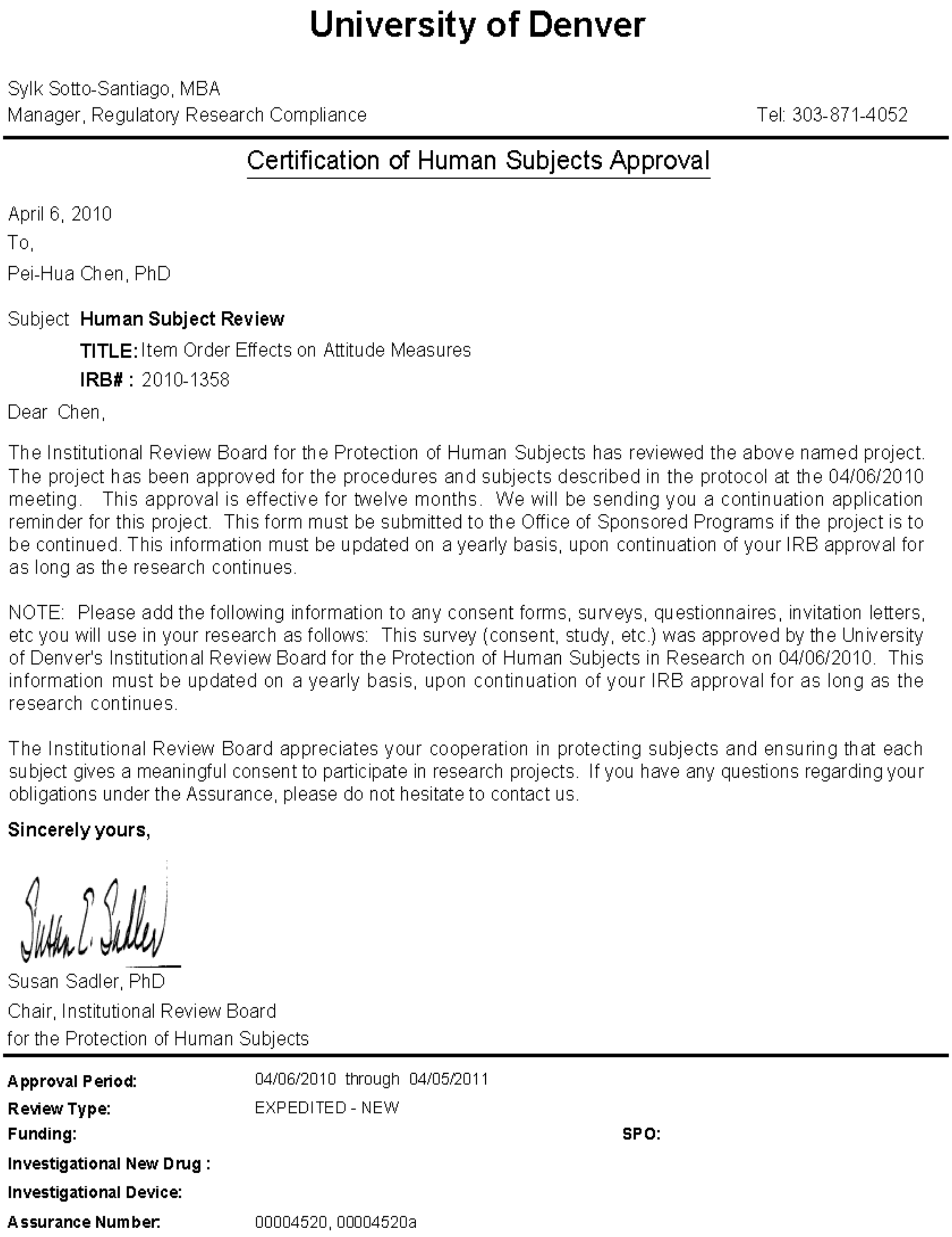
□ No.

□ Yes. My e-mail address is ___________________________________

THANK YOU FOR YOUR PARTICIPATION!

If you are interested in this topic, the relevant articles can be found on the website. The results of the research will also be posted in November (http://portfolio.du.edu/pchen30).

Appendix B



University of Denver

Sylk Sotto-Santiago, MBA
Manager, Regulatory Research Compliance Tel: 303-871-4052

Certification of Human Subjects Approval

April 6, 2010
To,
Pei-Hua Chen, PhD

Subject **Human Subject Review**
TITLE: Item Order Effects on Attitude Measures
IRB# : 2010-1358

Dear Chen,

The Institutional Review Board for the Protection of Human Subjects has reviewed the above named project. The project has been approved for the procedures and subjects described in the protocol at the 04/06/2010 meeting. This approval is effective for twelve months. We will be sending you a continuation application reminder for this project. This form must be submitted to the Office of Sponsored Programs if the project is to be continued. This information must be updated on a yearly basis, upon continuation of your IRB approval for as long as the research continues.

NOTE: Please add the following information to any consent forms, surveys, questionnaires, invitation letters, etc you will use in your research as follows: This survey (consent, study, etc.) was approved by the University of Denver's Institutional Review Board for the Protection of Human Subjects in Research on 04/06/2010. This information must be updated on a yearly basis, upon continuation of your IRB approval for as long as the research continues.

The Institutional Review Board appreciates your cooperation in protecting subjects and ensuring that each subject gives a meaningful consent to participate in research projects. If you have any questions regarding your obligations under the Assurance, please do not hesitate to contact us.

Sincerely yours,

Susan E. Sadler

Susan Sadler, PhD
Chair, Institutional Review Board
for the Protection of Human Subjects

Approval Period:	04/06/2010 through 04/05/2011	
Review Type:	EXPEDITED - NEW	
Funding:		**SPO:**
Investigational New Drug :		
Investigational Device:		
Assurance Number:	00004520, 00004520a	