A Monte Carlo Study of the Optimal Rank Order Relationship with Criterion Scores

Hongyan Cui

Western Michigan University, hongyan.cui@gmail.com

Follow this and additional works at: http://scholarworks.wmich.edu/dissertations

Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation
http://scholarworks.wmich.edu/dissertations/371

This Dissertation-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Dissertations by an authorized administrator of ScholarWorks at WMU. For more information, please contact maira.bundza@wmich.edu.
A MONTE CARLO STUDY OF OPTIMAL RANK ORDER RELATIONSHIPS WITH CRITERION SCORES

by

Hongyan Cui

A dissertation submitted to the Graduate College in partial fulfillment of the requirements for the Degree of Doctor of Philosophy Educational Leadership, Research and Technology Western Michigan University December 2014

Doctoral Committee:
Brooks Applegate, Ph.D., Chair
David Hartmann, Ph.D.
Warren Lacefield, Ph.D.
A MONTE CARLO STUDY OF THE OPTIMAL RANK ORDER RELATIONSHIP WITH CRITERION SCORES

Hongyan Cui, Ph.D
Western Michigan University, 2014

Appropriate scoring methods for tests should be based on theories of the construct domains of such tests (Messick, 1989). This is called structural fidelity (Loevinger, 1957), a necessary but not sufficient condition for construct validity (Keith & Kranzler, 1999). A situation reaction test (SRT) consists of items with alternatives to be ranked according to subjects’ best judgment. Traditional rank scoring methods assume unidimensionality, implying one scoring key for each item – a single external criterion. This is not appropriate when items elicit multidimensional responses determined by subjects’ best judgments based on possibly multiple internal criteria. The purpose of this study is to determine when a complex ranking item is theoretically governed by multiple traits, whether multiple trait-keys can be identified and validated such that multiple item scores, one for each trait, can be derived from each item.

SRTs with 4 alternatives per item were examined using "optimal ranking order relationship with criterion scores" (ORORCS) in a Monte Carlo simulation under various test conditions: correlation among traits, sample size for calibration and validation, and number of items. Dependent variables were $c_k s_k$ (Fisher’s z between SRT and criterion $k$ scores) and corresponding CI widths obtained using a validation sample. MANOVA results ($N = 1000/cell$) demonstrated SRTs can be scored in a multidimensional manner;
that multiple traits can be measured simultaneously using multiple trait keys to score ranking items.

The ORORCS’s ability to resolve multiple traits in SRTs decreases as inter-trait correlations increase \((p < .0001)\). But for fixed correlation, SRT validity improves and sharpens (tighter CIs) with test length \((p < .0001)\) and with sample size \((p < .0001)\). Interactions were also significant. The results indicate the SRT ORORCS scoring method has structural fidelity. It can effectively measure subject states with respect to several primary causal factors, even when these factors are somewhat inter-correlated. Sample size, though significant, was not much of a factor above 25. Test length is important, but good results do not require long tests. Good design is more important. This study demonstrated an alternative scoring model where items may be scored with multiple keys corresponding to different traits identified using ORORCS procedure.
ACKNOWLEDGMENTS

“And we know that all things work together for good to those who love God, to those who are called according to His purpose. Because those whom He foreknew, He also predestinated to be conformed to the image of His Son, that He might be the Firstborn among many brothers; And those whom He predestinated, these He also called; and those whom He called, these He also justified; and those whom He justified, these He also glorified.”

– Romans 8:28-30

It will be very interesting if our life could be zoomed in and out like the computer screen. We would be able to see clearer how we should be grateful for everything, every situation and everyone in our life. There are so many people that have helped me to reach the peak of school education, a doctoral degree – all the teachers, staff and classmates from my elementary, middle school, high school, college and graduate school; all the believers of Jesus Christ who prayed and supported me; all my families and friends who have always encouraged me and never, not even once discouraged me on accomplishing my study.

I would thank Western Michigan University for being such a friendly student-centered campus. I got timely help, support and sponsorship from all the directly relevant department including the Graduate College, the Registration office, the Office of International students services, the Evaluation Center, and of course, my home, the Department of Educational Leadership, Research, and Technology at the College of
Acknowledgments—Continued

Education and Human Development. I am very grateful that all the faculty and staff have worked together to create such a warm and inviting education environment.

I can never thank enough my committee, Dr. Brooks Applegate, Dr. Warren Lacefield, and Dr. David Hartmann whose trust in me has never dwindled and whose diligent and serious attitude toward research work has always propelled me to go a step further. As a golden team, they did not only inspire me, but also encouraged and challenged me so much that they made me see that a dissertation is a commencement rather than a consummation of an academic journey. I will never forget the many hours we worked together. This dissertation is dedicated to my two dear academic counselors, Dr. Brooks Applegate and Dr. Warren Lacefield who have been working very inspiringly and intensively with me through my whole graduate study.

I am deeply indebted to my parents and siblings who have always thought highly of education and supported my education since my early age. My deep gratitude also goes to Dr. Timothy Light, without whose kindness and insight my doctoral study would not have had a beginning; and to my husband and our son, whose love and encouragement get me through the whole process of study. My gratitude extends to the Aiqin Chen, Colbers, the Deyoungs, the Durwoods, Yehong Du, the Erhardts, Ivonne Gandia, Xianglan He, Sarah Ho, Yuangzhong Huang, Dajiang Jiang, Yingchun Liu, Xuejin Lu, Yingbo Ma, Ling Pan, Nan Song, Hui Sun, Nan Sun, Yumei Tan, the Toppens, Feilin Wang, Chunling Wang, Haiping Wu, Xinlan Wu, Xin Su, Jiangang Xia, Chunchun Ye, Chuanzhen Zhanghong, Manying Zhangmu, Yanli Zhao, Zhenji Zhou, Pengwei Zhu, and all other friends.
Acknowledgments—Continued

At last, I would dedicate my doctoral degree to my ultimate Counselor, my Lord and Savior Jesus Christ, Who has been guiding me through every step in my life and study with or without my awareness. It is in Him that I have finally found the purpose of life and living – to gain more of Him and for His glorious expression. I thank you, Lord Jesus, for being with me all the time in my life and being the map of my travelling my journey. I thank You for making me love You more, for making me more desirable of Your wisdom and light, for making me more humble every day. You have showed me how much I need other people to accomplish something that You favor.

Hongyan Cui
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS .................................................................................. ii

LIST OF TABLES ............................................................................................ viii

LIST OF FIGURES ........................................................................................ ix

CHAPTER

I.  INTRODUCTION ...................................................................................... 1

   Statement of the Problem ...................................................................... 1

   Purpose of the Study ........................................................................... 6

   Assumptions of the Study .................................................................. 8

   Research Questions ............................................................................. 8

   Hypotheses ......................................................................................... 9

II. LITERATURE REVIEW ............................................................................. 11

   Basic Concepts and Main Characteristics of Rankings ....................... 12

   Teaching Situation Reaction Test ...................................................... 14

   A Measurement Problem with the TSRT ........................................... 17

   External Criterion vs. Internal Criterion .............................................. 17

   Keyed vs. Non-Keyed Ranking Items ................................................. 21

   Dimensional Structure of Internal Criterion-based Rankings .......... 22

   A Possible Solution to the TSRT Scoring Problem ......................... 24

   Modeling and Analysis of Non-keyed Rankings ............................... 26

   Tests of Ranking Items ..................................................................... 30
<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criterion Validity</td>
<td>35</td>
</tr>
<tr>
<td>Test Dimensionality</td>
<td>37</td>
</tr>
<tr>
<td>Test Scoring</td>
<td>44</td>
</tr>
<tr>
<td>Unidimensional Unfolding Theory</td>
<td>48</td>
</tr>
<tr>
<td>Multidimensional Unfolding Theory</td>
<td>53</td>
</tr>
<tr>
<td>Theories on Preference Constructions</td>
<td>56</td>
</tr>
<tr>
<td>Cross-Validation</td>
<td>69</td>
</tr>
</tbody>
</table>

### III. METHODOLOGY

- Experimental Conditions | 71
- Finding Multiple Criterion Trait-Keys for an Item through the ORORCS | 73
- The Monte Carlo Simulation Design | 76
- Data Generation | 79
- Data Analysis | 81
- Summary of Research Methodology | 83

### IV. RESULTS

- 3-Trait Model | 86
- 2- and 4-Trait Models | 111
- Summary of Findings for Three Models: | 114

### V. DISCUSSION

- Significance of the Study: | 117
- Implications for SRT Test Construction | 120
- Limitations of the Study | 121
LIST OF TABLES

1. Ranking Strategies Based on Personal Preference ........................................34
2. Two Sets of I-scales for a Qualitative J-scale of ABCD ..........................49
3. Pros and Cons of Different Cross-Validation Methods ..........................70
4. ACBD Scored Against ABCD Using Kendall's Distance .......................75
5. Summary of Factorial Design of the Study (N = 1000) .........................79
6. A Model of Item Alternatives Measured on 3 Traits .........................81
7. MANOVA Results for the 3-Trait Model .............................................101
8. Univariate 3-Way Interaction Findings for 3-Trait Model, df = 24, 59940 ....102
9. NITEMS Quadratic Estimates (3-Trait Model) ....................................109
10. Quadratic Contrast for Largest SSIZE Parameter Difference (3-Trait Model) ..........................................................109
11. TCORR Contrasts in NITEMS Quadratic Curve Fitting for Mean r_{ZCkS_k} Pooling Over SSIZE (3-Trait Model) .........................................................110
12. MANOVA Results for 2-Trait Model ....................................................111
13. MANOVA Results for 4-Trait Model ....................................................111
14. Univariate 2-Way Interactions for 2-Trait Model ................................112
15. Univariate 3-Way Interactions for the 4-Trait Model, df = 24, 59939 ....113
16. Correlations among the SRT Validity Coefficients in the 2-, 3-, and 4-Trait Models .................................................................121
LIST OF FIGURES

1. Comparison of Traditional Scoring Method with the Trait-key Scoring Method.....7
2. Personal Compensatory Model.............................................................................64
3. City Block Model. .............................................................................................65
4. Identify Trait-keys through ORORCS ..............................................................76
5. Factors and Sample Sizes in the Study (3-Trait Model, N = 1000/cell). .............86
6. CI Plots for SRT Validity Coefficient Using the Best Key Found for C₁ in the Calibration Sample, as a Function of the Sample Size and Test Length with Inter-trait Correlation = .00 (1,000 Replications of a 3-Trait Model). ......................88
7. CI Plots for the SRT Validity Coefficient for Trait C₁ as a Function of the Sample Size and Test Length with Inter-trait Correlation = .40 (1,000 Replications of a 3-Trait Model). .......................................89
8. CI Plots for the SRT Validity Coefficient for Trait C₁ as a Function of the Sample Size and Test Length with Inter-trait Correlation = .80 (1,000 Replications of a 3-Trait Model). ..................................................90
9. Mean SRT CI Width as a Function of Inter-trait Correlation, Test Length and Sample Size .........................................................................................................................91
10. CI Plots for the SRT Validity Coefficient for Trait C₁ as a Function of the Inter-trait Correlation and the Test Length with Sample Size = 50 (1,000 Replications of a 3-Trait Model). ....................................................92
11. CI Plots for the SRT Validity Coefficient for Trait C₁ as a Function of the Inter-trait Correlation and the Test Length with Sample Size = 250 (1,000 Replications of a 3-Trait Model). ........................................................................93
12. CI Plots for the SRT Validity Coefficient for Trait C₁ as a Function of the Inter-trait Correlation and the Test Length with Sample Size = 1000 (1,000 Replications of a 3-Trait Model). .......................................................94
13. CI Plots for the SRT Validity Coefficient for Trait C₁ as a Function of the Inter-trait Correlation and the Test Length with Sample Size = 5000 (1,000 Replications of a 3-Trait Model) ....................................................95
List of Figures—Continued

14. CI Plots for the SRT Validity Coefficient for Trait C1 as a Function of the Inter-trait Correlation and the Sample Size with Test Length = 1 Item (1,000 Replications of a 3-Trait Model). ..........................................................96

15. CI Plots for the SRT Validity Coefficient for Trait C1 as a Function of the Inter-trait Correlation and the Sample Size with Test Length = 5 Item (1,000 Replications of a 3-Trait Model). ..........................................................97

16. CI Plots for the SRT Validity Coefficient for Trait C1 as a Function of the Inter-trait Correlation and the Sample Size with Test Length = 10 Item (1,000 Replications of a 3-Trait Model). ..........................................................98

17. CI Plots for the SRT Validity Coefficient for Trait C1 as a Function of the Inter-trait Correlation and the Sample Size with Test Length = 15 Item (1,000 Replications of a 3-Trait Model). ..........................................................99

18. CI Plots for the SRT Validity Coefficient for Trait C1 as a Function of the Inter-trait Correlation and the Sample Size with Test Length = 20 Item (1,000 Replications of a 3-Trait Model). .........................................................100

19. SRT Validity as a Function of 3-way Interaction between Trait Inter-correlations, Sample Size and Test Length (3-Trait Model, Trait C1)..............103

20. SRT Validity as a Function of Trait Inter-correlations and Test Length at Each Sample Size (3-Trait Model, Trait C1). .................................................................104

21. SRT Validity as a Function of 3-way Interaction between Trait Inter-correlations, Sample Size and Test Length (3-Trait Model, Trait C2).................................105

22. SRT Validity as a Function of Trait Inter-correlations and Test Length at Each Sample Size (3-Trait Model, Trait C2). .................................................................106

23. SRT Validity as a Function of 3-way Interaction between Trait Inter-correlations, Sample Size and Test Length (3-Trait Model, Trait C3).................................107

24. SRT Validity as a Function of Trait Inter-correlations and Test Length at Each Sample Size (3-Trait Model, Trait C3). .................................................................108

25. SRT Validity as a Function of 3-way Interaction between Trait Inter-correlations, Sample Size and Test Length (2-Trait Model, Trait C1).................................113

26. SRT Validity as a Function of 3-way Interaction between Trait Inter-correlations, Sample Size and Test Length (4-Trait Model, Trait C1).................................114
CHAPTER I

INTRODUCTION

Test scoring is a very important issue in educational test development. If the scoring method for a test systematically disadvantages certain groups of examinees over others, such as students of color or students who are native speakers of a language, then this test is considered biased and fairness of this test could be jeopardized (Great School Partnership, 2014; Thissen & Wainer, 2001). This dissertation investigates a test-scoring problem which happens when a unidimensional key is used to score multidimensional responses. That is, when multiple latent traits or attributes are involved in the responses to a test item, a key highlighting one trait or attribute would disadvantage examinees weaker on this trait but possibly stronger on the other traits tested by the item. The information collected from such items through such scoring methods is neither comprehensive nor complete.

This study proposed and tested a new scoring method for such multidimensional tests through Monte Carlo simulation in an exploratory way. It examines specifically the applicability of the new scoring method for complex ranking items in testing subjects’ cognitive abilities and latent traits.

Statement of the Problem

Structural Fidelity of Test Score

Constructs or latent variables are not directly observable, but are manifested by test scores (Crocker & Algina, 1986; McDonald, 1999). In order to create an item score, the scoring criteria and rubrics need to be developed based on the theory of construct
domain or dimensional structure inherent in the responses (Loevinger, 1957; Messick, 1989; Peak, 1953; Thissen & Wainer, 2001). Both the construction of assessment tasks and the development of scoring criteria and rubrics should be based on the theory of the construct domain. The internal structure of test score (i.e., interrelations among the scored aspects of task performance) should be consistent with the known internal structure of the construct domain (Messick, 1989). This relation of the test score structure to the domain structure is called *structural fidelity* (Loevinger, 1957).

*Structural fidelity* is a necessary but not sufficient condition for construct validity. An absence of *structural fidelity* indicates that the scores derived from the test do not correspond to the underlying theory upon which the test is based (Keith & Kranzler, 1999). When the response matrix of a test only involves one attribute, e.g., mathematics ability, it is appropriate to use unidimensional scoring methods. Likewise, when the test scores are multidimensional, e.g., both reading and mathematics, multidimensional or multi-attribute models should be used (Böckenholt, 1996; Loevinger, 1957). If multiple domains or traits are measured in a test, but the scoring method assumes unidimensionality, then it might be difficult to interpret the test scores.

**A Scoring Problem of Rank Item Tests**

A ranking item is a type of item that contains three or more stimuli for respondents to rank with respect to an *external* criterion (a named attribute) or to an *internal* criterion (personal preference and judgment) (McDonald, 1999). Ranking items have been used to measure respondents’ preference choices in many fields, including economics, sociology, biology, marketing, geology, and education (Marden, 1995). The ranking items in situation reaction tests (SRT), e.g., the *Teaching Situation Reaction Test*
(TSRT) (Hough & Duncan, 1965), usually feature contextual item stems and sets of alternative actions to be ranked by subjects according to personal preference and judgment (hence, they are internal criterion-based). Having greater richness and less tedium, SRT ranking items are more interesting for the respondent to read and they have the potential to provide multiple insights into how the respondent thinks since they force the respondent to think through the situation before they respond (O’Kuma, Maloney & Hieggelke, 1999).

Ranking items, external criterion-based or internal criterion-based, are often scored using the same method. The traditional scoring method for ranking items requires test writers to provide one key for each item and responses will be scored according to their agreement with the key, with full agreement receiving a full score, partial agreement receiving a partial score, and no agreement receiving zero. For example, for the TSRT, there is one key (e.g., “2341”) for each item. A response can get a score of 0~6 depending on the degree of agreement with the key.

The problem with the traditional “one key for each item” scoring method for ranking items is that it does not differentiate the dimensional structure of the responses from external criterion-based rankings and those from internal criterion-based rankings. However, these two kinds of rankings have different dimensional structures. On the one hand, the external criterion-based rankings are based on one explicitly specified attribute and can be treated as unidimensional (Roussos, Stout & Marden, 1998). The ranking of the item alternatives results through within-attribute comparisons, i.e., different alternatives are evaluated on the same attribute. On the other hand, the dimensional structure of internal criterion-based rankings is likely to be rather complex since there is
no common specified attribute for the respondents to react to (Gessaroli & De Champlain, 2005; Hattie, 1985; Nandakumar, 1991; Reckase, 1990; Tate, 2003). It is plausible that a second, third, or $k$th dimension may actually be dominating or significantly influencing a respondent’s ranking choices.

Ranking decisions usually require cross-attributes comparisons, i.e., the options are compared on different attributes which mostly results in multidimensional rankings (Böckenholt, 1996). Moreover, different subjects may react differently to different dimensions in the item space since people may have different response frames. Even if subjects have similar response frames, the configuration of the options within their frames might still be different depending on how a subject weighs the relevant attributes when he/she compares the options. For instance, a subject may not care about attribute $K$ and makes his/her preference and judgment exclusively on the basis of attribute $L$. Hence on his/her individual preference scale, the scale value of an option having more $L$ but less $K$ will be larger than the scale value of another option with little $L$ but a lot of $K$.

A principle characteristic of the traditional “one key for each item” scoring method is that it is largely based on the assumption of unidimensionality. When there are multiple attributes possible in the response matrix and a response is only scored against one key, then the produced scores would not fully manifest the multidimensionality of the response space. According to Torgerson (1952), a multidimensional model should be considered if the preference rankings space is not sufficiently unidimensional to be modeled by a single scale. When there is only one key which highlights only one attribute, the information on the other attributes embedded in an item will be discarded (or treated as noise) and a respondent is not given the credit he/she deserves if another key was used.
In other words, the subject is inappropriately characterized by a single score collapsed over all other considerations. In order to avoid discarding potentially useful information and provide more fair or balanced or nuanced scoring treatment to respondents, a scoring method should be flexible enough to accommodate a varying number of dimensional structures of the response matrix of rankings.

This study proposed to investigate this problem by allowing for multiple scoring keys to be applied against a multidimensional ranking task. That is, instead of an item-key, trait-keys are provided for each item. If we suspect a ranking item measures multiple traits, we will find multiple trait-keys for this item, each trait-key highlighting one trait. A multiple key scoring method is possible with ranking items because of their unique format – \( m \) alternatives in a ranking item can be permuted to produce \( m! \) different ranking patterns (each of which may be used as a key). Imbedded in the \( m! \) ranking patterns are the unique and differentiable trait-keys and this provides the foundation for this new scoring method. It should be noted that only half of the \( m! \) all possible scores measure distinct dimensions due to the mirror image effect. Thus, when a ranking item has multi-attribute dimensionality, there can be multiple plausible ranking patterns, depending on which attribute(s) is used by the respondent as the criterion. Suppose, options A, B, C, and D are known to be related to two attributes \( K \) and \( L \). It is possible that when options A, B, C, and D are ordered based on attribute \( K \), the correct answer is “2341”, but when they are ranked on attribute \( L \), the correct answer is then “1243”. Still, if the composite of both attributes are considered, the correct answer is again different, say “2143”.
Purpose of the Study

This dissertation study explored a more efficient method by investigating a scoring solution that takes into account the dimensionality and the individual differences in the subjects’ preference rankings. The purpose of this study is to demonstrate that given a set of multidimensional rankings, there exists a unique item key for each dimension. Furthermore, this key can be established empirically from test data and validated against a new sample. The new scoring method constructs a set of “best” trait-keys (one set for each item) to score the respondents’ rankings along several predetermined dimensions/traits. Each trait-key maximally spreads out the respondents, maximizing the score variance and the correlation, along its respective dimension/trait, based on one single set of ranking data. The result is a method of constructing and scoring ranking item tests in a multi-dimensional way such that by using different trait-keys for an item, multiple measures of respondent features/attributes can be extracted from one administration of the ranking item test (Figure 1).
Figure 1. Comparison of traditional scoring method with the trait-key scoring method.

It should be noted that this study did not try to find all of the traits involved in a set of preference rankings. Rather, the new scoring method aimed at obtaining the optimal criterion validity of a ranking test by finding the best trait-key from all possible keys to score the data along each prior specified criterion continuum. That is, this study only focused on the traits that had been concurrently measured by some other known and previously validated criterion instruments.

The trait-key for each item will be identified through a procedure that produces the optimal rank order relationship with criterion scores (ORORCS) using Pearson’s correlation. A prerequisite for the ORORCS procedure is some well-established criterion instruments that can be administered to obtain subjects’ scores on these known criterion traits. Subjects need to take both the ranking item test and the known criterion instruments so that the scores can be correlated with each other. By scoring each item with all possible keys and choosing the one that maximizes the correlations between the
ranking test scores and a particular criterion measure scores, the ORORCS method
should yield a criterion validity coefficient similar to or possibly higher than a “one key
for each item” scoring method. Moreover, the ORORCS should be able to collect
information regarding multiple attributes involved in the test by using multiple criterion
trait-keys.

**Assumptions of the Study**

The assumptions in this study include:

1) Each of the multiple items in a test measures to some extent the same set of traits
called a state space.

2) People differ in their expression among the given set of traits.

3) When asked to provide preference ranking decisions, people maximally apply their
trait level(s) to the task and hence their responses reveal their level(s) of trait
expression.

4) When there are multiple items, people bring the same set of traits to consecutive
items and, hence, are transitive with respect to test items and alternatives.

5) The scales underlying preference ranking decisions are quantitative and metric
multidimensional scaling is applicable.

6) People’s traits and item alternatives can be depicted in the same trait state space of
which their preference ranking decisions are a function and their item responses are in
proportion to their specific position on the attributes defining this space.

**Research Questions**

The general research question posed by this study is to determine when a complex
ranking task/item (i.e., multiple attributes are involved in both the options of the item and
the subject’s response) is theoretically governed by multiple traits, whether multiple trait-keys can be identified and validated such that multiple ranking item scores, one for each of the multiple traits, can be derived from each ranking item. Once multiple ranking item scores are derived, their correlation with the criterion trait score (provided by asking the subjects to take some established criterion instruments) can be evaluated. If the trait-key is applicable to the trait then the Pearson correlation ($r_{C_kS_k}$) between the ranking test response scores on trait $k$ ($S_k$) and the trait scores from the given criterion test on trait $k$ ($C_k$) should approach unity as the number of test items increase. And the criterion validity coefficient obtained in this manner is defined as the SRT criterion validity coefficient and is the primary dependent variable in this study. Data from 1,000 replications of the experiment, with random subjects and random tests, under various sets of test conditions was examined. Specific to this study, factors investigated to determine how they might affect the SRT criterion validity correlation between the ranking test scores ($S_k$) and the criterion trait ($C_k$) included the: number of criterion traits (NTRAITS), correlation among the criterion traits (TCORR), test length in terms of the number of items (NITEMS), and subject sample size (SSIZE).

**Hypotheses**

The research hypothesis of this study is that $S_{kp}$ (person $p$’s ranking test score on criterion $k$ from a multidimensional preference ranking test SRT), is an accurate estimate of $C_{kp}$ (person $p$’s criterion score estimated by criterion instrument $k$). The accuracy of the estimation can be evaluated by examining the SRT criterion validity coefficient, $r_{ZCS}$ (in terms of the Fisher’s $z$ transformation of Pearson correlation coefficient
between $C_k$ and $S_k$) and its confidence interval (CI) width (in terms of the Pearson correlation coefficient).

Specifically, this study posed the following null hypothesis:

1) Given the number of criterion traits, there is no independent variable (TCORR, NITEMS and SSIZE) main effects on $r_{ZCS}$;

2) Given the number of criterion traits, there are no 2-way independent variable (TCORR, NITEMS and SSIZE) interactions effect on $r_{ZCS}$.

3) Given the number of criterion traits, there is no 3-way independent variable (TCORR, NITEMS and SSIZE) interaction effect on $r_{ZCS}$.
CHAPTER II

LITERATURE REVIEW

It is scoring is dependent on item format. Nine main item formats have been distinguished, including the completion item, the multiple-choice item, and the ranking item, etc. (McDonald, 1999). For each item format, there is a usual method of scoring, and a key is provided for each item. A correct/keyed answer will receive a unity score, while a non-keyed answer receives a score of zero. This kind of scoring method assumes a known dimension and a direction on that dimension. It is suitable for items that measure a known dimension. However, ranking items, especially those yielding internal criterion-based rankings, usually elicit responses with an inherently complex dimensional structure and therefore demand more complicated scoring procedures (McDonald, 1999).

Moreover, the statistical treatment of the rank numbers (e.g., 1, 2, 3, and 4 for four options) is not simple due to the ipsative (or forced choice) data issue (Baron, 1996; Maydeu-Olivares & Brown, 2010). In order to have one key for a ranking item, the dimensional structure of the rankings should be examined first and found to be unidimensional (e.g., an ordered set such as a set of magnitudes). Otherwise, other scoring methods should be considered.

---

1 Nine item formats include completion or constituted-response items, multiple-choice items, dichotomous items, checklists/ frequency-count items, matching items, ordered-category/rating scale/ Likert scales/ graded response items, unordered categorical items, forced-choice items, and rankings (McDonald, 1999, chapter 2).
Basic Concepts and Main Characteristics of Rankings

A ranking is a set of alternatives ordered by a subject based on certain criteria (Kendall & Smith, 1940; McDonald, 1999). There are several important characteristics of rankings data (Marden, 1995) listed below.

1) The basic analysis unit of ranking consists of \( n \) judges ranking a set of \( m \) stimuli, the set of \( m \) stimuli denoted by \( S = \{S_1, S_2, \ldots, S_m\} \). The stimuli are identified with integers.

2) Two most useful representations of a ranking are order vector and rank vector. A rank vector lists the ranks given to the stimulus, “1” denoting the most favored and “\( m \)” denoting the least favored. An order vector lists the stimuli in order from the most favored to the least favored. Both orderings and rankings will be permutations of the first \( m \) integers. Rankings are generally reside in the space \( R_m = \{\text{Permutations of the ranks } \{1, 2, \ldots, m\}\} \), e.g., \( R_3 = \{123, 132, 213, 231, 312, 321\} \), whereas orderings generally reside in the space \( T_m = \{\text{Permutations of the objects}\} \), e.g., for A, B, and C, \( T_3 = \{ABC, ACB, BAC, BCA, CAB, CBA\} \).

For both expressions of rankings and ordering, there needs to be an \textit{a priori} identification of the objects with the integers, e.g., assigning 1 to A, 2 to B and 3 to C. People typically do not deal with all the rank permutations but only a sample of \( n \) rank vectors, e.g., \( y^{(1)}, y^{(2)}, \ldots, y^{(n)} \in S_m \).

3) Distance between two rankings is defined as spatial element and is a function of rank vectors \( d(y, y^*) \) of \( y^* \) and \( y \in H_m \), \( y \in H_m \) denotes ranking \( y \) is one of all possible rankings \( H_m \). If \( y^* \neq y \), then \( d(y^*, y^*) = 0, d(y^*, y) = d(y, y^*) > 0 \). A
spatial distance between two rank vectors depends on how one travels: along a straight line (Euclidean distance); along a great circle on the sphere; moving in unit steps parallel to the axes (Spearman or Footrule distances); or stepping between connected vertices of the polytope (Kendall’s distance).

- **Euclidean Distance.** In Cartesian coordinates, if \( p = (p_1, p_2, \ldots, p_d) \) and \( q = (q_1, q_2, \ldots, q_d) \) are two points in \( d \)-dimensional Euclidean space, then the distance between \( p \) and \( q \) is:

\[
 d_E(p, q) = \sqrt{\sum_{i=1}^{d} (q_i - p_i)^2}
\]

- **Mahalanobis Distance.** The Mahalanobis distance of a multivariate random variable \( x = (x_1, x_2, x_3, \ldots, x_N)^T \) from a group of values with mean \( \mu = (\mu_1, \mu_2, \mu_3, \ldots, \mu_N)^T \) and covariance matrix \( S \) is defined as:

\[
 D_M(x) = \sqrt{(x - \mu)^T S^{-1} (x - \mu)}
\]

\( D_M \) reduces to the Euclidean distance if the covariance matrix is the identity matrix.

- **Spearman and Footrule Distance.** Both Spearman distance and Footrule distance are a function of a pair of different ranks assigned to one stimulus. Spearman distance is the square of Euclidean distance:

\[
 d_{\text{Spear}}(y^*, y) = \sum_{s=1}^{m} |y_s^* - y_s|^2
\]

Footrule distance goes as:

\[
 d_{\text{Foot}}(y^*, y) = \sum_{s=1}^{m} |y_s^* - y_s|
\]

where \( s \) denotes stimulus, \( y \) is the rank assigned to stimulus. \( |y_s^* - y_s| \) is paired comparison of ranks for a specific stimulus \( s \).
• *Kendall’s Distance.* For two stimuli $i$ and $j$ ($i \neq j$) assigned different ranks by different subjects, $y^*$ and $y$, for $(y^*_i - y^*_j)(y_i - y_j) < 0$, the Kendall’s distance between two rank vectors is the number of discordant pairs,

$$d_{Kend}(y^*, y) = \sum_{i \leq j \leq m} I[(y^*_i - y^*_j)(y_i - y_j) < 0].$$

If two judges rank order stimuli A B C, Kendall distance is the distance between two judges.

4) Rankings data are multivariate data, the stimuli representing the variables. Hence, multivariate method can be applied to the rank data. Meanwhile, simple descriptive statistics, cluster analysis, multidimensional scaling, factor analysis, etc., all help to analyze the ranked data.

5) Rank data are highly structured. The fundamental structure is the “permutation polytope”, which is created by plotting the rank vectors in Euclidean space, then connecting the points.

6) Full ranking assigns a complete ordering to the objects from the best one of the objects to the worst one of the objects.

7) Partial ranking is a ranking process when judges rank their favorite $q$ out of $m$ ($q < m$) stimuli.

**Teaching Situation Reaction Test**

The *Teaching Situation Reaction Test* (TSRT) provides a context and an example of SRTs for the problem that this study investigates. The TSRT measures teaching related-attributes and the type of structure that the subject uses in the classroom by way how the subject reacts to various teaching situations. The subject reacts to common aspects of teaching such as planning, classroom management, and teacher-pupil
relationships. The ranking items in the test were intended to balance among the dimensions of objectivity, control, sociability, reflectiveness, empathy, and confidence although the test does not measure discretely for each of these dimensions (Hough & Duncan, 1965). The TSRT was administered with other criterion instruments (e.g., the objectivity scale and the sociability scale of the Guilford Zimmerman Temperament Survey, the confidence scale of the 16-PF Questionnaire, the reflective scale of the Thurstone Temperament Schedule, and the Intraception (Empathy) Scale of the Edwards Personal Preference Schedule) in attempts to identify the dimensions/attributes and the theory underlying it (Amidon & Powell, 1966; Duncan, Hough, & Thompson, 1966; Frye, 1972; Hough & Duncan, 1965; Hough, 1965a; Murray, 1969). These studies showed the TSRT measures multiple attributes/dimensions. For example, Murray’s study (1969) found the TSRT scores consistently appeared to be negatively related to control score and positively related to empathy scores.

The TSRT has been studied frequently in pre-service teacher education and in-service teacher evaluation in many thesis and dissertations (Flora, 1970; Furst, 1965; Gallo, 1970; Gold, 1966; Graening, 1971; Hall, 1965; Murray, 1967, 1969; Sobelman, 1971; Strawitz, 1970; Stauffer, 1973), in combination with a number of papers and presentations (Carline, 1970; Clayton, Thomas, et al. 1969; Lawson & McClernon, 1969; Moskowitz, 1967; Murray & Treanor, 1972; Treaner & Murray, 1971). For example, when studying the verbal and cognitive behavior patterns of selected secondary social studies teachers, Treaner and Murray (1971) used the TSRT score as a control variable to assist in the selection of subjects, i.e., those whose TSRT scores were in the upper and those whose TSRT scores were in the lower quartiles of the group. The TSRT was also
used to identify effective and less effective secondary school biology teachers in combination with the *Teacher Rating Scale* and the *Student Opinion Questionnaire* (Gold, 1966). In a study on in-service elementary school teachers' teaching behavior and students' achievement (Carline, 1970), pretests included intelligence tests and *Stanford Achievement Tests* in arithmetic as well as the TSRT.

The TSRT begins with an overall classroom situation narrative describing and establishing the general context (Appendix A, quoted from Murray, 1969). That is followed by 48 situational items, each with four behavioral options to be ranked by the respondent in order of preference. Each item presents a more narrowly defined classroom situation or encounter. The four alternative options (A, B, C, and D) to be ranked describe actions the subject might choose to take in response to the item situation. The subject is to supply the ranks that he/she assigned to the alternatives: i.e., 1, 2, 3, or 4 for each item, by inserting respectively the numbers in the spaces provided in the answer sheets under (a) (b) (c) and (d). The most desirable choice should be labeled 1 and the least desirable 4.

For example, if he/she ranks Alternative A first, Alternative C second, Alternative B third, and Alternative D last, his/her decision of the ranking item will be “1324”. This ranking will be compared to the key provided by the test writers for this item, e.g., “2134”. The key and the scoring system guide the raters to determine the number of correct decisions a respondent makes for each item. The maximum number of correct pairwise comparison decisions for each item is 6; and the minimum number of correct decisions is 0. Thus the highest possible score for each item is 6 and the lowest possible score for an item is 0. The highest possible test score is 48x6 = 288.
A Measurement Problem with the TSRT

As mentioned above, the TSRT was designed to embed several attributes in the items, such as control, sociability, and confidence. Some studies validated the TSRT with other criterion instruments that measure these attributes respectively, e.g., control, sociability, or confidence (Amidon & Powell, 1966; Duncan, Hough, & Thompson, 1966; Frye, 1972; Hough & Duncan, 1965; Hough, 1965a; Murray, 1969). The findings were significant but the criterion validity coefficients between the TSRT score and these criterion tests were surprisingly low ($r = -.282 \sim .013$). Criterion validity is established when the scores obtained on the test of interest are similar to the scores obtained on one or more external well-established measures (the criterion) (Engel & Schutt, 2010). Low criterion validity indicates either the test was not measuring the trait that was intended or the scoring key applied to the test responses did not allow for the test score composite to manifest the dimensionality of the traits. It is possible that the scoring method was problematic or that alternative scoring methods may result in different criterion validity estimates.

The “one key for each item” scoring method that the TSRT used assumes the response continuum is unidimensional, which does not match the multidimensional domains designed in the response matrix for the TSRT. This may be problematic in that it creates a possible discontinuity between the internal-referent of the respondent and the external-referent of the expert key.

External Criterion vs. Internal Criterion

The dimensional structure of rankings results from respondents interacting with stimulus options in the items (Gessaroli & De Champlain, 2005; Hattie, 1985;
Nandakumar, 1991; Reckase, 1990; Tate, 2003). The interactions between subjects and ranking items are guided by some ranking criterion. Distinguishing the different types of criteria for ranking decisions is very important for estimating the dimensionality of the response matrix. Therefore, this study explicitly distinguishes two kinds of rankings based on the criterion on which the ranks are decided – internal vs. external criterion-based rankings, despite that “ranking data” and “preference ranking data” have been used interchangeably (Luce, 1959; Bennett & Hays, 1960; Coombs, 1964; Critchlow, 1985; Diaconis, 1988; Fligner & Verducci, 1993; Gormley & Murphy, 2006; and Marden, 1995).

When test writers provide an external specific criterion in a ranking item for the respondent to compare the options against, the elicited rankings are based on external criterion. Typically, there is only one external criterion, i.e., one named attribute for an item. For example, in a chemistry test item, examinees were told that they were expected to arrange substances represented by their chemical formula in order of increasing value of a given property (Maeyer & Talanquer, 2013). There also could be several external criteria, one criterion at a time for an item (e.g., Böckenholt, 1992, 1996; Hershey & Tanaka, 1992). In such cases, a respondent will provide multiple rankings for an item, but each ranking will be based on one criterion. For example, in a tri-attribute ranking study about voting behavior, respondents were asked to rank presidential candidates three times, first with respect to overall preference, second with respect to the extent to which they agreed with the candidates’ positions on economic, and third on education issues (Böckenholt, 1992). In this case, each respondent gave three rankings, each of which was based on preference, economic, or education.
In external criterion-based ranking, the respondent is clearly instructed about which attribute should be involved in the ranking decisions. Therefore, it can be expected that the observed rankings will likely cluster around or involve the dimension of the named attribute, e.g., economic issues or education issues in the tri-attribute ranking study (Böckenholt, 1992).

When there is no external criterion specified for the respondent, the respondent can interact freely with the ranking item. In this case, the subject-item behavioral sample is an illustration of dimensional trait expression of that subject. That is, item alternatives are ranked based on the subject’s own personal preferences or judgment regarding the importance of the alternatives (McDonald, 1999). The resulting rankings, then, are internal criterion-based rankings; often referred to as preference rankings in the literature (Marden, 1995). For example, when respondents were asked to rank (from 1st to 8th choice) various possible life-sustaining therapies based on their judgment and without being supplied with a specified criterion, the resulting rankings formed internal criterion-based or preference rankings (Christakis & Asch, 1994; 1995).

The freedom in internal criterion-based or preference ranking decisions does not come without complication. For example, the dimensionality of the observed rankings is totally unknown since the criteria for the rankings are not explicitly indicated to the respondent, which can result in any dimension(s). Thus, the dimensional structure of the internal criterion-based rankings is more complex than the external criterion-based rankings. In point of fact, this structure reflects the internal structure of the subject rather than the external imposed structure of the item alternatives. From a test designer perspective, it must be assumed that the target audience for the test will perceive those
alternatives similarly within the situational context but will evaluate them personally. Such concerns about the dimensional structure of preference rankings should be considered whenever a scoring method for internal criterion-based rankings is being developed.

Without any external criterion provided to respondents, the internal criteria considered by different respondents may or may not be the same. According to Thurstone (1927), to compare stimuli, people generally first identify what quantitative or qualitative attribute(s) they will use to compare the stimuli/objects. This discriminial process assumes that a subject aggregates stimuli information into a “unidimensional” attribute and assigns different amounts of this attribute to each stimulus. It is unidimensional only in the sense that each subject may be considered to rank order a set of stimuli on the basis of his/her preference. This is different from stating that a scale is measuring one attribute. This “unidimensional” attribute defines the psychological continuum/scale and the compared stimuli are allocated on the continuum. It can be conceived of as an average scale for subjects. The position of an object on that dimension is characterized by a stochastic variable.

Coombs (1964) named this scale “J (Joint) scale” in the sense that both subjects and stimuli are jointly represented as points on that continuum. When the J-scale is folded at the point representing a subject, it becomes this subject’s I (individual) scale which displays the preference ranking of the stimuli for this subject. The point representing the subject is the subject’s ideal stimulus and hence is called “ideal point”. Each subject ranks all of the stimuli in order of their increasing distances from his/her ideal point. When rankings are already yielded, unfolding the rankings (hence the name)
allows to infer the order and spacing of subjects and stimuli on the continuum. Coombs’ assumption of unfolding as a quantitative theory of psychology has been supported by empirical studies (e.g., Kyngdon, 2006). For example, voters’ preferences for presidential candidates if, for instance, each voter could be characterized by his/her position on, say, the conservatism-liberalism continuum and he/she favored the candidates according to the distance between their positions on the continuum.

Preference data reveal the respondent’s state or level on the attribute(s) on which the competing stimuli are compared to each other. Preference rankings assume that the more reasonable reactions are caused by high levels on the involved traits (Thurstone, 1928; Likert, 1952; Newcomb, 1953; Rokeach, 1968; Combs & Snygg, 1959; Flanagan, 1954; Lewin, 1936). These "involved" traits are primary causal factors. This is a very important assumption because it provides a sufficient condition for this study to infer a subject’s latent traits based on his/her preference rankings.

**Keyed vs. Non-Keyed Ranking Items**

In addition to the distinction of external vs. internal criterion-based rankings, this study also differentiated ranking items with or without a key. A ranking item in a questionnaire or survey instrument usually focuses on options and and is not graded according to a specified key. For example, the Rokeach Value Survey (Rokeach, 1967), the patient treatment decision-making preference cards (Degner & Sloan, 1992; Ramfelt & Lützén, 2005), and the World Values Surveys (Inglehart et al. 2000) all include such non-keyed ranking items.

In contrast, a ranking item in a test normally focuses on subjects and requires a key against which a respondent’s answer can be scored, e.g., the Ranking Task Exercises
in Physics (ed. O’Kuma, Maloney & Hieggelke, 2003), the Virtual Environments for Dismounted Soldier Simulation, Training, and Mission Rehearsal (Knerr et al., 2003), the pre-employment screening personality test (Blinkhorn, Johnson, & Wood, 1988), and the TSRT (Hough & Duncan, 1965). Normally, there is a key which is a ranking pattern provided for each item by the test writer. For example, the TSRT has 48 items and an expert-key is provided for each item. A subject’s preference ranking will be scored against the key. A response in full agreement with the key will get a full score; a response in partial agreement with the key will get a partial score; and if there is no agreement, then a score of zero. The current study concentrated on situation reaction tests composed of keyed-ranking items.

**Dimensional Structure of Internal Criterion-based Rankings**

There are two types of dimensional structure of preference rankings: simple structure (SS) and approximate simple structure (APSS) (Roussos, Stout & Marden, 1998). SS occurs when each ranking item is commonly related to only one trait/attribute/construct. APSS occurs when each item has some amount of discrimination on all the dimensions of the test but is (expected to be) *primarily* measuring just one or several of those dimensions. This primary dimension(s) is the dominant or major dimension and the other dimensions are regarded as minor dimensions. SS normally indicates unidimensionality in the test data while APSS usually indicates that a test is multidimensional. APSS is more realistic than SS because in practice it is very difficult to construct perfectly unidimensional items, especially for tests that measure multiple constructs like the internal criterion-based ranking item tests (Gessaroli & De Champlain, 2005; Humphreys, 1952, 1962, 1970, 1981, 1985, 1986). The dimensional structure
underlying a response matrix of internal criterion-based rankings is often dependent upon
several minor dimensions in addition to the primary dimension (Messick, 1995; Traub, 1983).

Two major sources contribute to the dimensional structure of internal criterion-
based rankings: the dimensionality of items (Böckenholt, 1996; Roussos, Stout & Marden, 1998) and the dimensionality of subjects’ interaction with items (Gessaroli & De Champlain, 2005; Hattie, 1985; Nandakumar, 1991; Reckase, 1990; Tate, 2003). On one hand, when a ranking item is written, the options in it might be designed to involve
different number of dimensions, unidimensional or multidimensional (Böckenholt, 1996). On the other hand, with unidimensional options (i.e., every option in an item only
corresponds to a common attribute/dimension), the ranking of the options results through within-attribute comparisons (different options are evaluated on the same attribute).

Simple structure normally results from having unidimensional options in an item. With multidimensional option (i.e., each option corresponds to a different attribute or every option corresponds to several attributes measured by a test), the ranking decision requires crossing-attributes comparisons (the options are compared on different attributes) and mostly results in multidimensional rankings. Multidimensional options in a ranking item often result in approximate simple structure (APSS) or in more complex dimensional structures.

Nonetheless, some researchers noted that test dimensionality should not refer to
the particular set of items of a test. Rather, it is a property of the subjects interacting with
or responding to the set of items (Gessaroli & De Champlain, 2005; Hattie, 1985;
Nandakumar, 1991; Reckase, 1990; Tate, 2003). Different subjects might respond to the
same set of items differently. When subjects rank a list of options according to their own preference without any specified criterion, there could be various amounts of multiple criteria involved depending on how a subject interacts with the items. It is very likely that some people may evaluate a list of multi-attribute options only on one of the attribute at issue while others consider all the attributes for their ranking decision. This will complicate the dimensional structure of the test scores (and is one reason for the frequent assumption of communality among subjects' perception of the alternative stimuli in terms of their common situational context).

**A Possible Solution to the TSRT Scoring Problem**

When ranking a given set of \( m \) multi-attribute alternatives, a subject reacts as described below:

- The subject favors/weighs/rates differently \( k \) characteristics/attributes of these alternatives.
- The subject rates each alternative (e.g., a set of teaching situation reactions) according to each attribute.
- The subject essentially performs a personal multi-dimensional analysis (consciously or mostly unconsciously) in deciding how to rank the alternatives.

Suppose there are several ranking items measuring multiple attributes including three known ones: objectivity, empathy, and control. If there are four alternatives (e.g., \( A \), \( B \), \( C \), and \( D \)) in each item for a subject to rank, there are 24 different ways to fully rank four things – 24 permutations. Hence, this subject could have expressed any one of 24 different preference patterns. Any of these 24 patterns can be used as a key depending
what attribute is being measured. By choosing “ABCD” as key to measure an attribute, say, objectivity, and by comparing a subject’s response to “ABCD” (using a scoring rule), a number (a measurement from 0 to 6 if the “scoring rule” follows the Kendall distance) can be obtained that estimates this subject’s position in the state space along the dimension of objectivity – a measure of the importance this subject gives to objectivity or the degree to which he/she values objectivity.

And the same can be done for the other two attributes. “ABCD” was chosen as the key for objectivity because it is the “best” of 24 possible key choices. As the best key, it would orient, position, and spread out a population like this subject’s along the "objectivity" dimension correctly in an ordinal sense. Some other key would do the same for "empathy" or "control."

The situation, the alternatives, and the keys make up the basic item unit. This kind of item unit would yield three distinct measurements (e.g., objectivity, empathy, and control), not simply one composite measurement. Replicating this whole experiment with the same subject but another set of four situational reactions (another test item), using three more “best keys” would yield another set of similar measures or estimates of this subject's position in the “objectivity, empathy, and control” state space.

Combining these two sets of three measures together (as in a test) should produce an even more accurate estimates of this respondent’s true position than either set alone. This subject’s position on each of the attributes (in other words, the degree to which this subject values each attribute when choosing an alternative) would be measured in the same way in both cases by both situational items. Thus measurement quality would be a function of test length.
Modeling and Analysis of Non-keyed Rankings

The non-keyed approaches to analyze rank data mostly focus on aggregating multiple rankings (Marden, 1995). Some approaches focus on classifying options or subjects on the compound “unidimensional” preference scale by Thurstone’s law of comparative judgment (Brennett, 1960; DeSarbo, Young & Rangaswamy, 1997; Falahee & MacRae, 1995; Lehmann, 1975; Marden, 1995; Schönemann & Wang, 1972; Thurstone, 1927). This compound “unidimensional” scale was developed into a multidimensional preference space, i.e., by extending the one dimensional scale to $n$ dimensions (Richardson, 1939; Young & Householder, 1938) and by setting stimuli and subjects jointly in that multidimensional space.

Other approaches seek to define the center and spread of a data set and to find clusters of stimuli or subjects (Yemelichev, Kovalev & Kravtsov, 1984; Cohen & Mallows, 1980; Thompson, 1993). The analysis task in these cases focused on determining the dimensionality of the rankings in order to obtain the configuration of the subjects and/or stimuli in this preference space (e.g., Bennett & Hays, 1960; Coombs, 1964; Delbeke, 1968; Richardson, 1939; Shepard, 1962a, b, 1972; Torgerson, 1952).

Modeling of non-keyed rank data consists of finding the probability distribution $p$ that is closest to the observed probability distribution in the space of permutation of ranks (Marden, 1995). Many ranked data studies involve aggregated statistics (Fligner & Verducci, 1993; Marden, 1995), such as how to estimate the differences between means of hypothetical variables assumed to underlie the judgments (Guttman, 1944; Croon & Luijkkx, 1993), how to determine group consensus (Goldberg, 1976; McCullagh, 1993; van Blokland-Vogelesang, 1993; Walker, 1993), how to aggregate information about a
specific alternative (Marley, 1993), or how to describe an individual’s characteristics in
general (Beaver et al, 1996; Beaver, Bogg & Luker, 2001; Ramfelt & Lützén, 2005).

In the early days, ranked data were mainly collected in experiments involving
paired comparisons between treatments. For example, Thurstone (1927) studied this
problem on the assumptions that a linear variable is involved in ranking and that
perceptible differences exist among the stimuli to be compared. A variety of models
have been developed to analyze ranked data (Luce, 1959; Critchlow, 1985; Diaconis,
1988; Fligner & Verducci, 1993; Gormley & Murphy, 2006; Marden, 1995, etc.) since
then. Ranked data has been studied as a problem of testing hypotheses, e.g., tests of
consistency of a judge and tests of agreement among several observers (Kendall & Smith,
1940) and also as a problem of estimation, e.g., estimating the true ratings or preferences
of stimuli on a particular subjective continuum (Guttman, 1944; Bradley & Terry, 1952).

Guttman developed a method to quantify paired comparisons, assigning a
numerical value to each of a number of items which is believed to best represent the
comparisons in some sense. The ranked data modeling include the general probability
models, Thurstonian order statistic models, distance-based models, paired comparison
models, multistage models, sufficient statistic models, loglinear models, ANOVA-like
models, and unfolding models, etc. (Marden, 1995). In general, the fundamental structure
of ranked data is the permutation polytope (Marden, 1995), which uses distances
throughout the permutation of the ranks to define the center and spread of a data set, and
to find clusters of judges (Yemelichev, Kovalev, & Kravtsov, 1984; Cohen & Mallows,
1980; Thompson, 1993).
The most common category of ranking models is the probability model (e.g., Luce, 1959; Plackett, 1975). The probability of a ranking is interpreted as utility ordering, where the utility of the alternative ranked first is larger than the utility of that ranked second which is larger than that ranked third, and so on. From the probability model derived a number of more complicated models that are closer to real life situations. For instance, the multilevel logistic regression model (Skrondal & Rabe-Hesketh, 2003) addressed multilevel rankings; Deng (2007) developed similarity preference models with a similarity-based approach (e.g., Coombs, 1954) to rank multi-criteria alternatives for solving discrete multi-criteria problems. The overall performance index of each alternative across all criteria is determined by the degree of similarity between each alternative and the ideal solution. Logistic regression is currently the standard method to model ranked data (Skrondal & Rabe-Hesketh, 2003). But this has been demonstrated to involve a questionable independence assumption known as “Independence from Irrelevant Alternatives” (IIA). The analysis of multivariate choice data is now becoming a routine matter because of significant computational advances in estimating the parameters of the latent utility distribution (Böckenholt & Tsai, 2006; Caffo & Griswold, 2006).

New developments of models include a framework of multilevel logistic regression for polytomous rankings (Skrondal & Rabe-Hesketh, 2003) which accommodates dependence induced by factor or attribute at different level; mixture models based on multistage ranking models to investigate the types of application behavior that are exhibited by college applicants (Gormley & Murphy, 2006); and so on. Even the joint modeling of discrete and continuous choice outcomes is starting to be
commonplace (Caffo & Griswold, 2006). Some studies focused on the stimuli/objects/events, such as the number of each object attaining each rank or each pair of objects attaining each pair of ranks (e.g. Bradley & Terry, 1952; David, 1988; Diaconis, 1988, 1989; Silverberg, 1980; Critchlow & Verducci, 1992; Gormley & Murphy, 2006) or the decomposition of the data into factors based on objects comparisons (McCullagh, 1993).

Some studies intended to visualize or numeracize the frequency of a ranking (Schulman, 1979; Thompson, 1993; Cohen & Mallows, 1980; Croon, 1989). Huber (1985) and Goldberg (1976) explored projections of polytope in order to pick out the features of the data: e.g. whether most of the data are clustered near a single ranking or whether there are groups of judges clustered around disparate rankings. Hollander and Sethuraman (1978) were interested in studying the judges, but their focus was on comparing populations of judges, not individual differences between judges.

Some models are more adept at revealing groupings of judges than others: e.g. cluster analysis of rank data aims at the identification of groups of judges with a common preference behavior. The distance model, Plackett-Luce, and Mallows models presume a certain degree of unimodality in the judges’ rankings. More complicated models such as Barbington Smith models and the orthogonal contrast models (Marden, 1995) are flexible to capture additional dimensionality of the polytope, but at the sacrifice of interpretation. Alternatively, judges’ responses to stimuli could be measured through unfolding techniques (Coombs, 1964) which can study the objects and the judges and/or their interactions.
Most of these studies examine large sample ranking patterns on a single “item” with multiple, possible many, alternatives. This current study focuses on instruments developed using multiple but similar situational “items,” each with relatively few (e.g., 2, 3, 4, or 5) alternatives to be ranked. In this sense this study relies on replication (multiple items) rather than sample size to magnify/intensify the effects being measured.

In addition to various mathematical models, graphical techniques have also been developed to visualize rankings in polytopes (Cohen & Mallows, 1980; Croon, 1989; Schulman, 1979; McCullough, 1992; Thompson, 1993). Suppose there are m objects to rank. If we take each of the m! permutations of the numbers (1, 2,..., m) as the coordinates of a point in M-space, then obviously all the m! points are equidistant from the origin. For instance, for m = 4, the "shape" made by joining each point to its nearest neighbors is what's called a "truncated octahedron" with six square faces and eight hexagonal faces, a total of 14 faces.

**Tests of Ranking Items**

Keyed ranking items – which are often presented as ranking tasks in situation reaction tests - are used by educators to elicit students’ comparative judgments about a variety of arrangements of a specified situation (Maloney, 1987; Siegler, 1976). A ranking task commonly presents students with a list of (three to eight) slightly different options relevant or related to a particular situation. The students compare and rank order the options based on a specified criterion or on their best judgment. The basic structure of ranking tasks usually has four elements: a description of the situation, a set of options to be compared, a place for the respondents to write their answers, and a place to write out an explanation for their answers (Maloney, 1987).
There are two major concerns about such ranking item tests. First, ranking tasks demand high cognitive load (Streiner & Norman, 2003; Dillman, 2009), including the limited capacity of human working memory, the visual scan time associated with surveying the list of options, and the difficulty of assigning a ranking to the options when multiple options are present. When a respondent begins a ranking task, he/she must familiarize himself or herself with the options by visually scanning them (perception). This process is repeated as options are removed from the source list and placed in the ordered list. Assigning a rank to an option can be a long and tedious process (Alreck & Settle, 2004; Berlucchi & Aglioti, 1997). The burden on respondents increases with the number of optional tasks (Dillman, 2009). It is recommended that ranking tasks contain 10 or fewer options if they are given in written form (Alreck & Settle, 2004) and no more than 6 options (Miller, 1956; Sudma & Bradburn, 1982) if they are given via the telephone. Failure to recognize these human limitations in item design may lead to cognitive fatigue of respondents who then may become careless when responding (Parten, 1950).

Second, because the ranking does not allow for inconsistent preferences, it may fail to provide a real picture either of the subject’s preference or of the variation of the attributes compared among the alternatives. This problem was studied by Kendall and Smith’s method of paired comparisons (1940). Their method allows for inconsistencies in decision and they proposed that if judgments are highly concordant and most of the inconsistencies are confined to certain alternatives, the alternatives can be considered indistinguishable on the attributes at issue. But if the subjects scatter their inconsistencies over the whole space, the linearity of the attribute variables is under question.
Strengths of Ranking Item Tests

Compared to standard test items such as two-choice (T/F) or multiple-choice items, ranking tasks items are relatively novel to the respondents and serve as variations that encourage survey or test attention and response. However, such items have been receiving more recognition over the years in such subjects as physics, chemistry, astronomy, math, language instruction, and pre-service teacher education for facilitating teaching and learning (e.g., O’Kuma, Maloney & Hieggelke, 1999; Hudgins, 2005; Hough & Duncan, 1965; Nation, 1979, 1995). Ranking tasks may be used in many ways: homework assignments, group exercises, test questions, in class assessment and generation of class discussions.

O’Kuma, Maloney and Hieggelke (1999) summarized the major merits of ranking tasks as follows:

- Ranking tasks often elicit students’ common sense about a specified situation rather than a memorized response which provides insights into how students think, not just whether or not they obtain the correct answer, which can help teachers to adapt their teachings;
- Ranking tasks enforce students to think through the situation whereas students have got used to such more familiar format of items as multiple choices and developed coping mechanisms that allow them to answer without really thinking them through;
- Ranking tasks can help to lead students to legitimately think about and understand concepts, and develop the types of understandings and reasoning
skills that are required to generate more thoughtful responses in academic tasks.

Although the effects of ranking tasks on students’ learning have not been intensely studied, the available literature suggested that adding ranking task exercises to traditional classroom instruction can significantly improve student understanding of key introductory astronomy concepts (Hudgins, 2005).

Appendix B shows two examples of astronomy ranking task retrieved (Oct. 11, 2013) from Professor J. Allyn Smith’s Laboratory website at the department of Physics & Astronomy of Austin Peay State University, Clarksville, TN.

**Ranking Strategies**

Karth (2011) identified four possible strategies used by participants to complete the ranking process (Table 1). Two of these were called "chunking" strategies (Level Driven and Similar Option Driven) while two others were "non-chunking" strategies (Numeric Rank Driven and Individual Option Driven). He found that task format did not influence the frequency of selection of a participant’s strategy. In most cases, participants appeared initially to select a rank (first, ninth, etc.) and then to assign an option to that rank. This is a chunking strategy. Karth concluded that the chunking strategies were favored over the non-chunking strategies.
Table 1

**Ranking Strategies Based on Personal Preference**

<table>
<thead>
<tr>
<th>Chunking Strategies</th>
<th>Non-chunking Strategies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Level Driven Strategy: While or after reviewing all options, participants sort each option into a major category such as highly preferable (top choices), not preferable (bottom choices), and neither highly preferable nor not preferable (remaining choices). They then order the options within each category. For example, the participant ranks the top 3-5 options, then the bottom 3-5 options, and then orders the remaining options in the middle of the scale.</td>
<td></td>
</tr>
<tr>
<td>2. Similar Option Driven Strategy - Participants sort each option into a major category based on similarities between the options. Major categories are then ordered based on preference, and the options within each category are ordered. For example, the participants create major categories (e.g. financial, academic, or extra-curricular considerations) and then after prioritizing the categories, they order the options within each category.</td>
<td></td>
</tr>
<tr>
<td>3. Numeric Rank Driven Strategy - Participants rank options in a serial method starting at one end of the ranking scale (high or low) and progressing in the opposite direction. For example, the participant reviews all options and determines which one has a rank of “1,” then reviewing the remaining options to determine which option has a rank of “2,” and so on, until all options have been ranked.</td>
<td></td>
</tr>
<tr>
<td>4. Individual Option Driven Strategy - Participants evaluate an option and assign a rank based on their estimation of where it should be ranked. They repeat this process with each subsequent option.</td>
<td></td>
</tr>
</tbody>
</table>
Criterion Validity

Test validity refers to the usefulness of inferences drawn from test scores for a specific purpose under certain conditions (Crocker & Algina, 1986). There are three classes of evidence test validity: content validity (which shows how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn), construct validity (which investigates what qualities a test measures by determining the degree to which certain explanatory concepts or constructs account for performance on the test), and criterion validity (which is established when the scores obtained on one measure are similar to the scores obtained with one or more well-established external measures (the criterion) of the characteristic or behavior in question (Engel & Schutt, 2010; Messick, 1990, 1995). The validity of a test essentially depends on its construct validity – the extent to which a test measures what it purports to be measuring (Cronbach & Meehl, 1955; Messick, 1981). Neither content- nor criterion-validity can stand alone to support the specific test use because score meaning for both tests and criteria is needed to justify testing application (Loevinger, 1957; Thorndike, 1949). However, these two forms of validity can provide rational basis for the proposed use (Guion, 1976; Messick, 1990). What is referred to in this study as structural validity - the relation of the test to the construct domain - involves both construct and criterion validity

While all the three types of test validity are important, criterion validity is the one that is of particular interest of this study. Engel and Schutt (2010) noted that criterion validation greatly increased confidence that the measure was measuring what was
intended and that it was a stronger form of validity than content validity as it was based on research evidence rather than subjective assessment.

In order to obtain criterion validity, subjects need to be measured on both the test of interest and the criterion instrument. The criterion instrument that researchers select can itself be measured either at the same time as the test to be validated or after that time. There are two kinds of criterion-related validity, *concurrent validity* and *predictive validity*. When a measure yields scores that are closely related to a criterion measured at the same time, there exists *concurrent validity*. A measure of blood-alcohol concentration could serve as the concurrent criterion for validating a self-report measure of drinking as far as they happen at the same period. *Predictive validity* is the ability of a measure to predict scores on a criterion measured in the future. GRE scores as a measure of academic success in graduate program study can be validated when compared with graduate school performance.

Criterion-related validity is evaluated by the degree of empirical correlation between the test scores and criterion scores. This correlation or *validity coefficient* serves as a basis for using the test scores to predict an individual's position on a criterion measure of interest, e.g., grade-point average in college. Criterion-related validity only emphasizes the selected parts of the test's external structure, focusing on relationships with measures held to be critical for a particular applied purpose. Thus, there are as many criterion-related validity coefficients for a test as there are criterion measures for it (Messick, 1990, 1995; Thorndike, 1949). We can see that criterion validity is largely dependent on the quality of test scores (their dimensionality, sensitivity and reliability).
Test Dimensionality

Test dimensionality is also of particular interest of this study because it is very relevant to the appropriateness of test scoring methods. To understand exactly what dimensionality is requires some introduction to the construct validity of tests. Construct validity has six aspects: content, substantive, structural, generalizability, external, and consequential aspects (Messick, 1995). The structural aspect of construct validity mathematically appraises the extent to which the internal structure of the assessment reflected in test scores is consistent with the structure of the construct domain at issue (Loevinger, 1957). An examination of the consistency between the internal structure of the test scores (or the mathematical latent trait) and the psychological structure of the construct (or the psychological latent trait) at issue involves the assessment of test dimensionality.

Some definitions of test dimensionality are statistically based and assume an item response theory (IRT) modeling approach: if \( k \) is the minimal number of latent traits possible for a \( N \)-item test such that the principle of local independence – for a fixed level of a latent trait, the covariances between all pairs of items in the test is zero – holds in practice, then the dimensionality of this test is \( k \) (Nandakumar, 1991). Stout (1987, 1990) argued that the traditional IRT assumptions of unidimensionality and local independence can be replaced by the weaker and psychologically more appropriate assumptions of essential unidimensionality and essential independence respectively. Essential independence requires only that, as the number of items \( N \) goes to infinity, the average covariance over item pairs is small in magnitude for all ability. The IRT definition is different from non-IRT based definitions of test dimensionality, such as those out of
McDonald's nonlinear factor analysis (McDonald & Ahlawat, 1974; McDonald, 1981) where a set of \( n \) tests or items is unidimensional if and only if the tests or the items fit a common factor model, generally nonlinear, with one common factor that is one latent trait.

Statistically based test dimensionality is distinct from cognitively based psychological dimensionality, which is a notion describing the number of abilities required to correctly answer all \( N \) items of the test (Reckase, 1990). The modeling of ability tests based on the factor analytic tradition has attempted to assess the number of dominant dimensions, such as McDonald's nonlinear factor analysis (1967) and Tucker, Humphreys, and Roznowski’s (1986) linear factor analysis based procedures to assess whether the number of dominant dimensions is one or greater than one.

Test dimensionality directly impacts the approach for analysis of test data. Many studies (Ansley & Forsyth, 1985; Drasgow & Parsons, 1983; Harrison, 1986; Reckase; 1979) posited that analysis methods for unidimensional and multidimensional data should be used discriminately. Depending on the amount of multidimensionality present in a test data, multidimensionality can have a deleterious influence on the performance of unidimensionally based linear and nonlinear models. The pioneering Reckase article (1979) used both simulated and real data sets and found that IRT estimates ability well if there is exactly one dominant factor, but estimates ability poorly when two or more dominant dimensions generate the data because it tends to track only one component of ability.

Drasgow and Parsons (1983) demonstrated through hierarchical factor analysis modeling of simulated test data that LOGIST (an IRT estimation program popular in the
1980’s) recovers the general latent factor and the item structure well where the general factor is sufficiently dominant, but is drawn to the strongest of the group factors as the dominance of the general factor is decreased. Harrison (1986) replicated these findings for several various test models. Yen (1985) investigated tests in which more difficult items involve more skills and hence depend on more latent traits and found that LOGIST often caused a systematic shrinkage in the estimated (assumed unidimensional) ability scale. Wang (1987, 1988) suggested that despite whether or not there is a clearly dominant dimension, unidimensional analyses of multidimensional data likely produce estimates of ability that are weighted averages of the multidimensional latent abilities of the multidimensional model generating the data. Nandakumar’s study (1991) based on essential unidimensionality implied that in cases of lacking unidimensionality in test data, one might use the theory of testlet scoring (Thissen, Steinberg, & Mooney, 1989); the multidimensional compensatory logistic modeling approach (Reckase, 1985, 1989); or disintegrate the test into several essentially unidimensional subtests, each scored separately.

Therefore, it is important to assess test dimensionality before conducting data analysis. Early techniques for assessing test dimensionality were based on test score homogeneity, reliability, or principal components. Those proved problematic because they seemed to have confused the concepts of homogeneity, internal consistency, and dimensionality (Gessaroli & Champlain, 2005; Hambleton & Rovinelli, 1986). For example, it was thought that higher reliability was indicative of a more unidimensional test. However, Green, Lissitz, and Mulaik (1977) showed that that was not always the case. With a five-dimension test, the coefficient alpha could still be high. McDonald
(1981) particularly noted that coefficient alpha should not be used as a decision criterion for unidimensionality. Recent methods of dimensionality assessment are more theoretically sound, based on the principle of local independence (McDonald, 1981) or the principle of essential independence (Stout, 1990). The principle of local independence is achieved when, for a fixed level of a latent trait, items in the test are mutually independent and hence the covariance between item pairs is zero. This mathematical definition of test dimensionality based on latent traits has been argued sometimes to fail to capture all the dependencies among the item responses (Goldstein, 1980; Goldstein & Wood, 1989). Moreover, it is possible to satisfy the mathematical definition yet not fully account for all of the psychological traits influencing the item responses. Meanwhile, there might be two psychological traits that affect the responses although local independence is satisfied by a single mathematical latent trait.

It has been pointed out (Gessaroli & De Champlain, 2005; Hattie, 1985; Nandakumar, 1991; Reckase, 1990; Tate, 2003) that when test items are related to more than one trait and those traits are correlated, this combination leads to complex multidimensional structures that are often difficult to identify with the existent dimensionality assessment methods. They concluded that, in general, those dimensionality assessment methods function best when the data has simple structure and that methods based on factor analysis enable a more detailed analysis of the dimensional structure including the relative strengths of each dimension and the relative strengths of each dimension on individual test items.

The current study did not intend to find a new model to capture the dimensionality of a test. Rather, it investigated a scoring method that is based on the assumption of
multidimensionality of data and that allows the multiple dimensions in the data to be recognized one by one individually. Failing to recognize one embedded trait should not and would not impact the recognition of other traits, correlated or not.

**Unidimensionality vs. Multidimensionality**

The assumption that a test is unidimensional is made for many types of item response models, i.e., the one-, two-, or three-parameter logistic or normal ogive models. However, as Hambleton and Rovinelli (1986) pointed out, the definition of the term “unidimensionality” and the approaches for assessing the presence or absence of it in a test are somewhat confounded in the psychometric literature. A typical definition of undimensionality is that a single latent trait can explain examinee performance in a given test. This type of definition is criticized as non-operational and abstract, e.g., it does not distinguish the mathematical aspect and the psychological aspect of test scores (Hambleton & Rovinelli, 1986). A more helpful definition was esposed by McDonald (1980) and Hattie (1985) who argued that the principle of local independence should be the basis for a proper definition of the unidimensionality assumption. The practical interpretation of the principle of local independence is that the covariances between all pairs of items in the test are zero. It should be noted that the assumptions of unidimensionality and local independence are logically independent (Nandakumar, 1991).

Nonetheless, many studies showed that real test data often cannot be modeled by locally independent unidimensional models (Ackerman, 1987; Ansley & Forsyth, 1985; Harrison, 1986; Reckase, 1979, 1985; Reckase, Ackerman, & Carlson, 1988; Thissen, Steinberg, & Mooney, 1989; Wang, 1987, 1988; Yen, 1985). This is particularly true for achievement test scores (Hirsch, 1989; Reckase, 1979; Yen, 1985). A number of
researchers argue that tests are almost always multidimensional (Humphreys, 1985; Lumsden, 1961; Reckase, Ackerman, & Carlson, 1988; Yen, 1985). For example, Thissen, Steinberg, & Mooney (1989) stress that tests such as reading comprehension tests often have clusters of items that are still inter-dependent, even with the dominant dimension removed, thus indicating other relevant dimensions exist. Consider the example of a paragraph comprehension test consisting of four paragraphs, each of which is followed by several content questions. Let each paragraph be about different content, e.g., the first paragraph is about astronomy, the second about chemistry, etc. Reading ability is the common ability influencing all items in this example. In addition, each paragraph is influenced by one other ability – namely, knowledge of astronomy in the first paragraph, knowledge of chemistry in the second paragraph, and so on – thus creating four more psychological dimensions. If there is between-examinee variation regarding the five psychological dimensions and prior content knowledge, the item response dimensionality will be at least five. Other than a common language (English), it is difficult to argue here that perceptions are communal and responses are personal.

Traub (1983) explains that test data are very likely to be multidimensional due to differential educational background, differential test speededness strategies, and/or differential guessing strategies. The belief that achievement test data are multidimensional has motivated the development of multidimensional IRT (MIRT) models (e.g., see Reckase & McKinley, 1983; Carlson, 1987; Reckase, 1985, 1989). Humphreys (1952, 1962, 1970, 1981, 1985, 1986) argued for the inherent multidimensionality of real test data and suggested that from the aspect of validity, tests
should be deliberately constructed to include numerous minor factors in addition to the dominant dimension.

Test dimensionality is not only a property of a particular set of items in a test, but also is a psychological property of the examinees interacting with or responding to the set of items (Gessaroli & De Champlain, 2005; Hattie, 1985; Nandakumar, 1991; Reckase, 1990; Tate, 2003). In fact, that is what the test is supposed to be testing! The dimensional structure of a test is often dependent upon several minor dimensions in addition to the hypothesized primary dimension (Traub, 1983). It can be simple or complex. Roussos, Stout & Marden (1998) described two types of item dimensionality structure: simple structure (SS) and approximate simple structure (APSS). SS occurs when in a test each latent trait influences independent clusters of items. That means, each item is related to only one trait. When there is only one latent trait underlying the responses, the test is considered unidimensional. APSS occurs when there are major or dominant dimensions and minor dimensions in a set of test score. In situation of APSS, each item of the test has some amount of discrimination on all the dimensions of the test but is primarily measuring just one of the dimensions. That is, a test of APSS can also be divided into clusters, but each cluster corresponds not only to a separate major test dimension but also to a certain number of minor dimensions. This makes the test multidimensional.

**Dimensional Structure of Internal Criterion-based Rankings**

Bennett (1956), Bennett and Hays (1960), and Hays and Bennett (1961) discussed the estimation of the dimensionality of preference rankings given by subjects for a set of $m$ alternatives. The maximum number of rankings generated by $m$ alternatives in $d$ dimensions was displayed in a table (e.g., Bennett, 1956, p.38). This information can help
to determine if the responses are unidimensional or not. For example, based on the table, if for four alternatives, the respondents presented more than seven different ranking patterns, and then the responses reflect more than two dimensions.

Böckenholt (1996) distinguishes between two sources of response dependencies in multi-attribute preference ranking data: *within-attribute dependence* (different options are evaluated on the same attribute) and *between-attribute dependence* (the same options are compared with respect to different attributes). That is, options in a ranking item can be designed to involve different number of attributes/dimensions. When every option only corresponds to a single common attribute/dimension, the ranking of the options is the result of within-attribute comparisons. When each option corresponds to a single different attribute or when every option corresponds to several or even all of the attributes measured by a test, the ranking decision requires crossing-attributes comparisons.

**Test Scoring**

**Four Types of Observations and Data**

In psychological measurement, four types of observable features of attribute were recognized (Coombs, 1964): preferential choice (Type I) – a ranking of the stimuli/alternatives; single-stimulus (Type II) – a dichotomous response; stimulus-comparison (Type III) – one of a pair of stimuli possesses more of a perceived attribute than the other; and similarities (Type IV) – one pair of stimuli is more similar than the other. Coombs’s four types of observations are mapped into four types of data through some scoring and scaling process. The data displays distances between subjects, between stimuli/objects, or between subjects and stimuli/objects (McDonald, 1999). Type I data result from preferential observations that ask the subject to express his/her preference of
the stimuli. In Type II data, a single stimulus is presented to the subject who needs to provide a dichotomous answer, e.g., yes/no, true/false. Type III data result from stimulus-comparison observations in which the subject is asked to judge which one of a pair of stimuli possesses more of a named attribute. Type IV data represent a judgment that members of one pair of stimuli are more similar than members of another pair.

It should be noted that although Type I data and Type III data may look alike, they denote different decision-making process. Type I data represent a relation between a subject and a pair of stimuli without specifying the attribute and a subject’s preference is based on his/her level of the relevant attributes (a proximity relation). Type III data represent a relation between two stimuli as perceived by the subject regarding a specified attribute. The judgment that one stimulus has greater degree of an attribute than another can be independent of the degree of this attribute of the subject (McDonald, 1999). This study focused on Type I data, ordinary preference rankings.

**Scoring and Scaling**

Transforming observations into data requires some scoring and scaling process. Scoring links an observation with a number, whereas scaling specifies the properties of this number (McDonald, 1999; Torgerson, 1958). *Scaling* is the process of setting up the rule of correspondence between observations of attributes of the subject or the stimuli and the numerical values assigned to the observations. It maps the stimuli or/and the subjects onto the number line representing an attribute. The established correspondence is called *scale*. A *scaling rule* specifies how an observable feature corresponds to an element in the real number system. The assigned numbers are *scale values* or *metric*.
Based on the extent to which the metric retains the properties of the real number line (i.e., order, distance and origin), scales are identified by four levels:\(^2\):

A. Nominal Scales (mere classification, e.g., A or B, and permit any one-to-one transformation);

B. Ordinal Scales (i.e., \(n(A) > n(B)\), permitting any monotone transformation);

C. Interval Scales (i.e., \(n(A) + n(B) = n(C)\), permitting any linear transformation); and

D. Ratio Scales (i.e., \(n(A) + 0 = n(A)\) and \(n(A)/n(B)\), permitting any ratio transformation).

In general, researchers are content with having scales with arbitrary origins and units and ordinal properties. However, some people, e.g., psychometricians, desire to find methods and models for scaling attributes of people or stimuli that possess higher scale properties, e.g., interval or ratio scales (Coombs, 1964; McDonald, 1999).

Scaling methods should match the types of observations (Torgerson, 1958). According to McDonald (1999), in Type I (preferential choice) and Type II (single-stimulus) data, the subjects and the stimuli are represented as points along one or more dimensions. In Type I data, the point representing a subject is a hypothetical ideal stimulus for that subject. Stimuli are rank ordered according to their increasing distances from the subject’s ideal stimulus. Type I data is often modeled by Coombs’s unfolding model (Coombs, 1964). In Type II data a subject is represented by a hypothetical test

\(^2\) For a pair of stimuli (A, B), \(n(*)\) means a scale value assigned to a property of A and B, e.g., \((n(A) - n(B))\) measures the dissimilarity of A and B.
item that would be the most difficult the subject could pass, e.g., the Walker-Guttman Scale\(^3\) and the Rash-type Models (McDonald, 1999). In Type III (stimulus-comparison) and Type IV (similarities) data, only stimuli as perceived by subjects are represented by points on an attribute specified (Type III) or attribute(s) unspecified (Type IV). Type III data compare two stimuli as perceived by a subject. Stimuli are represented as points on a line delineating a named attribute. The Thurstone pair comparison model (1927) is generally used to scale Type III data. In Type IV data, the distance between a pair of points \(n(A) - n(B)\), where \(n(\cdot)\) means a scale value assigned to a property) measures the dissimilarity of this pair of stimuli \((A, B)\). The attribute(s) is not specified in Type IV observation, just like the case for Type I Data. Type IV data involves multidimensional scaling in models of spaces of more than one dimension or attribute. This study treated the preference rankings yielded by Type I observations as at least ordinal scale data (Coombs, 1964; Kyngdon, 2006).

---

\(^3\) The Walker-Guttman Scale (McDonald, 1999) was developed by Walker (1931) and was inherited by Guttman (1950). For a set of \(m\) binary items, the number of possible response patterns is \(2^m\). Walker (1931) studied the response patterns of binary items in the context of cognitive items and defined an ideal answer pattern as displaying only \(m+1\) permissible response patterns. If the examinees are ordered on their ability and the items on their difficulty, the permissible patterns, an ideal answer pattern should show that an examinee who passes the most difficult item passes all items, and an examinee who fails the easiest item fails all items. Guttman (1950) applied such notions to attitude items and came up with Guttman scale or perfect scale. The good feature of the perfect scale is that it provides a one-to-one correspondence between the response patterns and the total scores of a respondent. We know all about the behavior from the total score. The dilemma for the perfect scale is that the method collapses as more items from the behavior domain are added in attempt to increase the measurement precision, because in practice it is very difficult to find/write items that give only \(m+1\) permissible response patterns. If an ideal answer pattern is not obtained, the causes can be unique variation/error or/and multidimensionality. Multidimensional joint space has been attempted to represent responses that give impermissible patterns in one dimension, but these attempts restrictively require there is no error in the responses (McDonald, 1999).
**Unidimensional Unfolding Theory**

*J-scale and I scale*

The unidimensional unfolding theory postulates that there is a hypothesized psychological continuum on which jointly sit the subject represented by his/her ideal point – ideal stimulus and stimuli represented by their points. Unidimensional unfolding theory recovers the metric relation among the stimulus points, yielding an ordered metric scale of the stimuli (Hays & Bennett, 1961). Each subject orders the stimuli from most to least preferred based on the increasing absolute distances of the stimulus points from him/her (Bennett & Hays, 1960; Coombs, 1964; Hays & Bennett, 1961; McElwain & Keats, 1961). This joint psychological continuum is called the *J*-scale (which displays the order of the stimuli) and the preference ordering of the stimuli obtained by folding the *J*-scale at the subject’s ideal point is called the resulting *I*-scale (Coombs, 1964, p.80). For every *J*-scale, there are \( \binom{m}{2} + 1 \) permissible *I*-scales \( (I_1, I_2, \ldots, I_{\binom{m}{2} + 1}) \), \( m \) is the number of stimuli). For a given *J*-scale, the first two stimuli in each permissible ordering must be adjacent scale points (Crocker & Algina, 1986). For example, for a *J*-scale of “ABCD”, the 7 permissible *I*-scales are *ABCD*, *BACD*, *BCAD*, *BCDA*, *CBDA*, *CDBA*, and *DCBA*. The first two stimuli (which are underlined) in each *I*-scale are adjacent points on the *J*-scale.

Coombs (1964) distinguished qualitative *J*-scales and quantitative *J*-scales. The *J*-scale recovered at the ordinal scale level is a qualitative *J*-scale and quantitative *J*-scales have at least an ordered metric scale. Similarly, there are the difference between qualitative and quantitative *I*-scales. A *J*-scale of \( m \) stimuli is divided into \( \binom{m}{2} + 1 \) interval regions by the \( \binom{m}{2} \) midpoints between pairs of stimuli. These intervals define qualitative
permissible I scales. For example, each J-scale of four stimuli produces seven qualitative I scales. I scales may tell us: (1) whether there is a latent trait underlying the rankings, (2) what is the order of the stimuli on the latent trait, (3) the relative magnitudes of the distances between stimuli, (4) in which interval regions a subject sits and the order of the regions, and (5) relative magnitudes of these interval regions (Coombs, 1950).

The unique latent trait underlying a qualitative J-scale can be inferred by the set of I scales generated from the J-scale. However, a given qualitative J-scale does not mean a unique set of I scales. For example, a given J-scale of four stimuli A, B, C, and D can produce two sets of I scales differing on the 4th I scale depending on the magnitude of the distance between A and B compared with that between C and D (see Table 2). When the distance between A and B is greater, I4 is BCDA as in Set 1; if the distance between C and D is greater, I4 is CBAD as in Set 1. Theoretically, four stimuli on a hypothesized J-scale yield six midpoints that allow for maximal of seven I scales. Empirically, from the seven I scales the six midpoints can be ordered on the potential J-scale.

Table 2

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 2</td>
<td>ABCD</td>
<td>BACD</td>
<td>BCAD</td>
<td><strong>CBAD</strong></td>
<td>CBDA</td>
<td>CDBA</td>
<td>DCBA</td>
</tr>
</tbody>
</table>

In going down through I1: ABCD, I2: BACD, I3: BCAD, I4: BCDA, I5: CBDA, I6: CDBA and I7: DCBA, the pair of adjacent stimuli that has changed places from one I scale to the next I scale is the boundary between these two I scales on the J-scale. That is, from I1: ABCD to I2: BACD, the boundary is AB, the midpoint between A and B; from I2: BACD to I3: BCAD, the midpoint passed is between A and C; and in the same procedure
we get all the six midpoints in the order of $AB$, $AC$, $AD$, $BC$, $BD$, and $CD$. The six such boundaries section the $J$-scale into seven interval regions. Because midpoint $AD$ precedes $BC$ we know that the distance between $A$ and $B$ is greater than that between $C$ and $D$. If we observed a prevalent ordering of $BCDA$, we infer that the distance between $A$ and $B$ is greater than the distance between $C$ and $D$, because if $|A-B|<|C-D|$, the middle region should be $CBAD$. Other than this, all the other metric magnitudes on this $J$-scale are arbitrary. This is the only deduction about the distances between the stimuli that can be made for four stimuli.

Thus for four stimuli only two differentiable quantitative $J$ scales can be found for a given qualitative $J$-scale. But if there are five stimuli, a given qualitative $J$-scale can generate 12 quantitative $J$-scales, that is, 12 possible different sets of distinct permissible $J$ scales each of which would imply a different set of quantitative relations among the distances between stimuli on that $J$-scale. There will then be sufficient information to deduct the relative distances among all the stimuli. The resulted metric is between ordinal and interval levels because the distances between the units are not precise but the relative size of these distances is known (Coombs, 1950).

The determination of unidimensionality depends on whether the observed rankings of the stimuli conform to certain criteria (Coombs, 1964). A necessary condition for a common quantitative $J$-scale is that there is no intransitivity in the data. If examination of the metric information about the interval regions finds intransitivity, the rankings are not unidimensional.

Another criterion is the existence of a mirror-image reversal between the first and last preference orders. A dimension is not valid unless both ends of the dimension are
shown to exist. For instance, if a subject chooses the preference order “ABCD”, then another subject must choose the order “DCBA” to support for the existence of a unidimensional trait represented by J-scale of “ABCD”. If it is observed that a set of I scales generated from a J-scale has more than one pair of such mirror image I scales, it is sufficient to reject the hypothesis of unidimensionality. Consequently these two I scales immediately define the ordinal relations of the stimuli on the J-scale. Suppose four stimuli A, B, C, and D were ranked and I scales “ABCD” and “DCBA” were observed. Immediately these observations define the relations of the stimuli as “A>B >C>D” or “D<C <B< A”.

Another rather stringent ad hoc rule is the 50% plus 1, which states if the observed rankings that are permissible rankings for a J scale (a dimension) are one more than 50% of the total observations, then that J scale/dimension is supported. Suppose there are five stimuli, A, B, C, D, and E and 120 possible preference orderings. Five stimuli allow for the formation of 11 preference orders that are consistent with an underlying dimension of ABCDE-EDCBA. These 11 orders are ABCDE, BACDE, BCDAE, BCDAE, CDBAE, CDBEA, CDEBA, DCEBA, DECBA, and EDCBA. That is, 50% plus 1 of all individuals' preference orders must be consistent with the 11 preference orders listed herein.

**Number of Traits an Item Can Measure**

The analytical problem of unidimensional unfolding is how to unfold the observed I-scales to recover or discover the latent J-scale (Coombs, 1964). How many distinct unidimensional scales/continuum can a set of stimuli maximally delineate? It depends on how many distinct quantitative J-scales a set of m stimuli can maximally
produce. Coombs (1964) noted that the number of distinct qualitative and quantitative $J$ scales is not likely by chance. Recall that each qualitative $J$-scale is associated with a number of quantitative $J$-scales. $J$-scales for $m$ stimuli the number of distinct qualitative $J$-scales is $m!/2$ and the upper limit to the number of distinct quantitative $J$-scales for each qualitative $J$ scale is $\{[n(n-1)/2]!(n-2)!(n-3)!...2!]/[(2n-3)!(2n-5)!...3!]\}$ (Coombs, 1964).

For example, for four stimuli, there are 12 distinct qualitative and maximum of two distinct quantitative $J$-scales for each qualitative one. This makes a total of 24 maximum distinct quantitative $J$-scales. For five stimuli, there are 60 distinct qualitative and maximum of 12 distinct quantitative $J$-scales for each qualitative one, and thus making a total of 720 maximum distinct quantitative $J$-scales (Coombs, 1964, p.91). This study only considered the qualitative $J$-scale as it is easier to understand and more popularly accepted than the quantitative one (Coombs, 1964).

**Specification of the Dimension**

The next task is to specify the dimension: what is indicated by the ordinal and metric relations between the subjects and the stimuli? Because only the preference ranks of the stimuli (a set of $I$ scales) are observable, the existence of a latent dimension is inferred through unfolding the observed orderings of the stimuli ($I$ scales). The first step is naturally to examine the ordering of the stimuli. For example, if a subject’s preference ranking of a selection of cakes was unfolded and all the chocolate-ingredient cakes were at one end and all the non-chocolate at the other, we can infer this subject used a criterion

---

4 This is upper limit because it allows intransitivity metrics which may occur for $m \geq 6$. 
of “chocolatedness.” The next step recommended by Coombs (1964) is to conduct an independent experiment with other criteria to validate one's interpretation. The current study took this recommendation and used criterion measures to validate latent traits in the ranking data.

Schiffman, Reynolds, and Young (1981), in their book about multidimensional scaling, also discussed an ideal point model as one of the models for preference and property (attribute) analysis. Their ideal point was different from Coombs’ ideal point (1964) because it represents the optimal combination of stimulus characteristics for the attribute in question and the ideal point model should be used when the stimulus set does contain stimuli that have either too much or too little of at least one attribute. The ideal point model is used to find a point in a stimulus space that is most like an attribute, “It is the hypothetical stimulus which, if it existed, would contain the maximum amount of the attribute.” (p. 259). Whereas Coombs’ ideal point refers to a stimulus that is most preferred by a subject and each subject has his/her own ideal point, the ideal point model looks for a stimulus for a given attribute for all subjects.

**Multidimensional Unfolding Theory**

Multidimensional unfolding extends the unidimensional unfolding technique to include multidimensional preferences, e.g., J-scale in unidimensional unfolding extends to joint space in multidimensional unfolding. Multidimensional unfolding mainly addresses the issue of dimensionality determination and the determination of the configuration of the joint space (Bennett & Hays, 1960; Hays & Bennett, 1961). Multidimensional unfolding assumes that (1) the subjects and the stimuli are mapped into points in a common space; (2) a subject’s preference ordering between any two stimuli
reflects which stimulus point is closer to his/her ideal point; and (3) the distance in the space is Euclidean (Bennett & Hays, 1960; Hays & Bennett, 1961). In unidimensional unfolding theory, all stimulus points and any subject’s ideal point aligns along one line (the J-scale) in a space of one dimension. On the other hand, in multidimensional unfolding, any stimulus point or any subject’s ideal point is defined in a joint space composed of multiple J-scales. In general, the multidimensional theory conceives of the stimuli as having a fixed configuration in a space of $j$ dimensions and states that each subject’s preference ordering reflects the order of increasing Euclidean distance of the stimulus points from his/her ideal point. This is exactly how the preference data simulation in this Monte Carlo study was generated. (When traits covaried, Mahalanobis’s distance was used. When traits are orthogonal, this is equivalent to Euclidian distance).

Coombs (1964) noted there might be substantial variability in the widths of $I$ scale intervals. Information about the widths of the $I$ scale intervals on the $J$-scale can be estimated based on information about distances between stimuli. In other words, multidimensional unfolding theory allows for an interval estimate of the attribute for a subject in a relative space. This provides guidance for this study when inferring a judge’s multiple traits from his/her preference orderings in a joint space.

**Connection with Multiple Factor Analysis**

Multiple factor analysis was revealed as a substitute for multidimensional unfolding for the discovery of the latent attributes underlying preferences (Coombs & Kao, 1960). They defined the space obtained by unfolding as the genotypic space and the space obtained by factor analysis as the factor space. They suggested that factor analysis was particularly useful whenever (1) people wish to avoid the labor of multidimensional
unfolding or when (2) the assumptions of unfolding (e.g. every judges agrees on which attributes to rank the alternatives or that preferences are monotone in the distance, etc.) are objectionable or when (3) the data are too sparse (when rankings aggregate and only a small number of mirror image pairs can be found - recall each mirror image pair is one potential axis for describing the configuration of a multidimensional space) for multidimensional unfolding to yield a determinate solution. Their study also illustrated the techniques to transform ordinal data to numerical scale in a multidimensional space. Euclidean distances between individuals and random stimuli were computed and then the correlation matrix of individuals over stimuli was factored by the principal components method.

There are a few practical considerations with Coombs and Kao’s method. The first is that the basic data normally consist of rankings rather than the actual distances to stimulus points. This means that the product moment correlation can only be approximated. The second is that the distribution of stimulus points relative to the distribution for the individuals can distort the factor space. If the density of the stimulus points between two individuals is unusually high or low, the correlation between their preferences will be biased toward negative or positive correlation and they will appear farther apart or nearer together in the factor space than in the genotypic space. The third is that in any practical application, Coombs and Kao’s method arrives first at the factor space and seeks the genotypic space. This requires rotating the extra dimension in the factor space in order to work with just the genotypic space that remains. The problem is to locate the median individual in the space. Several solutions about this problem were described in Coombs (1964).
Theories on Preference Constructions

Delbeke (1968) developed a model that described individual differences in preference by constructing a multidimensional preference space. Davison (1992) displayed how to locate objects and subjects in a multidimensional space in terms of “ideal point” analysis. His solution is a model that gives a systematic description of the subject differences within a group to find those subjects who are similar to each other with respect to how they weigh each of these stimulus attributes. Coombs’s unfolding model (1964) is one of the techniques developed to construct a preference space that allow subjects weigh stimulus attributes differently.

The data generation in this study mostly followed the multidimensional unfolding theory (Bennett & Hays, 1960; Coombs & Kao, 1960; Coombs, 1964; Delbeke, 1968; Hays & Bennett, 1961; Kyngdon, 2006; Tucker, 1960). This theory assumes that both subjects and stimulus alternatives situate in the same state space (joint space) and a subject prefers a list of stimuli according to the increasing distances between him/her (as represented by his/her ideal stimulus \(i\)) and these stimuli. The stimulus with the shortest distance is mostly preferred. The multidimensional unfolding theory also assumes that the trait level of the ideal stimulus for a subject is regarded as the true trait score for that subject. The dimensions in such joint space are reflected by the attributes of the alternatives that subjects consider for their preference decisions. Such a multidimensional space, if constructed, will permit one to describe groups of subjects in terms of the alternatives they prefer and the attributes of the alternatives that determined preference. This is the method used in this Monte Carlo study to generate subject ranking patterns,
given randomly generated criterion scores and items weights as points located in a common trait-state space.

Subjects’ preference rankings of the alternatives for each item resulted from the interaction of the previous two data sets – the distance between a subject and each of the alternatives he/she was to rank. These distances were calculated using Euclidean distance for orthogonal traits and Mahalanobis distance for correlated traits (Stevens, 2002). These distances were used to sort the alternatives and to assign a rank based on “closeness” to the subject: i.e., the alternative that is closest to the subject is the first rank, 1, the alternative second closest is the second rank, 2, and so on.

Slovic (1995) illustrates the meaning of preference and the status of value by this well-known exchange among three baseball umpires. First one, "I call them as I see them" – values exist-like body temperature and people perceive and report them as best they can, possibly with bias. Second one, “I call them as they are” – people know their values and preferences directly - as they know the multiplication table. Third one, “They ain’t nothing till I call them” – values or preferences are commonly constructed in the process of elicitation. The research reviewed in this study is most compatible with the second view of preference and the nature of human values with the principle of procedure invariance that is fundamental to theories of rational choice.

Roskam (1981) distinguished between inferential choice and preferential choice: preferential choice refers to choice-decisions in terms of like/dislike or approach/avoidance; inferential choice refers to choice-decisions in terms of true/false responses as to “what is the case and what is not the case.” He also distinguished two forms of multidimensional scaling: smallest space analysis (SSA) to represent similarities
among behavioral variables or instances, and multidimensional scaling (MDS) in the strict sense, which implies a theory of behavior. SSA is primarily concerned with the mapping of relations among behavior variables whereas MDS is a measurement model and expresses a theory about the data-generating process. MDS is a means of disclosing cognitive dimensions and the measurement of objects with respect to those cognitive dimensions. It may tell which perceptual, cognitive or evaluative dimensions operate in the subject’s mind.

Comparing a series of objects, people generally first specify on what quantitative or qualitative attribute they are to compare the objects (Thurstone, 1927). Thurstone called this specification process the discriminial process. It implies the assumption that a subject aggregates objects information into a unidimensional attribute and assigns different amounts of this attribute to each object. This attribute defines the psychological continuum or scale and the compared objects are allocated on the continuum. The position of an object on that dimension is characterized by a stochastic variable. According to Thurstone (1927), as people compare two or more objects there is some kind of process in them by which they react differently to the several objects and identify different amount of characteristic under consideration. This is defined as discriminial process. For example, if one drawing sample seems to be better than a second one to an observer, then the two discriminial processes of the observer are different. He noted that the discriminial process corresponding to a given object was not fixed and that he/she might make different judgments on successive occasions about the same pair of objects. On the other hand, with a given object there might be one discriminial process that is
experienced more often than other processes. This most common process is called *the modal discriminal process for the given object* (Thurstone, 1927).

Thurstonian law of comparative judgment defines a psychological scale or continuum that allocates the compared objects on the continuum. It states that the experimentally observed proportion of judgments A is stronger (better, lighter, more excellent) than B is a function of the scale values of the objects, their respective discriminal dispersions, and the correlation between the paired discriminal deviations. The psychological continuum or scale is defined in terms of the frequencies of the respective discriminal processes for any given objects so that the frequencies of the respective discriminal process yield a normal distribution on the psychological scale.

The idea of utility, which is defined as the power to satisfy human wants by economists, proved to have general appeal to decision theorists (e.g., Miller & Starr, 1967). The objective of the individual is held to be the maximization of the total utility he/she can achieve with his limited sources of time, effort, and money. The choice that yields the greatest amount of utility will be preferred the most (Hogarth, 1987). However, people in reality seldom act in this rational way due to many reasons: e.g. the inability of the individual to duplicate the rather recondite mathematics which economists have used to solve the problem of maximization of utility, the effect of habits, the existence of other values, the influence of social emulations, the effect of social institutions (Miller & Starr,

---

5 The standard deviation of the distribution of discriminal processes on the scale for a particular specimen will be called its *discriminal dispersion* (Thurstone, 1927, p. 275).
6 The separation on the scale between the discriminal process for a given stimulus on any particular occasion and the modal discriminal process for that stimulus we shall call *the discriminal deviation* on that occasion (Thurstone, 1927, p. 275).
Economists, as well as psychologists and sociologists, have been trying to incorporate these aspects of behavior into choice or decision situations.

The measurement of individuals’ preference or utility functions for a specified stimulus has been the focus of decades of research in psychology and behavioral decision theory (Torgerson, 1958). It is desirable to achieve an interval scale measurement of utility whenever possible because then expected values can be calculated (Hogarth, 1987). In situations where a single dimension is involved, ordinal measurement can be obtained to measure utility of alternatives. When utility is used to measure the degree of satisfaction obtained, the problem is that there is no convenient measuring unit when it comes to describe utility of an intangible thing such as dignity. This difficulty of measuring the utilities prevents the theory from satisfactorily explaining observed behavior and decisions (Miller & Starr, 1967). Miller and Starr proposed that if the degree of achievement of the objective can be stated in quantitative terms, then alternative choices can be compared with each other. The first major problem facing a decision maker is how to precisely formulate his/her objective and specify its dimensions and values. However, even in terms of this objective, there still is a similar problem when measuring constructs such as good relations or job security. Even for objectives with a natural measure of degree of achievement, it is still necessary to recognize that the natural measure may not coincide with the utility the decision maker receives from the degree of achievement of his/her objective, e.g., the utility of dollars is not necessarily the same as the number of dollars. If it does not, it is the utility that governs the decision problem, not the natural measure. A variety of methods have been developed to measure the utility of a dollar, such as logarithmic representation of utility, the reciprocal of the
number of dollars, the square root of the number of dollars, etc., as the most satisfactory measure of utility. For example, Neumann and Morgenstern (1947) demonstrated a way, known as the standard-gamble method, to achieve an interval scale for the measurement of utility where utility is defined as the indifference value of probability.

**Multidimensional Scaling (MDS)**

Any particular model designed to recover the multidimensional space from the preference ranking data needs to specify the multidimensional scaling used in the model: i.e., how it measures the distance between pair of points in the space (Coombs, 1964). Multidimensional scaling (MDS) is the basis for non-keyed preference data modeling and analysis. The oldest method of multidimensional scaling is based on the work of Young and Householder (1938). Since then MDS has experienced development of two phases: metric MDS and nonmetric MDS. Important works of metric multidimensional scaling were accomplished by Messick and Abelson (1956), Torgerson (1952, 1958), Tucker and Messick (1963), Torgerson (1958), and Young and Hamer (1987) who first attempted to incorporate individual differences into multidimensional scaling procedure; while nonmetric multidimensional scaling method is significantly contributed to by Shepard (1962a, b), Kruskal (1964a, b), Carroll (1967, 1972), DeLeeuw (1981), Lingoes (1967), Young (1987), and others. Coombs’ unfolding models (1964) had influenced some aspects of the nonmetric phase of multidimensional scaling as a basis for the newer methods.

Torgerson’s model is a parametric model, assuming a normal distribution of the discriminant processes associated with the distance between two stimuli, and the mean value of this distribution is used as the scale value of the psychological distance and the
discriminal dispersion serves as the unit of the scale (cf. Thurstone’s law of comparative judgment). Therefore, before applying this space model, one needs to be sure that the distances satisfy the conditions for obtaining a metric space. The technique for transforming the scale with comparative distances into a scale with absolute distances is available (Messick & Abelson, 1956). Torgerson’s MD model permits one to find the perceptual structure for “the average subject.” For each pairwise comparison this average subject takes the average of the similarities estimated by the whole group of subjects. These average distances are put in the interstimulus distance matrix on which the MD analysis is performed.

The first nonparametric MD model was developed by Shepard (1962) to estimate parameters for the stimuli in a MD space from nonparametric data. Specifically, the data are proximities: i.e., any type of indices which indicate the similarity or dissimilarity between two stimuli, e.g., probabilities that a stimulus is distinguished from another one. The proximity measures themselves are not important for the model, but their rank order is. This model is able to determine the \( m \cdot r \) parameters (the projections of \( m \) stimuli on \( r \) dimensions) from the rank order of the \( m(m-1)/2 \) proximities.

A variety of methods have been developed to scale the mutual distances between pairs of points in a multidimensional space. Some of them come along with certain models for choice decision process, e.g., the data unfolding theory (Coombs, 1964), and the personal compensatory model (Tucker, 1960); some of them have not been seriously considered in the context of psychological theory, but mainly address mathematical expressions, e.g., the Landahl’s city-block model (Householder & Landahl, 1945; Landahl, 1945; Torgerson, 1958). The Euclidean distance function is almost universally
used in multidimensional scaling theory (Coombs, 1964). The multidimensional Euclidean distance between any pair of points \( x_1 \) and \( x_2 \) is the square root of

\[
\sum_{d=1}^{d} (x_i^d - x_2^d)^2
\]

for \( d \) dimensions.

The Hays-Bonnett multidimensional unfolding (1960, 1961) calculates the Euclidean distance between a stimulus point and a person’s ideal point (Hays & Bonnett, 1960; Coombs, 1964). This model can be regarded as an instance of a jointly compensatory model, where the individual and the stimulus together determine the rate of exchange between the dimensions.

The personal compensatory model (Tucker, 1960; Hogarth, 1987) and the lexicographic model (Coombs, 1964) also accounts for individual differences in preference ranking. Personal compensatory model states that each individual’s ideal point (X) and the origin define a line on which all the stimulus points (A, B, C and D) are projected (Figure 2). An individual’s preference ordering is generated by folding this line at his/her ideal point. For this person, the ranking goes DBCA. The psychological idea behind this model is that the primitive attributes interact in such a way that they compensate for each other at some rate that characterizes an individual. The ideal point represents a weighted sum, and individuals are indifferent to the relative weights of the alternatives. Tucker regards individuals and stimuli as vectors and defines the scale value, the utility of the stimulus for a specific individual as the vector product of a stimulus vector with the individual’s vector. This indicates that an individual’s utility or preference for any alternative increases monotonically with the alternative’s loading on any attribute. The psychological idea behind is that whatever the primitive attributes may be, an increase in any one of them increases the desirability of the stimulus.
The lexicographic model assumes that a hierarchy of importance exists on the primitive attributes. That is, alternative higher on the most important attribute will be preferred, irrespective of the relative positions of it on the other dimensions. If alternatives are tied on the most important dimension, a judge will turn to the second most important attribute, and so on. However, this model is untenable in its general form because no two alternatives are equal. It will work only if small differences are disregarded, or if the stimuli are partitioned into equivalence classes on each continuum (Hutt, 1954; Coombs, 1964). Although ties do occur in the Monte Carlo simulation and if the number of decimal places is not sufficient, then they are forced-resolved by chance alone.

Landahl’s city-block model is the consequence of a neurological mechanism designed to mediate judgments of similarity and difference. In the Landahl’s city-block model the distance between two stimuli is the simple arithmetic sum of their difference on the individual dimensions. According to this model the distance between an individual’s ideal point and the stimulus is
\[ |P_{hij}| = \sum_d \left| c_{hij}^{(d)} - q_{hij}^{(d)} \right| \]

where \( P_{hij} \) indicates at the moment \( h \), the distance of the point corresponding to alternative \( j \) and the ideal point of individual \( i \). \( c_{hij}^{(d)} \) is the projection of the vector \( c_i \) in the set of relevant dimension \( D \). \( q_{hij}^{(d)} \) is the projection of the vector \( q_j \) in the set of relevant dimensions \( D \). The individual’s preference between two stimuli is assumed to be mediated by a comparison of two distances, the distance of each of the stimuli from the individual’s ideal. The distance function is illustrated in Figure 3 with one individual (\( X \)) and three stimuli (\( A, B, C \)). The length of the dashed line from a stimulus to \( X \) is its distance from the ideal point.

Figure 3. City block model.

**Simulating Preference by Multidimensional Scaling**

The most popular theories of preferences in gambling propose that an individual maximizes some kind of expectancy such as a subjectively expected utility when he/she is offered a pair of gambles (see Edwards, 1955; Davidson, Suppes, and Siegel, 1957). The unfolding theory and all alternative theories considered so far has to do with first-choice preferences. Each successive choice is presumed to represent a first choice in the
absence of the previously chosen alternatives. However, this condition can be contradicted in reality. It is realistic to presume that a decision process is phased, with the first phase involving one model and a later phase a different model. This is where the multiattribute decision making models fit in.

**Simulation Preference Rankings by Multi-Attribute Decision Making**

Multi-Attributes Decision Making (MADM) is the study of methods and procedures by which concerns about multiple conflicting criteria can be formally incorporated into the management planning process, as defined by the International Society on Multiple Criteria Decision Making (website of the International Society on Multiple Criterion Decision Making). MADM is a decision-making procedure that aids making preference decisions over the alternatives that are characterized by multiple attributes (Hwang & Yoon, 1981). Yoon and Hwang (1995) have identified 13 MADM methods, such as non-compensatory methods that do not allow a trade-off of attributes, and Analytic Hierarchy Process (AHP) that will help the decision maker set up a hierarchy of attributes. Pardalos, Pitsoulis, and Resende (1995) classified the MADM methods into four categories: (1) multi-objective mathematical programming (MMP); (2) outranking relations approach; (3) multi-attribute utility theory (MAUT); and (4) preference disaggregation approach (PDA). While the MAUT represents/model the decision maker's preferences through a utility function $U_j$ (where $j$ denotes alternative option) aggregating all the evaluation criteria or attributes, the PDA disaggregates the global preferences of the decision maker in order to identify the criteria aggregation model that underlies the preference result (ranking or classification/sorting). The simplest way of employing the MAUT was done by Edwards and Newman (1982), and they called
it “multiattribute utility technology (MUT).” This study adopts the MUT method to simulate the preference data.

Multiattribute utility technology was used by Edwards and Newman as an approach for descriptive decision making applied in program evaluation that often requires dealing with multiple measures of effectiveness. With MUT, the value dimensions or attributes \( a \) will be elicited from the judges and are organized into a hierarchical structure called *value tree*. The relative importance of each of the attributes, or weights, \( (W_{ap}) \) will be assessed for each person or subgroup of people \( p \). Each alternative option \( (j) \) will be assessed by judge \( p \) on how well it serves each attribute at the lowest level of the value tree, and the numbers is called single-attribute utilities or location measures \( (u_{ajp}) \). Then these single-attribute utilities will be aggregated with measures of importance, weights. This aggregation produces overall or composite utility \( (U_{jp}) \) for each alternative \( j \) for person \( p \). The equation for aggregation goes as:

\[
U_{jp} = \sum W_{ap} u_{ajp}
\]

where \( U_{jp} \) is the overall composite utility for the \( j^{th} \) alternative rated by person \( p \); \( W_{ap} \) is the normalized weight assigned to the \( a^{th} \) attribute by person \( p \); and \( u_{ajp} \) is the utility of the \( j^{th} \) option on the \( a^{th} \) attribute for person \( p \). The larger the value of overall utility \( U_{jp} \) of an alternative, the better is the chance for top rank. For example, the utility of alternative \( J \) perceived by person 1 is the sum of the product of the relative importance of each of the attributes (e.g., \( a = 1, 2, 3 \)) involved in alternative A and the single attribute utilities of alternative A as perceived by person 1. It will go as:

\[
U_{J1} = W_{11}u_{1J1} + W_{21}u_{2J1} + W_{31}u_{3J1}
\]
The overall utility for each alternative will be obtained in the same way for one person.
And the alternative with the highest utility score is supposed to be preferred the most, and
alternative with the lowest utility score is preferred the least.

In the MUT model, the ideal measures of single-attribute utility is provided by experts so that they are independent of individual judges and so is independent of the value disagreements among judges. Things are little different in the proposed study in which the location measures are supposed to have moderately positive or negative correlation with the weights with the assumption that people do not randomly weigh attributes, but tend to interact with whatever attributes that they are weighing.

Weights for the attributes are assigned by individual judges and they sum to one at each level of the value tree. The final weights for each attribute at each twig of the tree are obtained by multiplying the normalized weights of this twig by the normalized weights of branches of this twig. In Edwards and Newman's study, individual difference is something that needs to be intentionally removed, e.g., by averaging the weights. However, this study is interested in the individual difference. Finding out how an individual judge weighs an unobservable attributes without being asked to provide his or her weighting is a core problem of this study. This information has to be inferred from the aggregated information a judge has provided, which is the preference rankings. To some degree, the ranking corresponds to the overall weighted values, $U_{jp}$, in the MUT model. There is no available data corresponding to the location measures or utilities in Edwards and Newman’s study, because this is exactly the information that this study intends to obtain.
Cross-Validation

Currently, cross-validation or rotation estimation, serves as a standard procedure for estimating the generalization performance of a model from available data (Refaeilzadeh, Tang & Liu, 2009). The basic form of cross-validation is k-fold cross-validation. Other forms of cross-validation are special cases of k-fold cross-validation or involve repeated rounds of k-fold cross-validation, including:

- Resubstitution Validation
- Hold-Out Validation
- K-Fold Cross-Validation
- Leave-One-Out Cross-Validation
- Repeated K-Fold Cross-Validation

Table 3 displays a summary of these approaches based on the paper by Refaeilzadeh, Tang and Liu (2009). Kohavi compared (as cited by Refaeilzadeh, Tang & Liu, 2009) cross validation (including regular cross validation, leave-one-out cross-validation, stratified cross-validation) and bootstrap (sample with replacement), and recommended stratified 10-fold cross validation as the best model selection method, as it tends to provide less biased estimation of the accuracy. The approach of 10-fold cross-validation (k = 10) is the most common in data mining and machine learning where each training set shares 8/9 of its instances with each of the other nine training sets.
<table>
<thead>
<tr>
<th>Validation Methods</th>
<th>Description</th>
<th>Pro</th>
<th>Con</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resubstitution Validation</td>
<td>The model is learned from all the available data and then tested on the same set of data.</td>
<td>Simple</td>
<td>Over-fitting and poor generalization.</td>
</tr>
<tr>
<td>Hold-out validation</td>
<td>The available data is split into two non-overlapped parts: one for training and the other for testing.</td>
<td>Avoids the overlap between training data and test data, yielding a more accurate estimate for generalization of the model.</td>
<td>Not use all the available data and the results are highly dependent on the choice for the training/test split; Large variance.</td>
</tr>
<tr>
<td>k-fold cross validation</td>
<td>The data is first partitioned into k equally sized segments or folds. Then k iterations of training and validation are performed such that within each iteration a different fold of the data is held-out for validation while the remaining k-1 folds are used for learning.</td>
<td>Maximum utilization of the available data; Accurate performance estimation.</td>
<td>Small samples of performance estimation; Overlapped training data; Elevated type I error for comparison; Underestimate variance or overestimated degrees of freedom for comparison</td>
</tr>
<tr>
<td>Leave-One-Out cross-validation</td>
<td>In each iteration, nearly all the data except for a single observation are used for training and the model is tested on that single observation.</td>
<td>Unbiased performance estimation</td>
<td>Very large variance</td>
</tr>
<tr>
<td>Repeated k-fold cross-validation</td>
<td>A commonly used method to increase the number of estimates by running k-fold cross-validation multiple times. The data is reshuffled and re-stratified before each round.</td>
<td>Large number of performance estimates</td>
<td>Overlapped training and test data between each round; Underestimated variance or overestimated degrees of freedom for comparison</td>
</tr>
</tbody>
</table>
CHAPTER III

METHODOLOGY

This study explored a scoring method that can suit both unidimensional and multidimensional response matrices of rankings via Monte Carlo simulation. Traditionally, in a test designed to measure multiple dimensions, each factor/trait is described by a subset of items and for each item there is one scoring key, the item key. This study hypothesized that when there is a well-established criterion instrument for a trait that a SRT ranking item test intends to measure, it is possible to maximize the ranking item test’s criterion validity by altering the scoring sequence of the response. Such a trait is called the “criterion trait,” and such maximal criterion validity is called “SRT criterion validity.” Instead of the traditional practice of “one key for each item,” this study proposed “one key for each trait for each item.” Thus, each item might have multiple criterion trait-keys, each corresponding to a different criterion trait. By applying the multiple criterion trait-keys to score subjects’ preference rankings, the subjects’ item scores will be obtained. Averages or totals of those item scores are the subjects’ test scores that then become the “best” estimates of the criterion traits.

Experimental Conditions

There are four main factors (independent variables) in this study – the number of criterion traits (NTRAITS), the correlation among the criterion traits (TCORR), the number of items (NITEMS), and the subject sample size (SSIZE). A Monte Carlo simulation of a SRT (or situational preference ranking test) data was utilized to examine the effects of these conditions on test results obtained using a discovered set of trait-keys.
for each item. In simulation studies, the amount of multidimensionality in the test data is generally manipulated by either increasing or decreasing the magnitude of the correlations among the latent traits (Gessaroli & De Champlain, 2005). Lower inter-trait correlations result in more multidimensionality. In the extreme case of perfectly correlated traits (or one trait), a unidimensional model occurs. Therefore, the correlation among the criterion traits, TCORR, to be studied was set to be $r = .00, .40, \text{ and } .80$.

The NTRAITS comprising the trait-space to be examined was set to 2, 3, and 4. This decision directly influenced the number of alternatives. According to Bennett and Hays (1960, p. 38), with 3 traits, the maximum number of ranking patterns generated by 3, 4, 5 and 6 alternatives are 6, 24, 96 and 326, respectively. Another consideration was Miller’s (1956) statement that the general population has a working memory capacity of seven plus or minus two alternatives. If there are five or six alternatives, the permutations of the alternatives will be too large and lead to too heavy computation work. Meanwhile, three alternatives with six ranking patterns will be too few to be representative. Therefore, the number of alternatives to be studied here was fixed to four.

The choice of SSIZE took into consideration both reliability of the study findings and the practice in the fields. Bennett and Hays (1960) proposed a general rule about the number of $n$ participating subjects in relation to the number of $m$ stimuli: $n > m!$. Salgado (1998) reviewed 64 criterion-oriented validity studies published in 12 years between 1983 and 1994 in the Journal of Applied Psychology (JAP) and other two journals. The sample size for validity studies carried out in civil settings had a mean around 153 and median around 113 participants, with a maximum of 1,097 participants and a minimum of 25. Sample distributions do tend to be not quite normal and downwardly biased. This
study employs 1:1 calibration and validation samples. Trait-keys are determined using a calibration sample and validated using a validation sample. This study examined results obtained using sample sizes $p = 50, 250, 1,000, \text{ and } 5,000$.

In summary, for the four main factors, there are 3 levels of NTRAITS ($k = 2, 3, 4$), 3 levels of TCORR ($r = .00, .40, .80$), 5 levels of NITEMS forming a test ($i = 1, 5, 10, 15, 20$), and 4 levels of SSIZE ($p = 50, 250, 1000, 5000$). The number of alternatives in each item was not of focus in this study and it is fixed to four.

**Finding Multiple Criterion Trait-Keys for an Item through the ORORCS**

The criterion trait-key for each item will be identified through a procedure that produces the optimal rank order relationship with criterion scores (ORORCS). A prerequisite for the ORORCS procedure is having some well-established criterion instruments that subjects can take in order to obtain subjects’ scores on these known criterion traits. Subjects need to take both the ranking item test and the known criterion instruments so that the scores from these tests can be correlated with each other. Once the “best” trait-keys are identified and validated, future subjects need only take the ranking item test that is then scored using those trait-keys.

The ORORCS involves three steps in the calibration phase in this study (Figure 4). Step one involves identifying all possible keys for a ranking item. For a ranking item composed of $m$ options, there are $m!$ all possible patterns of ranking which represent all possible keys for that item to score a response. The key’s are actually a vector of rankings.
For example, for a ranking item with options of A, B, C, and D, there are $4! = 24$ all possible keys from the permutation of four options:

\[
\begin{bmatrix}
1234 \\
1243 \\
1324 \\
\vdots \\
4321
\end{bmatrix}
\]

Step two involves producing all possible item scores. All possible keys for a ranking item are used to score subjects’ responses (option ranking patterns) for this item. The result is then all possible item scores for all of the subjects. For each item, there is a vector of scores for every subject and the number of ranking item scores in this vector is the same as the number of all possible keys in step one. Thus, one observation of preference ranking has now increased to $m!$ item scores ($S_1, S_2, \ldots, S_{m!}$).

The distance between an observed ranking response and a key is calculated using the well-established Kendall's distance method (Marden, 1995) as displayed in Table 4. Kendall distance counts the number of discordant pairs of options for a pair of rankings: i.e., the number of pairs of options that appear in the opposite relative order in the two rankings. The interpretation of ranking test scores follows such a principle: the closer the Kendall distance between a subject’s preference ranking and the keyed ranking, the better this subject’s performance. A ranking of $m$ options will be decomposed into $\frac{m(m-1)}{2}$ paired comparisons, e.g., for 4 options in an item, there are 6 paired comparisons and the Kendall distance between two rankings, “1324” and “1234” is 1. This can also be interpreted as the number of ranking decisions made by the subject that differs from those made by the test designer and represented by the key.

The traditional scoring method for rankings is the converse of the Kendall distance, counting the number of agreeing pairs of options for an observed ranking and
the keyed ranking. For example, a ranking of “1324” scored against key of “1234” will get a score of 5 (= 6-1) because the Kendall distance between the two rankings is 1. This subject made 5 decisions in concordance with the key. This converse method is used in this study so that higher scores (on an item or the test as a whole) reflect more agreement and hence are expected to correlate positively with the trait of interest.

Table 4

<table>
<thead>
<tr>
<th>ACBD Scored Against ABCD Using Kendall’s Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>Keyed Relations for “ABCD”</td>
</tr>
<tr>
<td>A &gt;B</td>
</tr>
<tr>
<td>Observed Relations for “ACBD”</td>
</tr>
<tr>
<td>A &gt;B</td>
</tr>
<tr>
<td>Right Decision</td>
</tr>
</tbody>
</table>

Step three involves producing all possible correlation coefficients, \( r_{c_k s_k} \) for the sample by correlating all possible scores from the ranking item test with the subjects’ criterion scores from the known criterion instruments, one criterion at a time. For each criterion, there will be a vector of all possible correlation coefficients. For instance, if there are only six possible keys for each item (as would be the case with only three alternatives to be ranked), then for every subject, there will be six possible items scores. Hence, for all subjects, there will be six correlation coefficients for each criterion measure and one vector of such correlations for each criterion measure.

The highest (most positive) correlation coefficient in this vector for each criterion is identified and this becomes the SRT criterion validity coefficient for each trait. The key behind it is the criterion trait-key, the "best key" choice for this item and this trait. In
similar fashion, each criterion trait-key for each criterion trait, one trait at a time, is identified for every item. Of the possible $m!$ keys, there is symmetry. Half of them are opposite the other half. (Both ends of the trait continuum are identified.) So the maximum number of traits that an item of $m$ alternatives can distinguish depends on the number of alternatives that this item has, i.e., $k = m!/2$ (Bennett & Hays, 1960; Coombs, 1964). For four alternatives, the maximum number of traits is 12.

![Diagram](https://via.placeholder.com/150)

**Figure 4.** Identify trait-keys through ORORCS

**The Monte Carlo Simulation Design**

The Monte Carlo simulation in this study included 1,000 replications of a simulated experiment modelling a random SRT under each set of experimental conditions. A sample of subjects, twice the "sample size" for the experiment, along with Gaussian random criterion trait scores, was created initially. This sample was divided in half with the first half used for calibration and the second half used for validation. In the calibration
phase, each experiment simulated a set of examinee criterion trait scores and sets of item alternative trait-weights. Given these, a simulated response ranking-pattern was produced for each item in the SRT and each subject (in the entire sample). For each item, calibration subject scores were generated by scoring the calibration subject's simulated ranking pattern against each of the 24 possible keys. The resulted 24 item score vectors (of length equal to the sample size of subjects) were correlated with the vectors of the calibration subjects' criterion trait scores. A “best” criterion trait-key with the highest positive correlation was identified for each criterion trait for each SRT item. The score vectors associated with the best trait-keys for item became the subject item scores and the sum of these scores for each trait became the subjects' test scores. This completed the calibration phase of the experiment.

The second phase of each experiment was the validation phase. Here, the ranking patterns produced by the other half of the experimental sample were scored using the criterion trait-keys determined in the calibration phase. Again, these new vectors of ranking scores were correlated with their respective criterion trait scores (obtained by taking the known criterion tests). The results were a set of SRT criterion validity coefficients \( r_{C_kS_k} \) for this particular experiment. This entire process was repeated independently 1,000 times to form a complete data set modelling 1,000 random SRTs in terms of the criterion validity coefficient and its 95% confidence interval under each set of experimental conditions. This complete data set became the basis for the multivariate analyses of variance (MANOVA) described below and allowed the major study hypotheses to be statistically tested.
The independent variables (IV) in this study include 3 levels of the number of criterion traits (NTraits: \( k = 2, 3, \text{ or } 4 \)), 3 levels of correlation among the criterion traits (TCorr: not correlated, \( r = .00 \); moderately positively correlated, \( r = .40 \); or highly positively correlated, \( r = .80 \)), 4 levels of sample size of calibration/validation samples (SSize: 50, 250, 1000, or 5000), and 5 levels of test length in terms of items (NItems = 1, 5, 10, 15, or 20). The dependent variables (DV) are vectors of SRT criterion validity coefficients \( r_{Zc_kS_k} \), one for each of the \( k \) traits. These were the Fisher’s \( z \) transforms of the Pearson correlations between the subjects’ criterion scores from the criterion instruments, \( C_k \) and their scores from the ranking item test, \( S_k \). Table 5 displays a summary of the factorial design with 3 parallel stratification models based on the number of criterion traits involved, i.e., 2, 3 and 4 traits. Each stratification model contains 3x4x5 = 60 cells. When the experiment was replicated 1,000 times for each cell in the model, 1,000 observations of the DV vector for each cell were obtained. These sets of validity coefficients and the widths of their 95% confidence intervals (of back transformed to Pearson correlation) are the data for further analysis. Descriptive statistics for each cell as well as main effects and interaction effects of the three IV experimental condition factors were the analytical outcomes estimated and evaluated using MANOVA methods and \textit{a priori} planned contrasts.
### Table 5

**Summary of Factorial Design of the Study (N = 1000)**

<table>
<thead>
<tr>
<th>Sample size</th>
<th># items (i)</th>
<th>Criterion correlation</th>
<th>2-Trait Model</th>
<th>3-Trait Model</th>
<th>4-Trait Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>( k=2 \rightarrow \begin{bmatrix} r_{ZC_1S_1} \ r_{ZC_2S_2} \end{bmatrix} )</td>
<td>( k=3 \rightarrow \begin{bmatrix} r_{ZC_1S_1} \ r_{ZC_2S_2} \ r_{ZC_3S_3} \end{bmatrix} )</td>
<td>( k=4 \rightarrow \begin{bmatrix} r_{ZC_1S_1} \ r_{ZC_2S_2} \ r_{ZC_3S_3} \ r_{ZC_4S_4} \end{bmatrix} )</td>
</tr>
<tr>
<td>( p=50 )</td>
<td></td>
<td>1 5 10 15 20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( r=.00 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( r=.40 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( r=.80 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p=250 )</td>
<td></td>
<td>1 5 10 15 20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( r=.00 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( r=.40 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( r=.80 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p=1000 )</td>
<td></td>
<td>1 5 10 15 20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( r=.00 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( r=.40 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( r=.80 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p=5000 )</td>
<td></td>
<td>1 5 10 15 20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( r=.00 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( r=.40 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( r=.80 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Data Generation**

The Monte Carlo simulation (MC) included 1,000 replications of a simulated experiment that contained a data generation section, a calibration section and a validation section. All the data was generated and analyzed using SAS 9.4 (English). A series of loops were set up to automate this procedure and loop through all of the combinations of run conditions. These steps resulted in SRT criterion validity coefficients (i.e., a correlation between the criterion trait scores and criterion scores from the ranking test obtained using the validation sample). The upper and lower CI bounds were calculated separately during the post-processing phase when the final analytical database was created. That is also where the Pearson correlations were transformed into Fisher zs.
Going into more detail, the Monte Carlo process generated a population of subjects from which a calibration sample and a validation sample were drawn based on a 1:1 ratio. This method of cross-validation is referred to as hold-out validation in the literature (Refaelzadeh, Tang & Liu, 2009). For each member in the population 2, 3 or 4 criterion scores \( C_k \) were simulated with a normal distribution \( (M = 50, SD = 10) \). These scores represent the coordinates of a subject point in the trait-state space. Item alternatives, four for each item and each also representing a point (with 2, 3 or 4 criterion weights \( W_k \) ) in the trait-state space, were generated by selecting randomly from a normal distribution \( (M = 50, SD = 10) \). The distance between each subject and the 4 alternatives in an item were calculated using the Mahalanobis distance (Coombs, 1964; Davidson, 1992; Delbeke, 1968; Edwards & Newman, 1982; Hogarth, 1987; Hwang & Yoon, 1981; Pardalos, Pitsoulis, & Resende, 1995; Tucker, 1960; Yoon & Hwang, 1995). These distances were then sorted and the closest alternative was ranked as the first choice this subject would make, and so on. When the trait space is orthogonal, the Mahalanobis distance is the same as the Euclidian distance. When the traits are correlated, the Mahalanobis distance method takes care of the correlation (Stevens, 2002; Wicklin, 2013). The same set of items and alternatives (i.e., the same test) were used for both calibration and validation phases, albeit with different random samples of subjects and their rankings. This allowed to examine whether the best trait-keys identified by the calibration sample would generalize across different groups of subjects.

This study used random alternatives for better generalization. The example displayed in Table 6 considered a ranking item where three criterion traits being modeled. It should be noted that if all alternatives were weighted the same on all traits, a subject's
preference ranking would contain no information (pure noise) as far as these traits are concerned. The best strategy might be attempting deliberately to capture as much variability in alternative characteristics as possible when selecting the 4 alternatives. That is instrument construction question that will be taken care of by instrument writers.

Table 6

<table>
<thead>
<tr>
<th>Alternative</th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$C_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 or A</td>
<td>60</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>2 or B</td>
<td>45</td>
<td>60</td>
<td>45</td>
</tr>
<tr>
<td>3 or C</td>
<td>45</td>
<td>45</td>
<td>60</td>
</tr>
<tr>
<td>4 or D</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

It is known that averaging correlations leads to underestimation because the sampling distribution of the correlation coefficient is skewed. However, if correlations are transformed by Fisher's $z$ prior to averaging, average $z$ backtransformed to $r$ is less biased positively than average $r$ is biased negatively (Silver & Dunlap, 1987). Therefore, Pearson product-moment correlation coefficient $r_{CkS_k}$ was transformed to $z_{CkS_k}$, the Fisher's $z$ in order to meet the normality assumption of MANOVA (Stevens, 2002). No "bias adjustment" was used because our sample sizes could be considered sufficiently large to avoid the positive skewedness (overestimation) of the Fisher's $z$ estimation (Silver & Dunlap, 1987).

**Data Analysis**

After the data for the 180 cells was generated, a complete database for analysis was constructed with the independent variables (experimental conditions) and the dependent variables: Fisher $z$ transformations (with 95% CIs for each replication) during a "post-processing" phase. Separate MANOVA's were examined for each of the three
trait-model stratifications (i.e., 2-Trait, 3-Trait, and 4-Trait) as well as follow-up univariate ANOVA for the planned contrasts. Each model was identical and tested main effects and interaction effects for the SRT criterion validity coefficient and for the 95% CI widths, as well as planned contrasts for quadratic effects. In addition, a series of "spaghetti plots" were produced to visually illustrate the interaction effects in the model. This study reports in detail on results obtained using the 3-trait model, while only summarizing the similar results observed for the other two models.

**Analysis of Precision of Estimation of SRT Criterion Validity**

One thousand replications of the experiment with random subjects and random tests provided DV vectors of length = 1,000 observations of the SRT criterion validity coefficient \( r_{Z_{Ckh}} \) and its confidence interval (CI), e.g., 95% CI \( z = z_r \pm 1.96/\sqrt{n - 3} \) where "\( n \)" refers to the sample size), under each set of experimental conditions. Descriptive statistics were calculated for each cell, including the means of \( r_{Z_{Ckh}} \) and the width of the CIs. In the end, the Fisher \( z \) values were back-transformed into Pearson correlation values for the visual plots.

The research hypothesis is that, as SSIZE goes up, the CI width will become narrower. Likewise, as NITEMS increases, the SRT criterion validity \( r_{Z_{Ckh}} \) (together with their CI lines) will increase and move to the right toward some upper limit. Furthermore, that upper limit is going to be an inverse function of NTRAITS and TCORR. Both NTRAITS and TCORR should diminish the SRT criterion validity coefficient \( r_{Z_{Ckh}} \) and increase its CI width. This curvilinear expectation was tested using a priori planned NITEMS quadratic function contrasts. If the model behaves as expected with random subjects and tests, then it is reasonable to conclude that the SRT scoring
method works effectively in general for well-designed and constructed SRT testing instruments or on-line testing situations.

**Analysis of IVs Effect on SRT Criterion Validity**

The main effects and the interaction effects of test conditions on test performance were examined using MANOVA to determine if the DV vectors are influenced by the manipulation of the four independent variables (IVs) – subject sample size (SSIZE), the number of ranking items (NITEMS), the correlation among the criterion traits (TCORR) and the number of criterion traits (NTRAITS). This analysis included the main effects and the interaction effects. A significant main effect of each of the IVs would indicate that the accuracy of the estimation of SRT criterion validity coefficient would differ under different IVs conditions. Significant interaction effects would indicate that one IV influences the DV differently at different conditions of the other IVs.

**Summary of Research Methodology**

This Monte Carlo study proposed and tested an idea to improve ranking item tests’ (e.g., SRTs) criterion validity by finding the best criterion trait-key for each item. The SRT ranking items (which are defined in terms of four alternatives loading on 2, 3, or 4 criterion traits) and the subjects (which are defined in terms of their 2, 3, or 4 criterion trait-scores corresponding to the criterion traits of the alternatives) were simulated. These values were simulated randomly from normal T-score distributions \( M = 50, SD = 10 \). Subjects’ SRT rankings responses resulted from calculating the Mahalanobis distance between each subject and the 4 alternatives. These distances were then sorted and the closest alternative was ranked as the first choice this subject would make, and so on. The same set of items and alternatives (i.e., the same test) was used for both calibration and
validation phases. Subjects were split randomly (50%) into a calibration group and a validation group. The calibration group identified the criterion trait-keys through the ORORCS procedure. These keys were then used to score the validation group to produce the validation subjects’ SRT criterion trait-scores. These were correlated with the validation subjects’ simulated criterion trait-scores using Pearson product-moment correlation. The Pearson correlation coefficients produced by the validation group are defined as the SRT criterion validity coefficients and are the DVs of this study. Correlation coefficients ($r_{c_kS_k}$) were transformed with the Fisher $z$, ($r_{z_{c_kS_k}}$) for the MANOVA study and back-transformed for the CI study.

The experimental conditions in this study included: the number of criterion traits (NTRAITS = 2, 3, or 4), the correlation among the criterion traits (TCORR = .00, .40, or .80), the number of test items (NITEMS = 1, 5, 10, 15, or 20), and the sample size of subject for both the calibration and the validation samples (SSIZE = 50, 250, 1000, or 5000). The data was stratified by NTRAITS = 2, 3 or 4, hence, there were three models of TCORRxNITEMSxSSIZE: 2-trait, 3-trait, and 4-trait models. Replicating this simulated experiment 1,000 times for each set of experimental conditions produced data for all the cells in the models. Descriptive statistics for each cell in each of the three models were calculated. MANOVAs were conducted for all three models to evaluate the main and interaction effects of the independent variables on the validity of the method. A priori planned quadratic function contrasts of various experimental conditions was used to study the curvilinear effect of NITEMS. Estimation precision was also explored through a set of visual illustration data plots.
CHAPTER IV

RESULTS

Once the data was generated as described above, post-processing first combined each data file of 1,000 replications. There were 180 data files altogether, one for each set of IV conditions (3 models, 5 test lengths, 4 sample sizes, and 3 inter-trait correlation conditions). These were combined in a database containing 180,000 records. Each record included the independent variables and the SRT criterion validity coefficient for each trait in the model obtained using the validation sample. Further post-processing added the Fisher z transformed value for each SRT criterion validity coefficient and the upper and lower bounds of the 95% confidence interval (given the sample size) around that Fisher z value. The confidence interval of the back transformed Fisher z data was used to produce the "spaghetti plots".

The rolled-up database was analyzed using MANOVA followed by univariate ANOVA and quadratic contrasts. The MANOVA used the SRT Fisher’s z validity coefficients as the DV vector in a 3-way general linear model. The main effect for each factor and the 2- and 3-way factor interactions were tested. It was also hypothesized that the test length by trait inter-correlation interaction (NITEMS*TCORR) effects (for either individual sample sizes or pooled over sample size) would be curvilinear and this effect can be seen quite clearly in the means plots. This hypothesis was statistically tested through a priori planned polynomial (quadratic) contrasts: first, by testing within each inter-trait correlation level, the curvilinear estimates for NITEMS were the same for all sample sizes and, second, after pooling sample sizes, that the curvilinear estimates were
significantly different and in the direction hypothesized for each level of inter-trait correlation.

**3-Trait Model**

The following sections present in detail the results of the visualizations and the analysis of the data for the 3-trait model and for trait $C_1$, the first of the three traits in that model. Additional analyses and plots for $C_2$ and $C_3$ were similar. For this model the design and data organization are presented in Figure 5.

![Figure 5](image)

**Figure 5.** Factors and sample sizes in the study (3-trait model, $N = 1000/cell$).

**Confidence Interval Plots**

Figures 6-18 present spaghetti plots for the 95% CI of the SRT validity coefficient (the Pearson correlation) as a function of sample size, test length, and inter-trait correlation of a 3-trait model. Three plots illustrate the sample size by test length interaction at three levels of inter-trait correlation, the inter-trait correlation by test length interaction at four levels of sample size and the inter-trait correlation by sample size at five levels of test length. The CI plots clearly followed the hypothesized patterns. For
given inter-trait correlation, increasing test length improved mean SRT validity coefficient (Figures 6-8), and greatly reduced the 95% CIs width until by 15 or more items, the SRT validity coefficients all range in the mid .90's with little variance at all sample sizes (Figure 9). Likewise, the CIs decreased sharply also with little visible variation. The effect of the inter-trait correlation, however, was to blur and weaken these effects at T\text{CORR} = .40 and especially at T\text{CORR} = .80. At these values, the asymptotic limit toward which the SRT validity coefficients converged reduced to about .75, even with test lengths of 15 or 20 and sample sizes of 1,000 or 5,000.

Similar patterns are visible in Figures 10-13, the SRT validity as a function of inter-trait correlation and test length for given sample size. Clearly SRT validity coefficient increases with test length and with sample size, while variability of SRT validity coefficient reduces. Just as clearly, inter-trait correlation blurs and diminishes both effects.

The effects of inter-trait correlation can be seen most clearly in the inter-trait correlation by sample size plots for each test length. Figure 14 shows that for a SRT comprised of one item, when traits are orthogonal, this does surprisingly well when the samples are 250 subjects or larger. But when multiple inter-correlated traits are being estimated, a single item does not provide a great deal of trait resolution, particularly at higher inter-correlation levels. On the other hand, increasing the test length to 5 or 10 or more items greatly improves resolution at all levels of inter-correlation and sample size. Beyond this point, when the inter-correlation approaches or exceeds .80, larger sample sizes and more items do not appear to improve SRT validity or its standard error a great deal more, at least in terms of individual experiments and their individual results.
<table>
<thead>
<tr>
<th>SSIZE</th>
<th>NITEMS=1</th>
<th>NITEMS=5</th>
<th>NITEMS=10</th>
<th>NITEMS=15</th>
<th>NITEMS=20</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td><img src="image1" alt="Plot" /></td>
<td><img src="image2" alt="Plot" /></td>
<td><img src="image3" alt="Plot" /></td>
<td><img src="image4" alt="Plot" /></td>
<td><img src="image5" alt="Plot" /></td>
</tr>
<tr>
<td>250</td>
<td><img src="image6" alt="Plot" /></td>
<td><img src="image7" alt="Plot" /></td>
<td><img src="image8" alt="Plot" /></td>
<td><img src="image9" alt="Plot" /></td>
<td><img src="image10" alt="Plot" /></td>
</tr>
<tr>
<td>1000</td>
<td><img src="image11" alt="Plot" /></td>
<td><img src="image12" alt="Plot" /></td>
<td><img src="image13" alt="Plot" /></td>
<td><img src="image14" alt="Plot" /></td>
<td><img src="image15" alt="Plot" /></td>
</tr>
<tr>
<td>5000</td>
<td><img src="image16" alt="Plot" /></td>
<td><img src="image17" alt="Plot" /></td>
<td><img src="image18" alt="Plot" /></td>
<td><img src="image19" alt="Plot" /></td>
<td><img src="image20" alt="Plot" /></td>
</tr>
</tbody>
</table>

*Figure 6.* CI plots for SRT validity coefficient using the best key found for $C_1$ in the calibration sample, as a function of the sample size and test length with inter-trait correlation = .00 (1,000 replications of a 3-trait model).
<table>
<thead>
<tr>
<th>SSIZE</th>
<th>NITEMS=1</th>
<th>NITEMS=5</th>
<th>NITEMS=10</th>
<th>NITEMS=15</th>
<th>NITEMS=20</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td><img src="image1.png" alt="Plot" /></td>
<td><img src="image2.png" alt="Plot" /></td>
<td><img src="image3.png" alt="Plot" /></td>
<td><img src="image4.png" alt="Plot" /></td>
<td><img src="image5.png" alt="Plot" /></td>
</tr>
<tr>
<td>250</td>
<td><img src="image6.png" alt="Plot" /></td>
<td><img src="image7.png" alt="Plot" /></td>
<td><img src="image8.png" alt="Plot" /></td>
<td><img src="image9.png" alt="Plot" /></td>
<td><img src="image10.png" alt="Plot" /></td>
</tr>
<tr>
<td>1000</td>
<td><img src="image11.png" alt="Plot" /></td>
<td><img src="image12.png" alt="Plot" /></td>
<td><img src="image13.png" alt="Plot" /></td>
<td><img src="image14.png" alt="Plot" /></td>
<td><img src="image15.png" alt="Plot" /></td>
</tr>
<tr>
<td>5000</td>
<td><img src="image16.png" alt="Plot" /></td>
<td><img src="image17.png" alt="Plot" /></td>
<td><img src="image18.png" alt="Plot" /></td>
<td><img src="image19.png" alt="Plot" /></td>
<td><img src="image20.png" alt="Plot" /></td>
</tr>
</tbody>
</table>

*Figure 7. CI plots for the SRT validity coefficient for trait $C_1$ as a function of the sample size and test length with inter-trait correlation $= .40$ (1,000 replications of a 3-trait model).*
SSIZE  |  NITEMS=1 |  NITEMS=5 |  NITEMS=10 |  NITEMS=15 |  NITEMS=20
--- | --- | --- | --- | --- | ---
50  |  |  |  |  |  
250 |  |  |  |  |  
1000 |  |  |  |  |  
5000 |  |  |  |  |  

*Figure 8.* CI plots for the SRT validity coefficient for trait $C_1$ as a function of the sample size and test length with inter-trait correlation $= .80$ (1,000 replications of a 3-trait model).
Figure 9. Mean SRT CI width as a function of inter-trait correlation, test length and sample size (1,000 replications of a 3-trait model).
Figure 10. CI plots for the SRT validity coefficient for trait $C_1$ as a function of the inter-trait correlation and the test length with sample size = 50 (1,000 replications of a 3-trait model).
<table>
<thead>
<tr>
<th>TCORR</th>
<th>NITEMS=1</th>
<th>NITEMS=5</th>
<th>NITEMS=10</th>
<th>NITEMS=15</th>
<th>NITEMS=20</th>
</tr>
</thead>
<tbody>
<tr>
<td>.00</td>
<td><img src="image1" alt="Plot" /></td>
<td><img src="image2" alt="Plot" /></td>
<td><img src="image3" alt="Plot" /></td>
<td><img src="image4" alt="Plot" /></td>
<td><img src="image5" alt="Plot" /></td>
</tr>
<tr>
<td>.40</td>
<td><img src="image6" alt="Plot" /></td>
<td><img src="image7" alt="Plot" /></td>
<td><img src="image8" alt="Plot" /></td>
<td><img src="image9" alt="Plot" /></td>
<td><img src="image10" alt="Plot" /></td>
</tr>
<tr>
<td>.80</td>
<td><img src="image11" alt="Plot" /></td>
<td><img src="image12" alt="Plot" /></td>
<td><img src="image13" alt="Plot" /></td>
<td><img src="image14" alt="Plot" /></td>
<td><img src="image15" alt="Plot" /></td>
</tr>
</tbody>
</table>

*Figure 11. CI plots for the SRT validity coefficient for trait $C_1$ as a function of the inter-trait correlation and the test length with sample size = 250 (1,000 replications of a 3-trait model).*
Figure 12. CI plots for the SRT validity coefficient for trait $C_1$ as a function of the inter-trait correlation and the test length with sample size = 1000 (1,000 replications of a 3-trait model).
<table>
<thead>
<tr>
<th>TCORR</th>
<th>NITEMS=1</th>
<th>NITEMS=5</th>
<th>NITEMS=10</th>
<th>NITEMS=15</th>
<th>NITEMS=20</th>
</tr>
</thead>
<tbody>
<tr>
<td>.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 13.* CI plots for the SRT validity coefficient for trait $C_1$ as a function of the inter-trait correlation and the test length with sample size = 5000 (1,000 replications of a 3-trait model).
Figure 14. CI plots for the SRT validity coefficient for trait $C_1$ as a function of the inter-trait correlation and the sample size with test length = 1 item (1,000 replications of a 3-trait model).
Figure 15. CI plots for the SRT validity coefficient for trait $C_1$ as a function of the inter-trait correlation and the sample size with test length = 5 items (1,000 replications of a 3-trait model).
Figure 16. CI plots for the SRT validity coefficient for trait $C_1$ as a function of the inter-trait correlation and the sample size with test length = 10 items (1,000 replications of a 3-trait model).
<table>
<thead>
<tr>
<th>TCORR</th>
<th>SAMPLE SIZE=50</th>
<th>SAMPLE SIZE=250</th>
<th>SAMPLE SIZE=1000</th>
<th>SAMPLE SIZE=5000</th>
</tr>
</thead>
<tbody>
<tr>
<td>.00</td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
<td><img src="image3" alt="Graph" /></td>
<td><img src="image4" alt="Graph" /></td>
</tr>
<tr>
<td>.40</td>
<td><img src="image5" alt="Graph" /></td>
<td><img src="image6" alt="Graph" /></td>
<td><img src="image7" alt="Graph" /></td>
<td><img src="image8" alt="Graph" /></td>
</tr>
<tr>
<td>.80</td>
<td><img src="image9" alt="Graph" /></td>
<td><img src="image10" alt="Graph" /></td>
<td><img src="image11" alt="Graph" /></td>
<td><img src="image12" alt="Graph" /></td>
</tr>
</tbody>
</table>

*Figure 17.* CI plots for the SRT validity coefficient for trait $C_1$ as a function of the inter-trait correlation and the sample size with test length $= 15$ items (1,000 replications of a 3-trait model).
<table>
<thead>
<tr>
<th>TCORR</th>
<th>SAMPLE SIZE=50</th>
<th>SAMPLE SIZE=250</th>
<th>SAMPLE SIZE=1000</th>
<th>SAMPLE SIZE=5000</th>
</tr>
</thead>
<tbody>
<tr>
<td>.00</td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
<td><img src="image3" alt="Graph" /></td>
<td><img src="image4" alt="Graph" /></td>
</tr>
<tr>
<td>.40</td>
<td><img src="image5" alt="Graph" /></td>
<td><img src="image6" alt="Graph" /></td>
<td><img src="image7" alt="Graph" /></td>
<td><img src="image8" alt="Graph" /></td>
</tr>
<tr>
<td>.80</td>
<td><img src="image9" alt="Graph" /></td>
<td><img src="image10" alt="Graph" /></td>
<td><img src="image11" alt="Graph" /></td>
<td><img src="image12" alt="Graph" /></td>
</tr>
</tbody>
</table>

*Figure 18. CI plots for the SRT validity coefficient for trait $C_1$ as a function of the inter-trait correlation and the sample size with test length = 20 items (1,000 replications of a 3-trait model).*
Descriptive Statistics

As a preliminary step in the quantitative analysis of study data, basic descriptive
statistics for all variables in all cells of the 3 different models were compiled and are
included in Appendix C. These statistics expand and document the results of the
MANOVA models described by the means plots and statistical analyses presented below.

Multivariate Analysis

For the 3-trait model the 3x5x4 between groups MANOVA rejected all
multivariate null hypotheses of this study on the dependent variable vector \[ \left[ \begin{array}{c} r_{z_1s_1} \\ r_{z_2s_2} \\ r_{z_3s_3} \end{array} \right] \], thus
supporting all main effect and interaction research hypotheses, see Table 7. Most
importantly, the multivariate 3-way interaction effect of TCORR, NITEMS and SSIZE
was statistically significant.

Table 7

<table>
<thead>
<tr>
<th>IV</th>
<th>Wilks’ ( \lambda )</th>
<th>F</th>
<th>Num df</th>
<th>Den df</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCORR</td>
<td>.178</td>
<td>27406.6</td>
<td>6</td>
<td>119876</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>NITEMS</td>
<td>.227</td>
<td>9926.58</td>
<td>12</td>
<td>158581</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>TCORR*NITEMS</td>
<td>.801</td>
<td>575.04</td>
<td>24</td>
<td>173839</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>SSIZE</td>
<td>.983</td>
<td>113.73</td>
<td>9</td>
<td>145873</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>TCORR*SSIZE</td>
<td>.999</td>
<td>1.66</td>
<td>18</td>
<td>169531</td>
<td>&lt;.0384</td>
</tr>
<tr>
<td>NITE*SSIZE</td>
<td>.998</td>
<td>1.99</td>
<td>36</td>
<td>177094</td>
<td>.0004</td>
</tr>
<tr>
<td>TCORR<em>NITEMS</em>SSIZE</td>
<td>.998</td>
<td>1.55</td>
<td>72</td>
<td>179125</td>
<td>.0019</td>
</tr>
</tbody>
</table>

Univariate Analyses

Parallel analyses univariate ANOVAs with a family-wise Bonferroni adjustment
on alpha (.05/3 = .0167) were conducted for each DV in each response vector following
the statistically significant multivariate findings. The focus of these univariate ANOVAs
was to establish the presence of a statistically significant 3-way interaction. In two of the
three univariate analyses, the 3-way interaction was statistically significant, see Table 8.

The presence of the statistically significant univariate 3-way interactions allow for the examination of a-priori quadratic effects. Of interest, specifically, were the values of the quadratic parameter estimates as a function of NITEMS.

Table 8

| Univariate 3-Way Interaction Findings for 3-Trait Model, df = 24, 59940 |
|-----------------|---------|--------|--------|
| r_{ZC_1 S_1}    | 24, 59940 | .0845  | 1.90   | .0050  |
| r_{ZC_2 S_2}    | 24, 59940 | .0803  | 1.791  | .0101  |
| r_{ZC_3 S_3}    | 24, 59940 | .0671  | 1.50   | .0562  |

The least squares mean plots for r_{ZC_1 S_1} in Figure 19 presents pronounced effects of the 3-way interaction between TCORR, SSIZE and NITEMS on SRT validity in a 3-trait model.

As hypothesized, there was a noted quadratic shape as a function of test length. This observation is better illustrated in Figure 20 where the interaction effects are plotted individually, along with their 95% confidence intervals for the effect sizes at the points (or the conditions) being examined, for each level of sample size. However, the effect of sample size, while statistically significant, is not as pronounced with unrelated or moderately related traits as it is with highly related traits (Figure 19). Its practical effect seems restricted to reducing the size of the confidence intervals around the estimates (Figure 20). Figures 21 and 22 present similar plots for r_{ZC_2 S_2} and Figures 23 and 24 for r_{ZC_3 S_3}.

Interaction Analysis: Quadratic Curve Fitting

Planned quadratic function contrasts over test length (NITEMS) were fit to evaluate the statistically significant univariate 3-way interactions. Tables 9-10 present the
estimated NITEMS quadratic effects for the Fisher $z$ transformed mean SRT criterion validity coefficients.

*Figure 19.* SRT validity as a function of 3-way interaction between trait inter-correlations, sample size and test length (3-trait model, trait $C_1$).
Figure 20. SRT validity as a function of trait inter-correlations and test length at each sample size (3-trait model, trait $C_1$).
Figure 21. SRT validity as a function of 3-way interaction between trait inter-correlations, sample size and test length (3-trait model, trait $C_2$).
Figure 22. SRT validity as a function of trait inter-correlations and test length at each sample size (3-trait model, trait $C_2$).
Figure 23. SRT validity as a function of 3-way interaction between trait inter-correlations, sample size and test length (3-trait model, trait $C_3$).
Figure 24. SRT validity as a function of trait inter-correlations and test length at each sample size (3-trait model, trait $C_3$).
Table 9

**NITEMS Quadratic Estimates (3-Trait Model)**

<table>
<thead>
<tr>
<th>SSIZE</th>
<th>TCORR = .00</th>
<th>TCORR = .40</th>
<th>TCORR = .80</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
<td>250</td>
<td>1000</td>
</tr>
<tr>
<td>( r_{ZC_1S_1} )</td>
<td>-1.089</td>
<td>-1.143</td>
<td>-1.152</td>
</tr>
<tr>
<td>( r_{ZC_2S_2} )</td>
<td>-1.103</td>
<td>-1.131</td>
<td>-1.193</td>
</tr>
<tr>
<td>( r_{ZC_3S_3} )</td>
<td>-1.084</td>
<td>-1.114</td>
<td>-1.128</td>
</tr>
</tbody>
</table>

*Note: Bold numbers represent the pair of estimates producing the largest difference.*

Table 10

**Quadratic Contrast for Largest SSIZE Parameter Difference (3-Trait Model)**

<table>
<thead>
<tr>
<th>TCORR</th>
<th>CONTRAST</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r_{ZC_1S_1} )</td>
<td>.00</td>
<td>1000 vs. 5000</td>
<td>4.52</td>
</tr>
<tr>
<td></td>
<td>.40</td>
<td>50 vs. 250</td>
<td>2.06</td>
</tr>
<tr>
<td></td>
<td>.80</td>
<td>50 vs. 1000</td>
<td>.66</td>
</tr>
<tr>
<td>( r_{ZC_2S_2} )</td>
<td>.00</td>
<td>50 vs. 1000</td>
<td>6.34</td>
</tr>
<tr>
<td></td>
<td>.40</td>
<td>250 vs. 1000</td>
<td>5.80</td>
</tr>
<tr>
<td></td>
<td>.80</td>
<td>50 vs. 1000</td>
<td>.35</td>
</tr>
<tr>
<td>( r_{ZC_3S_3} )</td>
<td>.00</td>
<td>50 vs. 5000</td>
<td>2.19</td>
</tr>
<tr>
<td></td>
<td>.40</td>
<td>50 vs. 1000</td>
<td>1.45</td>
</tr>
<tr>
<td></td>
<td>.80</td>
<td>50 vs. 250</td>
<td>1.43</td>
</tr>
</tbody>
</table>

*Note: Bold indicates significance at Bonferroni adjusted p-value of .0167.*

Note that three contrasts in Table 10 exceeded a .05 type I error, however at the Bonferroni adjusted level of .0167 the contrast for \( r_{ZC_1S_1} \) and \( r_{ZC_3S_3} \) failed to reach statistical significance. The presence of statistically significant quadratic differences implies that pooling over SSIZE for \( r_{ZC_2S_2} \) is not supported. Thus pooling over SSIZE
was done only for $r_{ZC_1S_1}$ and $r_{ZC_3S_3}$ to expedite further understanding on the quadratic effect of NITEMS on SRT criterion validity.

Table 11 below shows that with pooled sample sizes, the quadratic effect of NITEMS differed significantly for paired TCORR level contrasts indicating the TCORR levels was associated with different quadratic slope magnitudes in the $r_{ZC_1S_1}$ and $r_{ZC_3S_3}$ curves. The quadratic NITEMS negative estimates on mean $r_{ZC_kS_k}$, indicates that for every unit increase of quadratic NITEMS will decrease mean $r_{ZC_kS_k}$ by the corresponding estimates depending on TCORR levels. For example, the decrease of mean $r_{ZC1S1}$ is 1.115 for TCORR = .00; for TCORR = .80, the decrease is smaller, .65. The significant differences of the estimates between these TCORR levels indicate that as criterion traits are more intercorrelated, the quadratic effect of NITEMS is smaller. When TCORR changes from uncorrelated to moderately correlated, the change in quadratic NITEMS effect is small (e.g., .059), but when TCORR changed from moderately correlated to highly correlated, the change in quadratic NITEMS effect was much bigger (e.g., .406).

Table 11

<table>
<thead>
<tr>
<th>DV</th>
<th>TCORR</th>
<th>Quadratic Estimates</th>
<th>Difference</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_{ZC_1S_1}$</td>
<td>.00 vs .40</td>
<td>-1.115 vs. -1.056</td>
<td>.059</td>
<td>11.48</td>
<td>&lt;.0007</td>
</tr>
<tr>
<td></td>
<td>.00 vs .80</td>
<td>-1.115 vs. -.650</td>
<td>.456</td>
<td>695.82</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td></td>
<td>.40 vs .80</td>
<td>-1.056 vs. -.650</td>
<td>.406</td>
<td>528.54</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$r_{ZC_2S_2}$</td>
<td>.00 vs .40</td>
<td>-1.148 vs. -1.036</td>
<td>.112</td>
<td>40.64</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td></td>
<td>.00 vs .80</td>
<td>-1.148 vs. -.643</td>
<td>.505</td>
<td>814.77</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td></td>
<td>.40 vs .80</td>
<td>-1.036 vs. -.643</td>
<td>.339</td>
<td>491.47</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$r_{ZC_3S_3}$</td>
<td>.00 vs .40</td>
<td>-1.116 vs. -1.009</td>
<td>.107</td>
<td>36.26</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td></td>
<td>.00 vs .80</td>
<td>-1.116 vs. -.631</td>
<td>.485</td>
<td>746.32</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td></td>
<td>.40 vs .80</td>
<td>-1.009 vs. -.631</td>
<td>.478</td>
<td>453.58</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

*Note: SSIZE pooling for $r_{ZC_2S_2}$ is illustrated for continenity purposes only.*
2- and 4-Trait Models

Analysis of the 2- and 4-trait models provided similar findings as found in the 3-trait model with slight deviations in the 2-trait model. MANOVA results are presented in Table 12 for the 2-trait model and in Table 13 for the 4-trait model. It should be noted that as more traits were added into the SRT, TCORR began affecting the dependent variables more and more. However, as NITEMS increased, the SRT criterion validity coefficients increased in value and their CI widths narrowed in both 2- and 4-trait models.

Table 12

**MANOVA Results for 2-Trait Model**

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Independent Variable</th>
<th>Wilk’s λ</th>
<th>F Value</th>
<th>Num df</th>
<th>Den df</th>
<th>Pr&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCORR</td>
<td>.641</td>
<td>7471</td>
<td>4</td>
<td>119870</td>
<td>&lt;.0001*</td>
<td></td>
</tr>
<tr>
<td>NITEMS</td>
<td>.230</td>
<td>16246</td>
<td>8</td>
<td>119870</td>
<td>&lt;.0001*</td>
<td></td>
</tr>
<tr>
<td>TCORR*NITEMS</td>
<td>.990</td>
<td>37.45</td>
<td>16</td>
<td>119870</td>
<td>&lt;.0001*</td>
<td></td>
</tr>
<tr>
<td>SSIZE</td>
<td>.999</td>
<td>14.63</td>
<td>6</td>
<td>119870</td>
<td>&lt;.0001*</td>
<td></td>
</tr>
<tr>
<td>TCORR*SSIZE</td>
<td>.998</td>
<td>7.67</td>
<td>12</td>
<td>119870</td>
<td>&lt;.0001*</td>
<td></td>
</tr>
<tr>
<td>NITEMS*SSIZE</td>
<td>.999</td>
<td>1.06</td>
<td>24</td>
<td>119870</td>
<td>.382</td>
<td></td>
</tr>
<tr>
<td>TCORR<em>NITEMS</em>SSIZE</td>
<td>.999</td>
<td>1.12</td>
<td>48</td>
<td>119870</td>
<td>.259</td>
<td></td>
</tr>
</tbody>
</table>

Table 13

**MANOVA Results for 4-Trait Model**

<table>
<thead>
<tr>
<th>Outcome</th>
<th>IV</th>
<th>Wilk’s λ</th>
<th>F Value</th>
<th>Num df</th>
<th>Den df</th>
<th>Pr&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCORR</td>
<td>.092</td>
<td>34356</td>
<td>8</td>
<td>119872</td>
<td>&lt;.0001*</td>
<td></td>
</tr>
<tr>
<td>NITEMS</td>
<td>.164</td>
<td>9240</td>
<td>16</td>
<td>183108</td>
<td>&lt;.0001*</td>
<td></td>
</tr>
<tr>
<td>TCORR*NITEMS</td>
<td>.565</td>
<td>1156</td>
<td>32</td>
<td>221035</td>
<td>&lt;.0001*</td>
<td></td>
</tr>
<tr>
<td>SSIZE</td>
<td>.937</td>
<td>330</td>
<td>12</td>
<td>158576</td>
<td>&lt;.0001*</td>
<td></td>
</tr>
<tr>
<td>TCORR*SSIZE</td>
<td>.999</td>
<td>.99</td>
<td>24</td>
<td>209093</td>
<td>&lt;.0384*</td>
<td></td>
</tr>
<tr>
<td>NITEMS*SSIZE</td>
<td>.995</td>
<td>5.68</td>
<td>48</td>
<td>230882</td>
<td>.0004*</td>
<td></td>
</tr>
<tr>
<td>TCORR<em>NITEMS</em>SSIZE</td>
<td>.997</td>
<td>2.10</td>
<td>96</td>
<td>237438</td>
<td>.0019*</td>
<td></td>
</tr>
</tbody>
</table>

Similar to the 3-trait model statistically significant multivariate interactions were investigated for each univariate outcome after adjusting alpha for the family of dependent
variables in the multivariate vector, e.g., .05/2 = .025. Table 14 presents the univariate 2-way interaction findings. As expected both 2-way interactions were statistically significant.

Table 14

<table>
<thead>
<tr>
<th>Effect</th>
<th>Df</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>r_{ZC_1S_1}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCORR*NITEMS</td>
<td>8</td>
<td>1.927</td>
<td>46.11</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>TCORR*SSIZE</td>
<td>6</td>
<td>.347</td>
<td>8.31</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>r_{ZC_2S_2}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCORR*NITEMS</td>
<td>8</td>
<td>2.194</td>
<td>52.48</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>TCORR*SSIZE</td>
<td>6</td>
<td>.486</td>
<td>11.63</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Figure 25 presents the mean r_{ZC_1S_1} SRT criterion validity coefficients for the 2-trait model. While the 3-way interaction was not statistically significant (Table 12), Figure 25 presents the SRT validity means for comparitive purposes to the 3-trait model previously presented. As can be seen from this figure, SSIZE has no impact on the SRT criterion validity coefficients for r_{C_1S_1}, whereas SSIZE does impact the CI width. The behavior of the remaining independent variable shows similar patterns as in the 3-trait model.
Figure 25. SRT validity as a function of 3-way interaction between trait inter-correlations, sample size and test length (2-trait model, trait $C_1$).

Table 15 presents univariate 3-way interaction findings for the 4-trait model. As expected, all 3-way interactions were statistically significant even after adjusting alpha for the family of dependent variables in the multivariate vector, e.g., $0.05/4 = 0.0125$.

Table 15

<table>
<thead>
<tr>
<th>Outcome</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_{ZC_1S_1}$</td>
<td>0.107</td>
<td>2.91</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$r_{ZC_2S_2}$</td>
<td>0.109</td>
<td>2.94</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$r_{ZC_3S_3}$</td>
<td>0.085</td>
<td>2.31</td>
<td>&lt;.0003</td>
</tr>
<tr>
<td>$r_{ZC_4S_4}$</td>
<td>0.066</td>
<td>1.80</td>
<td>&lt;.0096</td>
</tr>
</tbody>
</table>
Figure 26 presents the mean $c_2s_1$ SRT criterion validity coefficients from the 4-trait model. As can be seen from this figure, SSIZE has a larger impact on the SRT criterion validity coefficients for $zC_1s_1$ such that the quadratic NITEMS forms begin to diverge in magnitude but not in slope particularly in the TCORR = .80 condition.

**Summary of Findings for Three Models**

1) Increasing the number of items or sample size will increase SRT criterion validity coefficients, holding the correlations among the multiple traits constant.

2) Increasing the correlations among the traits decreases the asymptotic limit of the SRT criterion validity coefficient (theoretically this limit is 1.0).
3) When TCORR is high (.80) there appear to be very negligible improvement in the SRT criterion validity coefficient above a test length of 15 items.
CHAPTER V

DISCUSSION

This study proposed and tested a new scoring method for SRT tests through Monte Carlo simulation. It examined, specifically, an approach that yielded multiple criterion trait-keys for simulated responses to multidimensional ranking items. The factors investigated hypothesized to affect $z_{C_kS_k}$, the correlation between the ranking test scores ($S_k$) and the criterion trait ($C_k$) were: number of criterion traits (NTRAITS), correlation among the criterion traits (TCORR), test length in terms of the number of items (NITEMS), and subject sample size (SSIZE). The research hypotheses were: as NITEMS increases, the SRT criterion validity coefficient $z_{C_kS_k}$ (together with their CI lines) will increase and move to the right toward the upper limit of unity. Likewise, as SSIZE goes up, the CI width will become narrower. Furthermore, the upper limit of $z_{C_kS_k}$ is going to be an inverse function of NTRAITS and TCORR (the correlation of those traits). Formal null hypotheses were:

1. Given the number of criterion traits, there is no IV (TCORR, NITEMS and SSIZE) main effects on the DV, $z_{C_kS_k}$.

2. Given the number of criterion traits, there are no 2-way independent variable (TCORR, NITEMS and SSIZE) interactions effect on the DV, $z_{C_kS_k}$.

3. Given the number of criterion traits, there is no 3-way independent variable (TCORR, NITEMS and SSIZE) interaction effect on the DV, $z_{C_kS_k}$. 
Results indicated that the scientific hypotheses were largely supported: increasing NTRAITS and TCORR values resulted in lower SRT criterion validity coefficient, \( r_{zCkS_k} \) and increases in CI width. Moreover, the mean of \( r_{zCkS_k} \) formed a curvilinear function over the values of NITEMS. This dissertation demonstrated that with randomly generated subject’s rank responses and randomly generated criterion test scores, there exists a unique and “best” trait-key for every criterion for each item.

**Significance of the Study**

Complex behaviors (e.g., teachers dealing with classroom situations, doctors dealing with emergency room patients, soldiers on a battlefield facing combat situations) are based on the subject’s learned elaborate sets of values and attitudes (neural subroutines adjusting weights or predispositions for certain types of actions). Modeling this collection produces a "theory of mind" for each subject, e.g., each teacher, doctor, soldier, etc. IRT models or learning curves can estimate the probability that a person's state model has reached a certain point that at last, with a probability of, say 95%, this student has learned a "skill" or "fact." Big data researchers gather such data now through on-line tutoring and observed classroom monitoring using algorithms to estimate "on-task" or "distracting" or momentary "group" behaviors as students' progress through lessons (Bughin, Chui, & Manyika, 2010). “Theory of mind” has also been modeled through SRT that consist of contextual item stems and sets of stimulus alternative actions to be ranked by subjects in preference order (e.g., the TSRT).

An extensive literature dealing with ranked data exists (Marden, 1995). However, expert advice is to avoid it due to analytical difficulties. This Monte Carlo study demonstrated how to score a hypothetical multidimensional SRT test. Given situations
involving multiple criterion traits such as $C_1$, $C_2$, $C_3$, we can choose sets of alternatives based on dimensionality and ability of differentiating subjects. The subjects’ ranking patterns can be scored against different keys to distinguish different traits, e.g., $C_1$-key, $C_2$-key, or $C_3$-key. Pearson product-moment correlations were used to identify sets of trait-keys for estimating corresponding traits of interest from each item in the test.

Traditionally, only one rank pattern is used as the key to score an examinee’s response for each item and one composite score is obtained to evaluate an examinee’s trait that is appropriate for an externally referenced task but potentially inappropriate for an internally based referent. This study proposed a mechanism that identifies a ranking pattern with optimal rank order relationship with a known criterion score (ORORCS) as the “best” key for this criterion trait. The ORORCS yields “one key for one trait” and “multiple keys for multiple traits” solution to a complex ranking task. By applying the multiple criterion trait-keys to score subjects’ multidimensional rankings, the subject’s item trait-scores will be obtained and the average or total of those scores are the subject’s test trait-scores which may be claimed the “best” estimates of the criterion traits of the subject.

This dissertation demonstrated the feasibility for scoring multidimensional ranking SRTs. The theory behind the ORORCS corroborated with other researchers, e.g., Dr. Andrew McAfee, the associate director of the Center for Digital Business at the MIT Sloan School of Management as quoted in Planes (2013) and a group of computer science scholars who recently presented a novel technique on dimensionality reduction (Stoyanovich et al, 2013). In an interview on education in the future, McAfee stated, “After all, most modern games (at least the well-designed ones) give players multiple
pathways to the completion of any given task, which turns the multiple-choice standardized testing paradigm on its head. Measuring learning doesn’t have to mean that we measure it with a single right answer to any given question.” Meanwhile, Stoyanovich and her group argued that aggregation of rankings was most meaningful in presence of structure (i.e., of agreement among the rankers revealed in the response matrix) and that structure must be identified before meaningful aggregation can take place. They focused on local structure in ranked data (i.e., the agreement of subsets of rankers over subsets of options) as a technique of reducing the size of the state space. They break up the state space into smaller sub-spaces by looking for multiple local models representing the distribution over a subset of options, rather than fitting a single global model to the whole set of data. For example, one model may represent the rankings of fruits, another – the rankings of protein rich foods (which include certain meats and vegetables), yet another – the rankings of sweet foods. Thus, multiple models can be defined over the same list of options to accommodate diversity of opinions that have the potential of fitting the data better. This is exactly what was proposed and evaluated within this dissertation.

The ORORCS method has practical meaning in fields including education, testing and survey research. First, it provides a way for test users to score students’ responses to ranking items so that teachers no longer need to avoid using ranking items due to the difficulty of analyzing such items and interpreting results from such items (O’Kuma, Maloney & Hieggelke, 2003). Second, it provides test developers with an alternative to unidimensionality, meanwhile facing the fact that multidimensionality in real test scores are more realistic while perhaps more intractable (Gessaroli & De Champlain, 2005; Hattie, 1985; Nandakumar, 1991). Although most of the time test items have been
carefully designed to achieve unidimensionality according to classical test theory or IRT, the inherent multidimensionality of real test data has been recognized and suggested to be deliberately constructed into a test (e.g., Humphreys, 1952, 1962, 1970, 1981, 1985, and 1986). Multidimensionality is no longer something to be avoided by test designers. Last, the ORORCS method enables inferences about individual’s specific latent traits that are hypothesized to underlie his/her preference rankings of a given set of stimuli. This information may have tremendous practical meaning for education and social studies, especially those that are diagnostic or selective studies.

**Implications for SRT Test Construction**

One of the most interesting implications for SRTs’ test construction is the fact that, within sets of experimental conditions, each based on a different random sample of subjects interacting with a *randomly designed SRT*, the correlations among SRT criterion validity coefficients among traits in each model studied were uniformly high, see Table 16. It was for this reason that the data analysis for this study first began with a multivariate model and multivariate effects and then secondly looked at a single trait ($C_1$) and univariate effects. This demonstrated that the scoring model generally works well with any SRT. However, that does not mean that the SRT criterion validity will be high or that its confidence interval will be narrow for any particular experiment. What it does imply is that if the SRT is first well-constructed and the subjects for calibration and validation are appropriate, then results using this scoring method for all traits in the test are likely to be measured with similar degrees of precision (CI width) and accuracy (magnitude of) the criterion validity coefficient. On the other hand, if the test is poorly constructed and/or the sample not appropriate, the results – even when scored using this
multi-trait method – will likely be poor. This, of course, can be said about any test using any scoring method. However, SRTs may not be particularly easy or simple to construct and test designers should take note.

Table 16

<table>
<thead>
<tr>
<th>Model</th>
<th>Trait Pair</th>
<th>SRT Validity</th>
<th>Fisher z Transform</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-Trait</td>
<td>C1S1 w C2S2</td>
<td>.7506</td>
<td>.8123</td>
</tr>
<tr>
<td>3-Trait</td>
<td>C1S1 w C2S2</td>
<td>.7799</td>
<td>.8555</td>
</tr>
<tr>
<td></td>
<td>C1S1 w C3S3</td>
<td>.7822</td>
<td>.8561</td>
</tr>
<tr>
<td></td>
<td>C2S2 w C3S3</td>
<td>.7764</td>
<td>.8545</td>
</tr>
<tr>
<td>4-Trait</td>
<td>C1S1 w C2S2</td>
<td>.8125</td>
<td>.8634</td>
</tr>
<tr>
<td></td>
<td>C1S1 w C3S3</td>
<td>.8134</td>
<td>.8640</td>
</tr>
<tr>
<td></td>
<td>C1S1 w C4S4</td>
<td>.8112</td>
<td>.8638</td>
</tr>
<tr>
<td></td>
<td>C2S2 w C3S3</td>
<td>.8120</td>
<td>.8627</td>
</tr>
<tr>
<td></td>
<td>C2S2 w C4S4</td>
<td>.8118</td>
<td>.8634</td>
</tr>
<tr>
<td></td>
<td>C3S3 w C4S4</td>
<td>.8143</td>
<td>.8640</td>
</tr>
</tbody>
</table>

Limitations of the Study

The major limitation of this study is that it did not examine all the relevant experimental conditions, e.g., the effect of the reliability of the criterion instruments or the effect of the number of alternatives in each item. This study only considered cases when there are 4 alternatives in each item. Also, the correlation among the 3 and 4 criterion traits to be examined in this study were all set to be of the same value, e.g., $\text{cov}(C_1, C_2) = \text{cov}(C_2, C_3)$, which is rarely the case in real tests.

Another limitation of this study is relevant to the simulation of preference rankings. Multidimensional preference decision is a complicated problem. For instance, normative theories of value maximization (e.g., Neumann & Morgenstern, 1953) posited that each subject possesses stable preferences for all possible stimulus options. However, there are a great number of studies that have questioned the existence of such a global
internal preference set (Reighluth & Moore, 1999). Strong empirical evidence shows that preferences are not simply revealed, but are actually constructed in the process of their elicitation (e.g., Slovic, Fischhoff, & Lichtenstein, 1979; Kahneman & Tversky, 2000; Slovic, Griffin & Tversky, 1990). It is likely that preferences become more firmly established with experience through a dynamic process (Herr, 1986; Shafir, Simonson & Tversky, 1993; Simonson & Tversky, 1992; Slovic, 1995). This suggests that individuals have personal strategies that they use to assemble their preferences, and these strategies are unlikely to produce preferences that are invariant over time. Nonetheless, this study assumed people used the same strategies over time and did not address the problem of invariant preferences.

Furthermore, there has not yet been a universally accepted best model to simulate individuals’ choice decisions, especially multidimensional ones. It seems to be under eternal debate in respect to how an individual weighs different criteria and ranks a set of multi-criterion alternatives (Tsetsos, Usher & McClelland, 2011). Most choice models assume a rational person making rational choices. Irrational decisions are very likely in real life but there has not been a model capable of integrating irrationality into a choice model (De Martino, Kumaran & Seymour, 2006). This study simply adopted the rational decision making model and the completely cognitive-science/neuro-science model of mostly unconscious decision weighting. This study generated preference rankings and simulated in only the most feasible existing models for multi-criterion decision making, while leaving unknown whether this choice of model is actually true to the reality or not. For example, no “bar” was considered in this study, which refers to a special kind of
criterion which – if an object does not meet – then this object will be automatically excluded from further considerations by a respondent.

Lastly, the ORORCS requires that, at least at test developing stage, the subject’s criterion trait score is already known so that the criterion trait-key for the items in the ranking test could be determined. The criterion trait scores have to be estimated by administering the criterion tests either concurrently or prospectively with the SRT.

**Conclusions and Suggestions for Further Research**

Appropriate scoring methods for tests should be based on theories of the construct domains of such tests (Messick, 1989). This is called *structural fidelity* (Loevinger, 1957), a necessary but not sufficient condition for construct validity (Keith & Kranzler, 1999). A situation reaction test (SRT) consists of items with alternatives to be ranked according to subjects’ best judgment. This study demonstrated a new multi-trait scoring method appropriate when SRT ranking items and their alternatives are theoretically governed by multiple traits and external criterion trait scores are known. This scoring method produces multiple trait scores from a single administration of an SRT test, e.g., there are multiple scoring keys, one specific for each trait. Study results indicate the SRT scoring method, the ORORCS, yields concurrent validity estimates that are structurally valid.

The ORORCS procedure can effectively measure subject states with respect to multiple dimension hypothesized to underlie ranking choice behaviors even when these factors vary in their correlation. Calibration and validation sample size, though significant, was not much of a factor above 250. Test length, number of ranking items, was important, but satisfactory results did not require extensively long tests. Good test/item design is more
fundamentally more important than structural issues such as trait correlation, sample size, and the number of items in an SRT test with four alternatives.
REFERENCES


Duncan, J. K., Hough, J. B., & Thompson, J. (1966). Exploratory studies of a teaching situation reaction test. A paper read at the annual AERA, Chicago, IL.


Gold, L. L. (1966). *Verbal interaction patterns in the classrooms of selected science teachers; biology* (Doctoral Dissertation). Ohio State University, Columbus, OH. (ED031400)


Hough, J. B., & Duncan, J. K. (1965, February). Exploratory studies of Teaching
Situation Reaction Test. Paper read at the Annual Conference of American


understanding of key astronomy topics* (Doctoral Dissertation). The University of
South Africa, Pretoria, South Africa.


(Ed.), *Current problems and techniques in multivariate psychology: Proceedings
of a conference honoring Professor Paul Horst* (pp. 23-32). Seattle: University of
Washington.

Humphreys, L. G. (1981). The primary mental ability. In M. P. Friedman, J. P. Das, & N.
O'Connor (Eds.), *Intelligence and learning* (pp. 87-102). New York: Plenum
Press.

theory. In B. B. Wolman (Ed.), *Handbook of intelligence* (pp. 201-224). New
York: Wiley. Humphreys, L. G. (1986). An analysis and evaluation of test and


culminating event (Technical Report 1138). United States Army Research
Institute for the Behavioral and Social Sciences. Retrieved from
http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA417360


*The report of the Institute of Manpower Utilization and Government Personnel*.

Loevinger, J. (1957). Objective tests as instruments of psychological theory.
*Psychological Reports, 3*, 635-694 (Monograph Supplement 9).


Murray, C. K. (1967). *A Study of the construct validity of the Teaching Situation Reaction Test* (Doctoral Dissertation). The Ohio State University, Columbus, OH.


Stauffer, M. L. J. (1973). *Impact of study abroad experience on prospective teachers* (Doctoral Dissertation). The Ohio State University, Columbus, OH.


Stoyanovich, J. et al. (2013). Understanding local structure in ranked datasets. This article is published under a Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/), which permits distribution and reproduction in any medium as well allowing derivative works, provided that you attribute the original work to the author(s) and CIDR 2013. 6th Biennial Conference on Innovative Data Systems Research (CIDR ’13) January 6-9, 2013, Asilomar, CA.


GLOSSARY

Test Validity

It refers to the usefulness of inferences drawn from test scores for a specific purpose under certain conditions.

Criterion Validity

This is a type of test validity. It is evaluated by comparing the test scores with one or more external variables (called criteria) considered to provide a direct measure of the characteristic or behavior in question.

Situation Reaction Test (SRT)

It is a kind of test composed of items that depict a specific situational context and provide a set of multi-dimensional alternative reactions to that situation for the subjects to rank according to their best judgment. It is believed that multiple traits of a respondent are involved/affect the way that he/she ranks the alternatives.

Situational Reaction Test (SRT) Criterion Validity Coefficient

SRT criterion validity coefficient $r_{CkSk}$ is obtained when the criterion trait-key for a given criterion $k$ is used to score the respondents’ rankings and the yielded test score for this criterion $S_k$ is correlated with the respondents’ score on this criterion trait scores $C_k$.

Rank

It is to assign numbers to objects, “1” indicating preferred the most. E.g., for 4 objects, $A$, $B$, $C$ and $D$, a possible rank is “1234” indicating $A$ is preferred to $B$, $B$ is preferred to $C$, and $C$ is preferred to $D$.

Order
It is to list objects with the most preferred set on the first position: e.g. “ABCD” indicating A is preferred to B, B is preferred to C, and C is preferred to D.

**Ranking**

It is a set of alternatives ordered by a subject based on certain criteria.

**Action Alternative**

In a certain situation there are more than one choice of reaction to a stimulus and each of these choices is an action alternative.

**State Space**

It is a multidimensional space in which a person’s position is to be estimated along multiple dimensions.

**Preference Rankings**

They are rankings resulting from unspecified or internal criteria and are based primarily on respondents’ own preference or best judgment regarding the importance of the alternatives. They are used to refer internal criterion-based rankings in this study.

**Ranking Item**

It is a type of item that elicits respondents’ rankings of the alternatives in the items.

**All Possible Rankings**

They are the permutation of a set of alternatives in an item.

**All Possible Keys**

For an item of \( m \) alternatives, the all possible keys are the \( m! \) patterns of rankings.

**All Possible Ranking Item Scores**
When a subject’s preference ranking of a set of alternatives in an item is scored against all-possible rankings of these alternatives, the resulting rank scores are called all possible ranking test scores.

**Ranking Test Score (RS)**

When a subject’s ranking of a list of alternatives in an item is scored against a key, the resulting score is a ranking test score for this subject. The ranking test score on trait $k$ for subject $p$ is denoted as $S_{kp}$.

**Criterion Trait Score**

Denoted as $C_{kp}$, it is the coordinate of subject $p$ along dimension $k$ in a multidimensional preference space. Rank score $S_{kp}$ is the estimate of $C_{kp}$. In practice, it is obtained by administering the criterion instruments to the subjects.

**Criterion Trait**

It is a trait of subjects that is measured both by the target ranking item test and by a known criterion instrument.

**Trait-Key**

It is the key that highlights a specific trait for an item. One item may have one or multiple trait-keys depending on the response pattern from the subjects.

**Criterion Trait-Key**

It is the best key among the all possible keys for a criterion trait in the sense that it produces ranking test scores most highly correlated with the score estimated by a well-established criterion instrument on that criterion trait.
Transitivity

It is a mathematical word signifying that if the subject prefers alternatives $A$ to $B$ and $B$ to $C$, then he/she should prefer $A$ to $C$, but not the opposite.

$J$ scale

In the unfolding theory concerning one dimension (Coombs, 1952, 1954, 1964; Kyngdon, 2006), each subject and each alternative may be represented by a point on a common dimension (psychological attribute), which formed a $J$ scale (joint scale). The number of possible qualitative $J$ scales equals the total number of the combinations of $m$ alternatives (i.e., $m!$) half of which are distinct.

Multidimensional Joint Space

A multidimensional psychological space in which both the subjects and the alternatives situates as points along multiple dimensions.

Action Alternative

In a certain situation there are more than one choice of reaction to a stimulus and each of these choices is an action alternative.
APPENDIX A

Teaching Situation Reaction Test Item Sample
“You have been employed by a school system which is engaged in a series of experimental studies. One of these studies involves a class designed to improve pupils' general adjustment to their environment. A heterogeneous group (physically, mentally, socially) of twenty-five 13- to 14-year-old youngsters have signed up for this class entitled, "Teen Topics" because they thought it would be interesting.

The class is scheduled to meet the last period of the day on Tuesday and Thursday during the second semester. Arrangements have been made so that the class might take trips and students might meet informally with the teacher after class. You have accepted the principal's invitation to take this class.

You have been given pretty much of a free hand to develop the course. You have a teacher-counselor to help you and a good supply of instructional materials available. Studies will be made of the personal adjustment gains evidenced by a selected number of your twenty-five students.”

Below are two sample items under one sample sub-situation (A):

“A. You have about eight weeks plus Christmas vacation to plan for your class:

1. When you begin planning the course you would:

   (a) Ask your teacher-counselor what he thinks should be in the course.

   (b) Examine the materials available to you and determine how they might be used by members of the class.

   (c) Read through the copies of publications describing other school programs of a similar nature and draw ideas from them.

   (d) Interview a randomly selected group of the young people signed up for the course and set your own tentative objectives based on those interviews.
2. During early December an important local civic group comes out against teaching sex education in the schools. Your planning had included some sex education. At this point in your planning you would:

(a) Continue planning as you have been.

(b) Ask the principal if you should include any sex education in your course.

(c) Remove the lessons dealing with sex education.

(d) Find out ways to get the sex education material across without causing an issue.”

We can see that the options under a ranking task item are all plausible choices, which is different from a multiple choice item in which there is only one right option and all the other options are wrong answer. This characteristic of ranking task items is very relevant to the scoring scheme and the assessment of the dimensionality of a ranking task item which will be discussed later.
APPENDIX B

Two Examples of Astronomy Ranking Task Retrieved from

http://www.apsu.edu/sites/apsu.edu/files/astronomy/DopplerShift-2_Solution.pdf
Astronomy Ranking Task:
Doppler Shift

Exercise #2

Description: The figure below shows the motion of five distant stars (A - E) relative to a stationary observer (telescope). The speed and direction of each star is indicated by the length and direction of the arrows shown.

Observer  Distant Stars

**Ranking Instructions:** Rank the Doppler shift of the light observed from each star (A – E) from greatest “blueshift”, through no shift, to greatest “redshift”.

**Ranking Order:**

Greatest blueshift 1 _E_  2 _D_  3 _B_  4 _C_  5 _A_  Greatest redshift

Or, the Doppler shift for each star is the same. _____ (indicate with check mark).

Carefully explain your reasoning for ranking this way:
Recall the Doppler shift is ONLY concerned with the radial velocity component (along the line of sight) between the observer and the source, not the distance between them. Also recall, a blueshift is approaching. From the instructions (top of the page) the arrow length and direction is the key.  E and D are both moving toward the observer, but E is moving fastest (longest arrow) so it has the highest blueshift.  B has no radial velocity component so it is next.  A and C are both moving away.  A is fastest so it has the highest redshift.
Astronomy Ranking Task:
Size & Scale

Exercise #2

Description: Consider the images of different astronomical objects below (A-G).

A. The Solar System

B. Globular Cluster

C. Neutron Star

D. Andromeda Galaxy

E. The Sun

F. Nebula

G. Galaxy Cluster

Ranking instructions: Rank the size (from largest to smallest) of the different objects (A-G).

Ranking Order: Largest 1 G 2 D 3 F 4 B 5 A 6 E 7 C Smallest

Or, all the objects have the same size. _____ (indicate with a check mark)

Carefully explain your reasoning for ranking this way:

Copyright © 2005 Conceptual Astronomy and Physics Education Research (CAPER) Team
University of Arizona
APPENDIX C

Descriptive Statistics for All Cells in the 3-Trait Analytical Model
### 2-Trait Model: Means and StDs for Cs by IV Combinations (N=1000/cell)

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Mean</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCORR=.00 SSIZE=50 NITEMS</td>
<td>$r_{s1s1}$ $r_{s2s2}$ $r_{s1s1}$ $r_{s2s2}$</td>
<td>$r_{s1s1}$ $r_{s2s2}$ $r_{s1s1}$ $r_{s2s2}$</td>
</tr>
<tr>
<td>1</td>
<td>1.116</td>
<td>1.106</td>
</tr>
<tr>
<td>5</td>
<td>1.647</td>
<td>1.656</td>
</tr>
<tr>
<td>10</td>
<td>1.805</td>
<td>1.805</td>
</tr>
<tr>
<td>15</td>
<td>1.868</td>
<td>1.873</td>
</tr>
<tr>
<td>20</td>
<td>1.905</td>
<td>1.907</td>
</tr>
<tr>
<td>SSIZE=250 NITEMS</td>
<td>$r_{s1s1}$ $r_{s2s2}$ $r_{s1s1}$ $r_{s2s2}$</td>
<td>$r_{s1s1}$ $r_{s2s2}$ $r_{s1s1}$ $r_{s2s2}$</td>
</tr>
<tr>
<td>1</td>
<td>1.126</td>
<td>1.112</td>
</tr>
<tr>
<td>5</td>
<td>1.650</td>
<td>1.657</td>
</tr>
<tr>
<td>10</td>
<td>1.808</td>
<td>1.810</td>
</tr>
<tr>
<td>15</td>
<td>1.874</td>
<td>1.872</td>
</tr>
<tr>
<td>20</td>
<td>1.903</td>
<td>1.903</td>
</tr>
<tr>
<td>SSIZE=1000 NITEMS</td>
<td>$r_{s1s1}$ $r_{s2s2}$ $r_{s1s1}$ $r_{s2s2}$</td>
<td>$r_{s1s1}$ $r_{s2s2}$ $r_{s1s1}$ $r_{s2s2}$</td>
</tr>
<tr>
<td>1</td>
<td>1.125</td>
<td>1.122</td>
</tr>
<tr>
<td>5</td>
<td>1.657</td>
<td>1.659</td>
</tr>
<tr>
<td>10</td>
<td>1.801</td>
<td>1.805</td>
</tr>
<tr>
<td>15</td>
<td>1.869</td>
<td>1.872</td>
</tr>
<tr>
<td>20</td>
<td>1.906</td>
<td>1.908</td>
</tr>
<tr>
<td>SSIZE=5000 NITEMS</td>
<td>$r_{s1s1}$ $r_{s2s2}$ $r_{s1s1}$ $r_{s2s2}$</td>
<td>$r_{s1s1}$ $r_{s2s2}$ $r_{s1s1}$ $r_{s2s2}$</td>
</tr>
<tr>
<td>1</td>
<td>1.110</td>
<td>1.124</td>
</tr>
<tr>
<td>5</td>
<td>1.656</td>
<td>1.659</td>
</tr>
<tr>
<td>10</td>
<td>1.811</td>
<td>1.811</td>
</tr>
<tr>
<td>15</td>
<td>1.871</td>
<td>1.870</td>
</tr>
<tr>
<td>20</td>
<td>1.909</td>
<td>1.904</td>
</tr>
<tr>
<td>TCORR=.40 SSIZE=50 NITEMS</td>
<td>$r_{s1s1}$ $r_{s2s2}$ $r_{s1s1}$ $r_{s2s2}$</td>
<td>$r_{s1s1}$ $r_{s2s2}$ $r_{s1s1}$ $r_{s2s2}$</td>
</tr>
<tr>
<td>1</td>
<td>1.062</td>
<td>1.501</td>
</tr>
<tr>
<td>5</td>
<td>1.599</td>
<td>1.609</td>
</tr>
<tr>
<td>10</td>
<td>1.768</td>
<td>1.773</td>
</tr>
<tr>
<td>15</td>
<td>1.836</td>
<td>1.831</td>
</tr>
<tr>
<td>20</td>
<td>1.881</td>
<td>1.876</td>
</tr>
<tr>
<td>SSIZE=250 NITEMS</td>
<td>$r_{s1s1}$ $r_{s2s2}$ $r_{s1s1}$ $r_{s2s2}$</td>
<td>$r_{s1s1}$ $r_{s2s2}$ $r_{s1s1}$ $r_{s2s2}$</td>
</tr>
<tr>
<td>1</td>
<td>1.070</td>
<td>1.062</td>
</tr>
<tr>
<td>5</td>
<td>1.601</td>
<td>1.615</td>
</tr>
<tr>
<td>10</td>
<td>1.771</td>
<td>1.772</td>
</tr>
<tr>
<td>15</td>
<td>1.834</td>
<td>1.844</td>
</tr>
<tr>
<td>20</td>
<td>1.880</td>
<td>1.873</td>
</tr>
<tr>
<td>SSIZE=1000 NITEMS</td>
<td>$r_{s1s1}$ $r_{s2s2}$ $r_{s1s1}$ $r_{s2s2}$</td>
<td>$r_{s1s1}$ $r_{s2s2}$ $r_{s1s1}$ $r_{s2s2}$</td>
</tr>
<tr>
<td>1</td>
<td>1.062</td>
<td>1.067</td>
</tr>
<tr>
<td>5</td>
<td>1.615</td>
<td>1.607</td>
</tr>
<tr>
<td>10</td>
<td>1.778</td>
<td>1.777</td>
</tr>
<tr>
<td>15</td>
<td>1.844</td>
<td>1.840</td>
</tr>
<tr>
<td>20</td>
<td>1.882</td>
<td>1.883</td>
</tr>
<tr>
<td>SSIZE=5000 NITEMS</td>
<td>$r_{s1s1}$ $r_{s2s2}$ $r_{s1s1}$ $r_{s2s2}$</td>
<td>$r_{s1s1}$ $r_{s2s2}$ $r_{s1s1}$ $r_{s2s2}$</td>
</tr>
<tr>
<td>1</td>
<td>1.054</td>
<td>1.057</td>
</tr>
<tr>
<td>5</td>
<td>1.609</td>
<td>1.611</td>
</tr>
<tr>
<td>10</td>
<td>1.776</td>
<td>1.773</td>
</tr>
<tr>
<td>15</td>
<td>1.846</td>
<td>1.844</td>
</tr>
<tr>
<td>20</td>
<td>1.881</td>
<td>1.878</td>
</tr>
<tr>
<td>TCORR=.80 SSIZE=50 NITEMS</td>
<td>$r_{s1s1}$ $r_{s2s2}$ $r_{s1s1}$ $r_{s2s2}$</td>
<td>$r_{s1s1}$ $r_{s2s2}$ $r_{s1s1}$ $r_{s2s2}$</td>
</tr>
<tr>
<td>1</td>
<td>.778</td>
<td>.775</td>
</tr>
<tr>
<td>5</td>
<td>1.317</td>
<td>1.288</td>
</tr>
<tr>
<td>10</td>
<td>1.494</td>
<td>1.495</td>
</tr>
<tr>
<td>15</td>
<td>1.584</td>
<td>1.582</td>
</tr>
<tr>
<td>20</td>
<td>1.657</td>
<td>1.663</td>
</tr>
<tr>
<td>SSIZE=250 NITEMS</td>
<td>$r_{s1s1}$ $r_{s2s2}$ $r_{s1s1}$ $r_{s2s2}$</td>
<td>$r_{s1s1}$ $r_{s2s2}$ $r_{s1s1}$ $r_{s2s2}$</td>
</tr>
<tr>
<td>1</td>
<td>.784</td>
<td>.786</td>
</tr>
<tr>
<td>5</td>
<td>1.347</td>
<td>1.341</td>
</tr>
<tr>
<td>10</td>
<td>1.547</td>
<td>1.544</td>
</tr>
<tr>
<td>15</td>
<td>1.627</td>
<td>1.628</td>
</tr>
<tr>
<td>20</td>
<td>1.671</td>
<td>1.685</td>
</tr>
</tbody>
</table>
### 3-Trait Model: Means and StDs for C by IV Combinations (N=1000/cell)

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Mean</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TCORR=.00</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SSIZE=50</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NITEMS</td>
<td>r_{zc1s1}</td>
<td>r_{zc2s2}</td>
</tr>
<tr>
<td>1</td>
<td>.858</td>
<td>.866</td>
</tr>
<tr>
<td>5</td>
<td>1.431</td>
<td>1.429</td>
</tr>
<tr>
<td>10</td>
<td>1.625</td>
<td>1.633</td>
</tr>
<tr>
<td>15</td>
<td>1.733</td>
<td>1.734</td>
</tr>
<tr>
<td>20</td>
<td>1.804</td>
<td>1.796</td>
</tr>
<tr>
<td><strong>SSIZE=250</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NITEMS</td>
<td>r_{zc1s1}</td>
<td>r_{zc2s2}</td>
</tr>
<tr>
<td>1</td>
<td>.870</td>
<td>.873</td>
</tr>
<tr>
<td>5</td>
<td>1.465</td>
<td>1.445</td>
</tr>
<tr>
<td>10</td>
<td>1.671</td>
<td>1.681</td>
</tr>
<tr>
<td>15</td>
<td>1.777</td>
<td>1.779</td>
</tr>
<tr>
<td>20</td>
<td>1.851</td>
<td>1.854</td>
</tr>
<tr>
<td><strong>SSIZE=1000</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NITEMS</td>
<td>r_{zc1s1}</td>
<td>r_{zc2s2}</td>
</tr>
<tr>
<td>1</td>
<td>.904</td>
<td>.869</td>
</tr>
<tr>
<td>5</td>
<td>1.456</td>
<td>1.467</td>
</tr>
<tr>
<td>10</td>
<td>1.683</td>
<td>1.684</td>
</tr>
<tr>
<td>15</td>
<td>1.793</td>
<td>1.791</td>
</tr>
<tr>
<td>20</td>
<td>1.865</td>
<td>1.861</td>
</tr>
<tr>
<td><strong>SSIZE=5000</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NITEMS</td>
<td>r_{zc1s1}</td>
<td>r_{zc2s2}</td>
</tr>
<tr>
<td>1</td>
<td>.749</td>
<td>.742</td>
</tr>
<tr>
<td>5</td>
<td>1.262</td>
<td>1.261</td>
</tr>
<tr>
<td>10</td>
<td>1.453</td>
<td>1.442</td>
</tr>
<tr>
<td>15</td>
<td>1.524</td>
<td>1.510</td>
</tr>
<tr>
<td>20</td>
<td>1.558</td>
<td>1.562</td>
</tr>
<tr>
<td><strong>TCORR=.40</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SSIZE=50</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NITEMS</td>
<td>r_{zc1s1}</td>
<td>r_{zc2s2}</td>
</tr>
<tr>
<td>1</td>
<td>.773</td>
<td>.787</td>
</tr>
<tr>
<td>5</td>
<td>1.289</td>
<td>1.288</td>
</tr>
<tr>
<td>10</td>
<td>1.482</td>
<td>1.473</td>
</tr>
<tr>
<td>15</td>
<td>1.570</td>
<td>1.570</td>
</tr>
<tr>
<td>20</td>
<td>1.630</td>
<td>1.626</td>
</tr>
<tr>
<td><strong>SSIZE=1000</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NITEMS</td>
<td>r_{zc1s1}</td>
<td>r_{zc2s2}</td>
</tr>
<tr>
<td>1</td>
<td>.766</td>
<td>.778</td>
</tr>
<tr>
<td>5</td>
<td>1.294</td>
<td>1.295</td>
</tr>
<tr>
<td>10</td>
<td>1.500</td>
<td>1.505</td>
</tr>
<tr>
<td>15</td>
<td>1.570</td>
<td>1.570</td>
</tr>
<tr>
<td>20</td>
<td>1.630</td>
<td>1.626</td>
</tr>
<tr>
<td><strong>SSIZE=5000</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NITEMS</td>
<td>r_{zc1s1}</td>
<td>r_{zc2s2}</td>
</tr>
<tr>
<td>1</td>
<td>.774</td>
<td>.777</td>
</tr>
<tr>
<td>5</td>
<td>1.301</td>
<td>1.294</td>
</tr>
<tr>
<td>10</td>
<td>1.483</td>
<td>1.485</td>
</tr>
<tr>
<td>15</td>
<td>1.580</td>
<td>1.581</td>
</tr>
<tr>
<td>20</td>
<td>1.626</td>
<td>1.628</td>
</tr>
</tbody>
</table>
### 4-Trait Model: Means and Stds for C by IV Combinations (N=1000/cell)

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Mean</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCORR=.00 SSIZE=50 NITEMS</td>
<td>r_{zc1s1} r_{zc2s2} r_{zc3s3} r_{zc4s4}</td>
<td>.712 .712 .700 .712</td>
</tr>
<tr>
<td>1</td>
<td>.712</td>
<td>.712</td>
</tr>
<tr>
<td>5</td>
<td>1.241</td>
<td>1.254</td>
</tr>
<tr>
<td>10</td>
<td>1.481</td>
<td>1.483</td>
</tr>
<tr>
<td>15</td>
<td>1.587</td>
<td>1.588</td>
</tr>
<tr>
<td>20</td>
<td>1.661</td>
<td>1.669</td>
</tr>
<tr>
<td>SSIZE=250 NITEMS</td>
<td>r_{zc1s1} r_{zc2s2} r_{zc3s3} r_{zc4s4}</td>
<td>.730</td>
</tr>
<tr>
<td>1</td>
<td>.730</td>
<td>.720</td>
</tr>
<tr>
<td>10</td>
<td>1.542</td>
<td>1.543</td>
</tr>
<tr>
<td>15</td>
<td>1.672</td>
<td>1.678</td>
</tr>
<tr>
<td>20</td>
<td>1.761</td>
<td>1.756</td>
</tr>
<tr>
<td>SSIZE=1000 NITEMS</td>
<td>r_{zc1s1} r_{zc2s2} r_{zc3s3} r_{zc4s4}</td>
<td>.750</td>
</tr>
<tr>
<td>1</td>
<td>.750</td>
<td>.730</td>
</tr>
<tr>
<td>5</td>
<td>1.292</td>
<td>1.307</td>
</tr>
<tr>
<td>10</td>
<td>1.551</td>
<td>1.547</td>
</tr>
<tr>
<td>15</td>
<td>1.684</td>
<td>1.687</td>
</tr>
<tr>
<td>20</td>
<td>1.774</td>
<td>1.767</td>
</tr>
<tr>
<td>SSIZE=5000 NITEMS</td>
<td>r_{zc1s1} r_{zc2s2} r_{zc3s3} r_{zc4s4}</td>
<td>.734</td>
</tr>
<tr>
<td>1</td>
<td>.734</td>
<td>.743</td>
</tr>
<tr>
<td>5</td>
<td>1.300</td>
<td>1.307</td>
</tr>
<tr>
<td>10</td>
<td>1.561</td>
<td>1.553</td>
</tr>
<tr>
<td>15</td>
<td>1.700</td>
<td>1.692</td>
</tr>
<tr>
<td>20</td>
<td>1.782</td>
<td>1.784</td>
</tr>
<tr>
<td>TCORR=.40 SSIZE=50 NITEMS</td>
<td>r_{zc1s1} r_{zc2s2} r_{zc3s3} r_{zc4s4}</td>
<td>.576</td>
</tr>
<tr>
<td>1</td>
<td>.576</td>
<td>.581</td>
</tr>
<tr>
<td>5</td>
<td>1.018</td>
<td>1.009</td>
</tr>
<tr>
<td>10</td>
<td>1.170</td>
<td>1.157</td>
</tr>
<tr>
<td>20</td>
<td>1.306</td>
<td>1.322</td>
</tr>
<tr>
<td>SSIZE=250 NITEMS</td>
<td>r_{zc1s1} r_{zc2s2} r_{zc3s3} r_{zc4s4}</td>
<td>.624</td>
</tr>
<tr>
<td>1</td>
<td>.624</td>
<td>.610</td>
</tr>
<tr>
<td>5</td>
<td>1.065</td>
<td>1.062</td>
</tr>
<tr>
<td>10</td>
<td>1.240</td>
<td>1.240</td>
</tr>
<tr>
<td>15</td>
<td>1.322</td>
<td>1.325</td>
</tr>
<tr>
<td>20</td>
<td>1.355</td>
<td>1.362</td>
</tr>
<tr>
<td>SSIZE=1000</td>
<td>NITEMS</td>
<td>( r_{c_1c_1} )</td>
</tr>
<tr>
<td>-------------</td>
<td>--------</td>
<td>------------------</td>
</tr>
<tr>
<td>1</td>
<td>.614</td>
<td>.612</td>
</tr>
<tr>
<td>5</td>
<td>1.067</td>
<td>1.077</td>
</tr>
<tr>
<td>10</td>
<td>1.247</td>
<td>1.250</td>
</tr>
<tr>
<td>15</td>
<td>1.329</td>
<td>1.336</td>
</tr>
<tr>
<td>20</td>
<td>1.382</td>
<td>1.383</td>
</tr>
<tr>
<td>SSIZE=5000</td>
<td>NITEMS</td>
<td>( r_{c_1c_1} )</td>
</tr>
<tr>
<td>1</td>
<td>.627</td>
<td>.621</td>
</tr>
<tr>
<td>5</td>
<td>1.070</td>
<td>1.064</td>
</tr>
<tr>
<td>10</td>
<td>1.262</td>
<td>1.262</td>
</tr>
<tr>
<td>15</td>
<td>1.335</td>
<td>1.337</td>
</tr>
<tr>
<td>20</td>
<td>1.373</td>
<td>1.375</td>
</tr>
<tr>
<td>TCORR=.80</td>
<td>SSIZE=50</td>
<td>NITEMS</td>
</tr>
<tr>
<td>1</td>
<td>.271</td>
<td>.268</td>
</tr>
<tr>
<td>10</td>
<td>.543</td>
<td>.543</td>
</tr>
<tr>
<td>15</td>
<td>.582</td>
<td>.574</td>
</tr>
<tr>
<td>20</td>
<td>.599</td>
<td>.600</td>
</tr>
<tr>
<td>SSIZE=250</td>
<td>NITEMS</td>
<td>( r_{c_1c_1} )</td>
</tr>
<tr>
<td>1</td>
<td>.308</td>
<td>.308</td>
</tr>
<tr>
<td>5</td>
<td>.535</td>
<td>.540</td>
</tr>
<tr>
<td>10</td>
<td>.613</td>
<td>.607</td>
</tr>
<tr>
<td>15</td>
<td>.638</td>
<td>.636</td>
</tr>
<tr>
<td>20</td>
<td>.657</td>
<td>.654</td>
</tr>
<tr>
<td>SSIZE=1000</td>
<td>NITEMS</td>
<td>( r_{c_1c_1} )</td>
</tr>
<tr>
<td>1</td>
<td>.319</td>
<td>.315</td>
</tr>
<tr>
<td>5</td>
<td>.546</td>
<td>.545</td>
</tr>
<tr>
<td>10</td>
<td>.618</td>
<td>.618</td>
</tr>
<tr>
<td>15</td>
<td>.650</td>
<td>.649</td>
</tr>
<tr>
<td>20</td>
<td>.670</td>
<td>.671</td>
</tr>
<tr>
<td>SSIZE=5000</td>
<td>NITEMS</td>
<td>( r_{c_1c_1} )</td>
</tr>
<tr>
<td>1</td>
<td>.317</td>
<td>.325</td>
</tr>
<tr>
<td>5</td>
<td>.549</td>
<td>.542</td>
</tr>
<tr>
<td>10</td>
<td>.630</td>
<td>.627</td>
</tr>
<tr>
<td>15</td>
<td>.654</td>
<td>.654</td>
</tr>
<tr>
<td>20</td>
<td>.671</td>
<td>.671</td>
</tr>
</tbody>
</table>
APPENDIX D

SAS Program Code of the Statistical Analysis of Study Data
/*--------------------------------------*/
/*   SITUATIONAL REACTION TESTS ANALYTICS  */
/*    GLM Analyses: 3 Traits            */
/*--------------------------------------*/

/* These opening commands need to be run at least once here */
/* here or somewhere to get things started right.          */
goptions reset=all;
*options reset=all device=png gsfname=plt;

options nodate nocenter pageno=1 linesize=100 pagesize=45;
*options nosource nonotes symbolgen;
*options source notes symbolgen;
*options symbolgen mlogic /*mfile*/ mprint nodate nocenter;

/* Change these to suit your environment */
/*--------------------------------------*/
/*Dr. Applegate environment MacBook-15*/
%let BASE = Z:\applegab\Documents\DATA\Students\Cui;
%let BASEPATH = Z:\applegab\Documents\SasFiles\Students\Cui;
libname srt "&BASE";
libname hydata "&BASEPATH\SimData";

/*Dr. Lacefield environment */
*%let BASE = C:\Documents and Settings\lacefieldw\Desktop;
*%let BASEPATH = &BASE\WMU\Department\Students\Hongyan Cui;
*libname srt "&BASEPATH\SRT";
*libname hydata "&BASEPATH\Data";

proc format;
  value groupfmt 1='Calibration Group'
                   2='Validation Group';
run;

proc sort data=hydata.HYfullSim out=HY3;
  by ntraits;
  where ntraits=3;
run;

***********************************************************************;
*****3 traits;
*ods pdf
  file="Z:\applegab\Documents\SasFiles\Students\Cui\DissOUT\MANOVA_1
3Traits.pdf";
*ods graphics on;
proc glm data=HY3;
  class tcorr nitems ssize;
  model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;
  manova h=_all_ ;
  run;
  lsmeans tcorr*nitems|ssize/out=LS3trait;
  run;
quit;
*ods graphics off;
*ods pdf close;
*****plot of 3-way IA;
*goptions reset=global gunit=cell border cback=white
colors=(black blue green red) ftext='Arial' papersize=landscape
xpixels=3600 ypixels=2400 /*xmax=8.5 in ymax=11 in*/
targetdevice=png300 device=png300 gsfmode=replace

lfactor=1;
symbol1 l=1 i=rq v=dot c=black;
symbol2 l=1 i=rq v=dot c=red;
symbol3 l=1 i=rq v=dot c=green;
symbol4 l=1 i=rq v=dot c=blue;
axis1 label=(a=90 h=1.5 "Z Transformed C1S1")
    order=(0 to 2.0 by .25)
    minor=none;
axis2 label=(h=1.5 "TCORR=0.0: Nitems*SSize")
    value=(*"1" *"5" *"10" *"15" *"20")
    minor=none;

proc gplot data=LS3trait;
    plot LSMEAN*nitems=ssize/vaxis=axis1 haxis=axis2;
    where _name_="Zc1s1";
run;
quit;

*****Traditional Simple Effect Post Hocs;
****** each level of TCORR;
proc glm data=HY3;
    class tcorr nitems ssize;
    model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;
    where tcorr=0;
run;
proc glm data=HY3;
    class tcorr nitems ssize;
    model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;
    where tcorr=40;
run;
proc glm data=HY3;
    class tcorr nitems ssize;
    model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;
    where tcorr=80;
run;

*****;

******Traditional Simple-Simple Effect Post Hocs;
****** each level of TCORR and SSIZE;
proc glm data=HY3;
    class tcorr nitems ssize;
    model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;
    means nitems;
    means nitems/line alpha=.0167 LSD;
    where tcorr=0 and SSize=50;
run;
proc glm data=HY3;
    class tcorr nitems ssize;
    model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;
    means nitems;
    means nitems/line alpha=.0167 LSD;
    where tcorr=0 and SSize=250;
run;
proc glm data=HY3;
class tcorr nitems ssize;
model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;
means nitems;
means nitems/line alpha=.0167 LSD;
where tcorr=0 and ssize=1000;
run;
proc glm data=HY3;
class tcorr nitems ssize;
model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;
means nitems;
means nitems/line alpha=.0167 LSD;
where tcorr=0 and ssize=5000;
run;
***** Change level of TCORR;
proc glm data=HY3;
class tcorr nitems ssize;
model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;
means nitems;
means nitems/line alpha=.0167 LSD;
where tcorr=40 and ssize=50;
run;
proc glm data=HY3;
class tcorr nitems ssize;
model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;
means nitems;
means nitems/line alpha=.0167 LSD;
where tcorr=40 and ssize=250;
run;
proc glm data=HY3;
class tcorr nitems ssize;
model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;
means nitems;
means nitems/line alpha=.0167 LSD;
where tcorr=40 and ssize=1000;
run;
proc glm data=HY3;
class tcorr nitems ssize;
model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;
means nitems;
means nitems/line alpha=.0167 LSD;
where tcorr=40 and ssize=5000;
run;
***** Change level of TCORR;
proc glm data=HY3;
class tcorr nitems ssize;
model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;
means nitems;
means nitems/line alpha=.0167 LSD;
where tcorr=80 and ssize=50;
run;
proc glm data=HY3;
class tcorr nitems ssize;
model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;
means nitems;
means nitems/line alpha=.0167 LSD;
where tcorr=80 and ssize=250;
run;
**proc glm** data=HY3;  
    class tcorr nitems ssize;  
    model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;  
    means nitems;  
    means nitems/line alpha=.0167 LSD;  
    where tcorr=80 and SSI=1000;  
run;

**proc glm** data=HY3;  
    class tcorr nitems ssize;  
    model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;  
    means nitems;  
    means nitems/line alpha=.0167 LSD;  
    where tcorr=80 and SS=5000;  
run;

/*  
**** each level of NITEMS;  
proc glm data=HY3;  
    class tcorr nitems ssize;  
    model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;  
    where NITEMS=1;  
run;

proc glm data=HY3;  
    class tcorr nitems ssize;  
    model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;  
    where NITEMS=5;  
run;

proc glm data=HY3;  
    class tcorr nitems ssize;  
    model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;  
    where NITEMS=10;  
run;

proc glm data=HY3;  
    class tcorr nitems ssize;  
    model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;  
    where NITEMS=15;  
run;

proc glm data=HY3;  
    class tcorr nitems ssize;  
    model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;  
    where NITEMS=20;  
run;

**** each level of SSI;  
proc glm data=HY3;  
    class tcorr nitems ssize;  
    model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;  
    where SSI=50;  
run;

proc glm data=HY3;  
    class tcorr nitems ssize;  
    model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;  
    where SSI=250;  
run;

proc glm data=HY3;  
    class tcorr nitems ssize;  
    model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;  
    where SSI=1000;
run;
proc glm data=HY3;
class tcorr nitems ssize;
model Zc1s1 Zc2s2 Zc3s3 = tcorr|nitems|ssize;
where SSIZE=5000;
run;
quit;
*/
*****Curve Fitting Post Hocs;
*ods pdf
file="Z:\applegab\Documents\SasFiles\Students\Cui\DissOUT\MANOVA_Poly
3Traits.pdf";
*ods graphics on;
proc glm data=HY3;
class tcorr nitems ssize;
model Zc1s1 /*Zc2s2 Zc3s3*/ = tcorr|nitems|ssize;
contrast "Item Quadratic, TCORR 0 vs .4, pooled SSize"
  tcorr 0 0 0
  nitems 0 0 0 0 0
  tcorr*nitems 2 -1 -2 -1 2
  -2 1 2 1 -2
  0 0 0 0 0;
* run;
estimate "Item Quadratic @ TCORR=0.0, pooled SSize"
  nitems 2 -1 -2 -1 2
  tcorr*nitems 2 -1 -2 -1 2;
* run;
estimate "Item Quadratic @ TCORR=0.4, pooled SSize"
  nitems 2 -1 -2 -1 2
  tcorr*nitems 0 0 0 0 0
  2 -1 -2 -1 2;
* run;
contrast "Item Quadratic, TCORR 0 vs .8, pooled SSize"
  tcorr 0 0 0
  nitems 0 0 0 0 0
  tcorr*nitems 2 -1 -2 -1 2
  0 0 0 0 0
  -2 1 2 1 -2;
* run;
estimate "Item Quadratic @ TCORR=0.8, pooled SSize"
  nitems 2 -1 -2 -1 2
  tcorr*nitems 0 0 0 0 0
  0 0 0 0 0
  2 -1 -2 -1 2;
* run;
contrast "Item Quadratic, TCORR .4 vs .8, pooled SSize"
  tcorr 0 0 0
  nitems 0 0 0 0 0
  tcorr*nitems 0 0 0 0 0
  2 -1 -2 -1 2
  -2 1 2 1 -2;
* run;
***** Unpacking Item Quadratic for each sample size;
***** TCORR=0.0;
estimate "Item Quadratic @ TCORR=0.0 & 50"
  tcorr 0 0 0
nitems  2 -1 -2 -1 2
tcorr*nitems  2 -1 -2 -1 2
  0 0 0 0 0
  0 0 0 0 0
ssize  0 0 0 0
tcorr*ssize  0 0 0 0
  0 0 0 0
nitems*ssize  2 0 0 0
  -1 0 0 0
  -2 0 0 0
  -1 0 0 0
  2 0 0 0
tcorr*nitems*ssize  2 0 0 0
  -1 0 0 0
  -2 0 0 0
  -1 0 0 0
  2 0 0 0

*  run;
estimate "Item Quadratic @ TCORR=0.0 & 250"
  tcorr  0 0 0
  nitems  2 -1 -2 -1 2
tcorr*nitems  2 -1 -2 -1 2
  0 0 0 0 0
  0 0 0 0 0
ssize  0 0 0 0
tcorr*ssize  0 0 0 0
  0 0 0 0
nitems*ssize  0 2 0 0
  0 -1 0 0
  0 -2 0 0
  0 -1 0 0
  0 2 0 0
tcorr*nitems*ssize  0 2 0 0
  0 -1 0 0
  0 -2 0 0
* run;
estimate "Item Quadratic @ TCORR=0.0 & 1000"
  
<table>
<thead>
<tr>
<th>tcorr</th>
<th>nitems</th>
<th>tcorr*nitems</th>
<th>nitems*ssize</th>
<th>tcorr<em>nitems</em>ssize</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

* run;
estimate "Item Quadratic @ TCORR=0.0 & 5000"
  
<table>
<thead>
<tr>
<th>tcorr</th>
<th>nitems</th>
<th>tcorr*nitems</th>
<th>nitems*ssize</th>
<th>tcorr<em>nitems</em>ssize</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
***** TCORR=0.4;
estimate "Item Quadratic @ TCORR=0.4 & 50"
   tcorr  0 0 0
   nitems 2 -1 -2  2
   tcorr*nitems 0 0 0 0
                 2 -1 -2 -1  2
                 0 0 0 0  0
   ssize 0 0 0 0
   tcorr*ssize 0 0 0 0
                0 0 0 0
   nitemsssize 2 0 0 0
                -1 0 0 0
                -2 0 0 0
                -1 0 0 0
                2 0 0 0
   tcorr*nitemsssize 0 0 0 0
                       0 0 0 0
                       0 0 0 0
                       0 0 0 0
                       0 0 0 0
                       2 0 0 0
                       -1 0 0 0
                       -2 0 0 0

* run;
run;
estimate "Item Quadratic @ TCORR=0.4 & 250"
   tcorr                  0 0 0 0
   nitems                2 -1 -2 -1 2
   tcorr*nitems          0 0 0 0 0
                           2 -1 -2 -1 2
                           0 0 0 0 0
   ssize                 0 0 0 0
   tcorr*ssize           0 0 0 0
                           0 0 0 0
   nitems*ssize         0 2 0 0
                           0 -1 0 0
                           0 -2 0 0
                           0 -1 0 0
                           0 2 0 0
   tcorr*nitems*ssize   0 0 0 0
                           0 0 0 0
                           0 0 0 0
                           0 0 0 0
                           0 0 0 0
                           0 0 0 0
                           0 0 0 0
                           0 0 0 0
                           0 0 0 0
                           0 0 0 0
                           0 0 0 0
                           0 0 0 0
                           0 0 0 0
                           0 0 0 0
                           0 0 0 0
                           0 0 0 0
                           0 0 0 0
                           0 0 0 0
* run;
  estimate "Item Quadratic @ TCORR=0.4 & 5000"
  tcorr 0 0 0
  nitems 2 -1 -2 -1 2
  tcorr*nitems 0 0 0 0
          2 -1 -2 -1 2
          0 0 0 0 0
  ssize 0 0 0 0
  tcorr ssize 0 0 0 0
             0 0 0 0
  nitems ssize 0 0 0 2
             0 0 0 -1
             0 0 0 -2
             0 0 0 -1
             0 0 0 2
  tcorr nitems ssize 0 0 0 0
                  0 0 0 0
                  0 0 0 0
                  0 0 0 0
                  0 0 0 0
                  0 0 0 0
                  0 0 0 0

* run;
***** TCORR=0.8;
estimate "Item Quadratic @ TCORR=0.8 & 50"
tcorr 0 0 0
nitems 2 -1 -2 -1 2
tcorr*nitems 0 0 0 0 0
2 -1 -2 -1 2
ssize 0 0 0 0
tcorr*ssize 0 0 0 0
0 0 0 0
0 0 0 0
nitems*ssize 2 0 0 0
-1 0 0 0
-2 0 0 0
-1 0 0 0
2 0 0 0
tcorr*nitems*ssize 0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0
2 0 0 0
-1 0 0 0
-2 0 0 0
-1 0 0 0
2 0 0 0;
* run;
estimate "Item Quadratic @ TCORR=0.8 & 250"
tcorr 0 0 0
nitems 2 -1 -2 -1 2
tcorr*nitems 0 0 0 0 0
2 -1 -2 -1 2
ssize 0 0 0 0
tcorr*ssize 0 0 0 0
0 0 0 0
0 0 0 0
nitems*ssize 0 2 0 0
0 -1 0 0
0 -2 0 0
0 -1 0 0
run;
estimate "Item Quadratic @ TCORR=0.8 & 1000"
  tcorr 0 0 0
        nitems 2 -1 -2 -1 2
        tcorr*nitems 0 0 0 0 0
                        0 0 0 0 0
                        2 -1 -2 -1 2
        ssize 0 0 0 0
        tcorr*ssize 0 0 0 0
                        0 0 0 0 0
                        0 0 0 0 0
        nitems*ssize 0 0 2 0
                        0 0 -1 0
                        0 0 -2 0
                        0 0 -1 0
                        0 0 2 0
        tcorr*nitems*ssize 0 0 0 0 0
                        0 0 0 0 0
                        0 0 0 0 0
                        0 0 0 0 0
                        0 0 0 0 0
                        0 0 0 0 0
                        0 0 0 0 0
                        0 0 0 0 0
                        0 0 0 0 0
                        0 0 2 0
                        0 0 -1 0
                        0 0 -2 0
                        0 0 -1 0
                        0 0 2 0;
run;
estimate "Item Quadratic @ TCORR=0.8 & 5000"
  tcorr 0 0 0
        nitems 2 -1 -2 -1 2
tcorr*nitems  0 0 0 0 0
      0 0 0 0 0
      2 -1 -2 -1 2
ssize      0 0 0 0

tcorr*ssize  0 0 0 0
      0 0 0 0
      0 0 0 0

nitems*ssize  0 0 0 2
      0 0 0 -1
      0 0 0 -2
      0 0 0 -1
      0 0 0 2
tcorr*nitems*ssize  0 0 0 0
      0 0 0 0
      0 0 0 0
      0 0 0 0
      0 0 0 0
      0 0 0 0
      0 0 0 0
      0 0 0 0
      0 0 0 0
      0 0 0 0
      0 0 0 0
      0 0 0 0
      0 0 0 0
      0 0 0 0
      0 0 0 0
      0 0 0 2
      0 0 0 -1
      0 0 0 -2
      0 0 0 -1
      0 0 0 2;

run;
*ods graphics off;
*ods pdf close;
quit;