COMPARING LIKERT-TYPE AND FORCED-CHOICE FORMATS
FOR ASSESSING PREFERENCE:
A VALIDATION STUDY ON
THE COLLEGE MAJOR PREFERENCE ASSESSMENT (CMPA)

by

Sirui Wu

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Measurement, Evaluation, and Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2021

© Sirui Wu, 2021
The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, a thesis entitled:


submitted by Sirui Wu in partial fulfillment of the requirements for
the degree of Master of Arts
in Measurement, Evaluation and Research Methodology

Examining Committee:

Amery Wu, Associate Professor, Measurement, Evaluation and Research Methodology, UBC
Supervisor

Alison Taylor, Professor, Educational Studies, UBC
Supervisory Committee Member

Yan Liu, Assistant Professor, Measurement, Evaluation and Research Methodology, UBC
Supervisory Committee Member
Abstract

This study validated the College Major Preference Assessment (CMPA) by evaluating its effectiveness of using a combination of Likert-type rating scale and forced-choice formats. Following the argument-based approach to validation by Kane (1992), I articulated the interpretation use argument for CMPA and its corresponding warrants and assumptions. The argument was evaluated by the three research questions based on datasets for Education and Psychology: 1) Did the respondents display response set due to Likert-type format of self-assessment? 2) Can the Likert rounds effectively screen out individuals’ “not preferred” majors? 3) Can the forced-choice rounds effectively identify individuals’ favorite major? A two-dimensional latent trait model was tested using structural equation modeling. The results, by and large, confirmed my hypotheses. Both Likert and forced-choice assessments had a very high level of discrimination power for assessing the major trait Favorite (individuals’ inclination towards a major being their favorite). The Likert rounds were relatively less difficult than the forced-choice rounds. The Likert rounds were more reliable for individuals with a relatively lower level of Favorite, and the forced-choice rounds were more reliable for a relatively higher level of Favorite. Moreover, respondents exhibited a response set, shown by the secondary latent trait, which reflected individuals’ perception about the majors when answering Likert-type items. Nonetheless, the assessment of Favorite was not contaminated by perception. All the findings matched the assumptions and warrants for the argument, showing that CMPA worked in the way it was designed for achieving a high level of effectiveness.
Lay Summary

Assessing individuals’ preference over a large number of college majors is a challenging task due to the problem of individuals’ perception about the majors and the burden on them. College Major Preference Assessment was designed to solve the problems by combining rating scales and forced-choice formats in sequence. This thesis evaluated its effectiveness based on the response data for Education and Psychology. The findings showed that including both formats in a sequence was an effective method for assisting individuals in identifying their favorite college majors. The ratings-scale method was effective for screening out individuals’ not-preferred college majors. The forced-choice method was effective for narrowing down individuals’ favorites. The effect of perception was eliminated almost entirely.
Preface

I am the primary author of the study presented in this thesis. This study used the anonymous, secondary data provided by iKoda Research (https://www.i-koda.com/ikoda-college-website). I am solely responsible for all the work in this study, including literature review and synthesis, research questions identification, as well as data analysis and reports, with clear guidance from my supervisor Dr. Wu. She critically reviewed and revised all the writing in this thesis before I presented it herein.

Please note that Chapter 3 in this thesis was written as a manuscript that is to be submitted to a peer-reviewed journal. Some of its content may be redundant with that in Chapter 1 and 2.
Table of Contents

Abstract .......................................................................................................................................... iii

Lay Summary .................................................................................................................................. iv

Preface ............................................................................................................................................. v

Table of Contents ............................................................................................................................ xi

List of Tables ..................................................................................................................................... x

List of Figures .................................................................................................................................... xi

List of Abbreviations ....................................................................................................................... xii

Acknowledgments ............................................................................................................................ xiii

Dedication ......................................................................................................................................... xiv

Chapter 1: Introduction .................................................................................................................... 1

1.1 Choosing a College Major .......................................................................................................... 1

1.2 College Major Preference Assessment ....................................................................................... 2

1.3 Intended Interpretation and Use of CMPA ............................................................................. 5

Chapter 2: Literature Review ............................................................................................................ 7

2.1 Factors Influencing Students’ College Major Choice ................................................................. 7

   2.1.1 Demographics ..................................................................................................................... 7

   2.1.2 Practical Concerns ............................................................................................................. 8

   2.1.3 Interpersonal Factors ....................................................................................................... 9

   2.1.4 Intrinsic Factors: Ability and Preferences ....................................................................... 9

2.2 Current Tools for Assisting in Identifying Occupational Interests ........................................ 10
2.2.1 Holland’s Theory and Self-Directed Search............................................... 11
2.2.2 Strong Interest Inventory Assessment ......................................................... 13
2.2.3 Non-theory-based Approaches to Assessing Career Interests .................... 15
2.2.4 Using Occupational Personality to Assess College Major Preference .......... 16
2.3 Likert and Forced-choice in Assessing Attitude and Interest ......................... 17
  2.3.1 Likert Scales and its Variation .................................................................. 17
  2.3.2 Forced to Choose from Multiple Dimensions ......................................... 19
2.4 Validation Framework ..................................................................................... 23
2.5 Latent Trait Model in Structural Equation Modeling (LTM-SEM) .................. 26
  2.5.1 Latent Trait Models .................................................................................. 26
  2.5.2 Analytical Approach of LTM-SEM ............................................................ 27

Chapter 3: A Validation Study on the College Major Preference Assessment (CMPA) .... 29
3.1 Introduction ..................................................................................................... 29
  3.1.1 The Design of CMPA ................................................................................. 31
  3.1.2 Validation Approaches: Argument-based Approach .................................. 35
3.2 Research Questions ......................................................................................... 36
  3.2.1 Question-1: Did the respondents display a response set due to the Likert-type
      format of self-assessment. If so, was it associated with the Favorite trait? .......... 37
  3.2.2 Question-2: Can the Likert rounds (with the first two screening steps) effectively
      screen out individuals’ “not preferred” majors? .............................................. 37
3.2.3 Question-3: Can the forced-choice rounds (with the third screening step) effectively identify individuals’ favorite major? .................................................. 38

3.3 Methods ............................................................................................................ 39

3.3.1 Participants ........................................................................................................ 39

3.3.2 The CMPA Assessment Data for the Current Study ........................................... 40

3.3.3 Data Analysis: Latent Trait Model in Structural Equation Modeling (LTM-SEM) .................................................................................................................. 41

3.4 Results ..................................................................................................................... 47

3.4.1 Question-1: Did the respondents display response set due to the Likert-type format of self-assessment. If so, was it associated with the Favorite trait? .............. 47

3.4.2 Question-2: Can the Likert rounds (with the first two screening steps) effectively screen out individuals’ “not preferred” majors? ................................................. 49

3.4.3 Question-3: Can the forced-choice rounds (with the third screening step) effectively identify individuals’ favorite major? ...................................................... 50

3.5 Summary ............................................................................................................... 56

Chapter 4: Conclusion & Discussions ........................................................................ 57

4.1 Conclusion ............................................................................................................... 57

4.2 Contributions ......................................................................................................... 61

4.3 Limitation ................................................................................................................. 62

4.4. Future Studies ........................................................................................................ 63

Bibliography .............................................................................................................. 65
Appendices..........................................................................................................................80

Appendix A. The Parallel Analysis Result for Education and Psychology Dataset..............80

Appendix B. Replicated Results for the Second Half of the Sample .................................81

Appendix C. An Example of Items about Education ..........................................................82
List of Tables

Table 2.1 An Example of Forced-choice Item Formats and the Outcome Scores ............... 20
Table 2.2 IUA for CMPA.................................................................................................. 25
Table 3.1 CMPA IUA: Warrants, Assumptions and Sources for Backing ...................... 36
Table 3.2 Demographics of the Sample ........................................................................... 40
Table 3.3 Model fit Comparison - Information Criteria and Chi-square Difference Tests ..... 48
Table 3.4 Discrimination, Difficulty and Location Parameters for the Seven Rounds ....... 51
Table 4.1 CMPA IUA: Warrants, Assumptions, Research Questions and Main Findings .... 60
List of Figures

Figure 2.1 Holland's Hexagon Model for Vocational Interests (Holland, 1973) .......................... 12

Figure 3.1 Example of Likert and Forced-choice Items ................................................................. 32

Figure 3.2 A Flowchart for the 7-round Procedure of CMPA ...................................................... 34

Figure 3.3 Specification of LTM-SEM for the CMPA Assessment Design and Data ......................... 45

Figure 3.4 Characteristic Curves - the Probability of the Major being Chosen in Each Round

...................................................................................................................................................... 52

Figure 3.5 Information Function for the Seven Rounds, Given Perception ($\theta_2$) = 0 ............... 54
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>Akaike information criterion</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian information criterion</td>
</tr>
<tr>
<td>CMPA</td>
<td>College Major Preference Assessment</td>
</tr>
<tr>
<td>IUA</td>
<td>interpretation and use argument</td>
</tr>
<tr>
<td>LTM</td>
<td>Latent Trait Model, a. k. a. Item Response Theory (IRT)</td>
</tr>
<tr>
<td>LTM-SEM</td>
<td>Latent Trait Model in Structural Equation Modeling</td>
</tr>
<tr>
<td>MLR</td>
<td>maximum likelihood robust</td>
</tr>
<tr>
<td>MLTM</td>
<td>Multidimensional Latent Trait Models, a.k.a. MIRT</td>
</tr>
<tr>
<td>O*NET</td>
<td>Occupational Information Network</td>
</tr>
<tr>
<td>RIASEC</td>
<td>Holland’s theory, including six types: realistic (R), investigative (I), Artistic (A), social (S), enterprising (E), and conventional (C)</td>
</tr>
<tr>
<td>SBIC</td>
<td>Sample-Size Adjusted BIC</td>
</tr>
<tr>
<td>SEM</td>
<td>Structural Equation Modeling</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>frequency–inverse document frequency</td>
</tr>
</tbody>
</table>
Acknowledgments

I would like to express my great thankfulness to my supervisor, Dr. Amery D. Wu, for her patient support and critical guidance throughout the entire journey of completing this thesis.

I would also like to thank iKoda Research for its provision of data for the study in Chapter 3.
Dedication

To my parents
Chapter 1: Introduction

College Major Preference Assessment (CMPA) is a new tool for assisting prospective students in finding their preferred college majors (iKoda Research, 2017). As a newly developed assessment, there is a need for validating its intended interpretation and use. This chapter provides a background for the development of CMPA as well as a general description of the approach this thesis takes for validation.

1.1 Choosing a College Major

Choosing a college major is one of the most important decisions in a person’s life (Beggs et al., 2008). One needs to devote a large amount of time and effort to complete a degree. Many factors may affect the decisions about choosing a college major. Among them, personal interest is one of the most crucial factors. The congruence between interest and college major was found to be associated with college students’ learning motivation, engagement (Leuwerke et al., 2004), and academic performance, as well as college and life-long success (Willcoxson & Wynder, 2010).

Currently, there are two main approaches to assessing interest in college majors. First, students can evaluate their preferences by simply reading the titles on a list of majors. However, students find it hard to judge whether they are interested in a major simply from its name (e.g., Strawn, 2014) before actually studying it. According to a survey by Wright (2018), not knowing the learning content of a major was reported by many participants to be the main reason for
hesitation in choosing or switching to a major. Moreover, the reliability of this approach is questionable because each major is only assessed once.

Another popular approach is to rely on theory-based instruments that were developed for identifying “vocational” personality types. Examples of such practices are using the Strong Interest Inventory (Campbell, 1971) or the Minnesota Interest Blank (Johnson et al., 1975). Once the vocational personality type(s) of a person is (are) identified, it is “mapped” or “matched” to the specific occupations that were judged to be suitable for the identified personality type. The problem with this indirect approach is that, although being connected, “interest in vocation” is not the same as “interest in college majors” after all.

Surprisingly, as far as I know, there had been no tools designed specifically for assisting individuals in directly assessing their preference for college majors. There was a need for a tool to fill in the gap in student guidance. In response to this gap, CMPA was developed with a goal of specifically helping individuals self-assess their preference over 51 college majors and reporting their top-ranked choices.

1.2 College Major Preference Assessment

CMPA took a practice-based approach to its development. The basic idea of practice-based assessment development is to count on practices to describe the phenomenon to be measured (i.e., preference for a major) rather than on a theory for a construct. This approach to assessment development was commonly seen in job-related or task-oriented performance assessments such as assessing teachers’ content knowledge for teaching elementary science (Mikeska et al., 2017) or
assessing nurses’ practical competencies in certification exams (Long et al., 2013). A practice
analysis is usually a starting point for practice-based assessment (Raymond, 2005).

The practice-based approach to assessment development is rare in assessing individuals’
interests. Most of the popular career-related assessments of interest were developed based on a
personality theory, such as the RIASEC occupational personality typology (Holland, 1959;
Holland, 1966). CMPA took a practical-based approach because the purpose of CMPA is not to
assess an abstract construct or typology of career interest. Rather the purpose is to assist the
prospective students in self-assessing their preference over 51 specifics of college majors.

To be as realistic as possible, the item contents of CMPA were extracted from the
descriptions of degree programs and the mandatory courses posted on the official websites of
over 40 Northern American Universities. For each major, a final list of ten phrases was
developed. The phrases were developed as follows. First, a corpus is gathered from the online
texts about the degree programs and course descriptions for the 51 majors. Each major was
treated as a document, then the term (word) frequency-inverse document frequency (TF-IDF) was
computed to identify the terms that best distinguished a major from others (Fan & Qin, 2018).
The 10 words with the highest TF-IDF were then referred back to their original textual contents to
find out the phrases that they were originally embedded in. For example, the term “earthquake”
had a high TF-IDF for the document Geology, and it was embedded in the phrase “how
earthquakes occur.” The identified phrases were characterized by the unique nature, activity,
work, topic, skills, etc. associated with a major. These phrases in their original texts were used as
the contents for creating CMPA items. If a phrase was highly specialist, it was simplified to the
language that a grade-10 high school student could understand.

The task of evaluating 10 key phrases for 51 majors will take a long time and cast a heavy cognitive burden on the respondents. To solve this problem effectively and efficiently, the design of CMPA employs seven rounds of assessment that will only take only about 20 minutes to complete. Each of the seven rounds uses a different set of key phrases. The assessment involves two types of response formats – Likert rating scale and forced-choice format, as well as three screening steps. The delivery of the items is computerized and adaptive as described below.

The first three rounds are in a Likert rating scale format. All items have four scale points (Strongly Disagree=0, Disagree=1, Agree=2, and Strongly Agree= 3). In the first two rounds, the individuals go through all the items for all the majors. Each round uses a different set of key phrases. The first screening step is done upon the completion of the first two rounds. Only the majors that have a mean score higher than a person’s own mean over all majors in the first two rounds will be included in the third Likert round. This screening rule is justifiable because the ultimate goal of this assessment is to find a person’s “favorite” choices. The majors that have a score below one’s own mean score are certainly not one’s “favorite” majors. This screening step is geared to maintaining a high level of efficiency by circumventing unnecessary questions in the third Likert round. The third Likert round uses a new set of key phrases. The same screening rule is applied again, however, this time is based on the average of the responses to all three Likert rounds. The remaining majors for everyone will enter the force-choice assessment.

1In total, there are 51 majors to be assessed by everyone. However, to be exact, only 33 majors are assessed in the Likert rounds. This is because some clusters of specific majors (e.g., Health Science, Kinesiology and Nursing) are assessed under the big umbrella major of General Health. If a general major sustains after the two screening steps in the three Likert rounds, the specific majors will be assessed in the forced-choice rounds.
The next four rounds are in pair-wise forced-choice format. The fourth, fifth, and sixth rounds employ the same random pairwise method for item creation; Each round uses a new set of key phrases of the remaining majors. That is, in each round, two phrases are randomly chosen, without replacement, from an individual’s remaining majors to create forced-choice items until running out of phrases to pair up. The third screening step is applied upon completion of three random-pair forced-choice rounds. The three majors chosen the most frequently (with possible ties) in the three random-pair rounds will enter the last round. In the last round, all possible pairwise comparisons will be made among all of the three (with ties) remaining majors. Figure 1 summarizes the seven-round procedure.

1.3 Intended Interpretation and Use of CMPA

Given the purpose of CMPA – assisting individuals in finding their most preferred college majors – the intended interpretation and use of the results are self-referenced, i.e., comparing the 51 majors within a person for preference. Accordingly, the report of CMPA ranks the three majors (with ties) based on their frequency of being chosen by an individual in the final force-choice round.

CMPA is a new tool for assisting students in finding their preferred majors. As a newly developed assessment, there is yet a report of validity for its intended interpretation and use. In addition, the design of including both Likert and force-choice formats for ensuring the efficiency and effectiveness in assessing preference over a large number of options needs to be evaluated. Therefore, the purpose of this thesis is to conduct an empirical validation study on CMPA by
evaluating its performance of combining the two response formats in sequence. Specifically, this thesis will focus on reporting the findings of two college majors – Education and Psychology – each based on two large samples. These two college majors were chosen because of their relevance to the Master’s degree in the Department of Educational & Counselling Psychology and Special Education.

An argument-based approach to validation proposed by Kane (1992) was adopted. This approach regards validation as a process of developing “a scientifically sound validity argument to support the intended interpretation of assessment scores and their relevance to the proposed use” (AREA, APA & NCME, 2014). This approach has three major steps. It starts with a proposed intended interpretation and use argument (IUA) for an assessment. Then, the assumptions and warrants for how an assessment fulfills the IUA are articulated. The final step is to collect evidence to evaluate the assumptions and warrants and make conclusions about the validity (Kane, 2001; Kane, 2016). In the next chapters, I will articulate my corresponding three steps for the current CMPA validation study. Before that, I will first summarize the related literature, explicating the design purpose of CMPA, and then make a reasonable plan for the validation study.
Chapter 2: Literature Review

The purpose of this Chapter is to provide a background understanding for the validation study presented in the next Chapter. First, I will review the literature regarding the factors, both intrinsic and extrinsic, that can affect individuals’ decisions about choosing a major. Then, I review other popular approaches for assessing career interest, and by comparison, explain the motivation behind the unique design of CMPA. Finally, I will explain my methodology for validation – Kane’s (1992) argument-based approach and the latent trait modeling.

2.1 Factors Influencing Students’ College Major Choice

There is a relatively large body of literature on identifying factors affecting individuals’ college major choices, given that choosing a college major is one of the most important decisions in a person’s personal life. They are summarized below.

2.1.1 Demographics

Previous research has found that demographic variables could influence people’s decisions on their choice of college majors. It was reported that students of the same gender and from the same demographics and social background tend to choose similar fields for their college majors (Ma, 2011). Song and Glick (2004) reported that Asian women were found to be more likely to choose more lucrative college majors than white men. African-American and Hispanic students were less likely to choose pharmacy as their major than Caucasians and Asian-Americans.
(Keshishian et al., 2010). However, it was argued that gender and ethnicity were simply correlates of choice of college major, not the direct factors influencing students’ decisions. For example, Keshishian (2010) found that females had more desire to help others, and it was this desire that led them to choose pharmacy as their major. Further study also found that when other factors, such as academic self-efficacy and preparation, were controlled for, the gender and racial differences decreased sharply (Porter & Umbach, 2006).

As for social-economic status (SES), Ware and Lee (1988) found that students from high SES were more likely to choose science majors, and females from high SES were less likely to choose a major in business (Leppel et al., 2001). However, similar to gender and racial background, some researchers found SES to be less influential when controlling for other factors, like parenting styles (Leppel et al., 2001) and school involvement (Lowinger & Song, 2017).

2.1.2 Practical Concerns

Expected earnings were one of the most often-reported external factors influencing students’ choice of college major (Altonji et al., 2012; Arcidiacono et al., 2012; Montmarquette et al., 2002). In addition to the pivotal role that prospective earnings might play, other financial issues, such as financial aid and debt burdens, significantly increased the possibility for students to choose professional majors instead of humanities and social sciences (Stater, 2011). Other studies, on the contrary, reported that practical concerns were not significant factors among students who chose majors such as psychology and pharmacy (Marrs et al., 2007; Keshishian, 2010). The practical concern only influenced students’ interest in some majors but not others,
showing that it may not be the key factor influencing individuals’ choices of college majors. Instead, interpersonal and personal influences were shown to be more important factors (Edmonds, 2012; Hill & Garner, 1991).

2.1.3 Interpersonal Factors

The apparent external influencers are guidance counsellors and high school teachers who provide career advice to students (Ware & Lee, 1988). Encouragement from close others, including peers and parents, was also found to have a compelling impact (Marrs et al., 2007; Singaravelu et al., 2005). During the process of choosing a major, the involvement and support from close others could make students feel more gratified about their final choice.

Despite these influencers, students still regard themselves to be the most influential person in their choice of college major (Yazici & Yazici, 2010). Most likely, students had thought about their strengths and interests before they sought advice from other people. Two more often reported intrinsic factors are ability and interest.

2.1.4 Intrinsic Factors: Ability and Preferences

Ability was one of the intrinsic factors that students would consider first, especially in choosing math and science as their college majors (Arcidiacono et al., 2012; Bartolj & Polanec, 2012; Maple & Stage, 1991). Malgwi et al.(2005) found that females, in particular, were more concerned about the level of aptitude needed for learning a subject, while men were more concerned about their expected career advancement. Nevertheless, for both genders, these authors
reported that students’ personal interest was a more influential factor than their aptitude.

Without a doubt, personal interest is one of the determining factors (Altonji et al., 2012; Worthington & Higgs, 2004; Yazici & Yazici, 2010). Students who choose their majors mainly out of interest had a higher chance of retention in their choices (Liao & Ji, 2015) and reported a higher level of satisfaction in their campus life and future career (Eun et al., 2013; Xu, 2013). The congruence between major and interest was found to be predictive of persistence. Tracey and Robbins (2006) reported that among those who were less interested in academics, those who chose their major out of interest were less likely to drop or quit courses.

In summary, the extant literature has reported extrinsic and intrinsic factors or correlates for students’ choice of college majors. Personal interest is no doubt the key factor to consider. Given the positive evidence that personal interest has on satisfaction in college life and career success, it is surprising that, to my best knowledge, there had been no tools available that could “directly” assist individuals in identifying their preference for college majors before CMPA was introduced. The most frequently resorted to and heavily researched tools, such as the Self-Directed Search (Holland & Rayman, 1984) and Strong Interest Inventory (Strong, 1935), were not designed specifically for such purpose. In what follows I will review these “indirect” tools because of their popularity in research and practice and, hence, their contributions to the current literature.

2.2 Current Tools for Assisting in Identifying Occupational Interests

As said, there had been no tools specializing in assessing personal preference for college majors. Instead, generic assessments that aim to assess “career” interests such as Self-Directed
Search (Holland & Rayman, 1984) and Strong Interest Inventory (Strong, 1935) were commonly adopted to handle the task (Hansen & Tan, 1992). This is perhaps due to the reported relationship between students’ choice of college majors and their future career path (Eun et al., 2013). Most of these tools were developed based on Holland’s well-known RIASEC theory, which will be briefly reviewed below.

2.2.1 Holland’s Theory and Self-Directed Search

In Holland’s RIASEC theory (Holland, 1959; Holland, 1966), there are six types of occupational personalities that individuals can belong to: realistic (R), investigative (I), artistic (A), social (S), enterprising (E) and conventional (C). They are abbreviated as RIASEC. Each of the six types of personality was shown to be more prone to choose certain occupations. For example, soldiers tend to be the Realistic type and artists tend to be the Artistic type. Holland (1985) further showed that the six personality types can be assigned to six vertices of a hexagon in a specific order (see Figure 2.1). The two types adjacent to a given type were more closely related to that given type, compared to that in the opposite vertex. For example, Realistic is more closely related to Investigative and Conventional but distant from Social. Many studies tested Holland’s hexagonal relationships and found that, in general, Holland’s theory worked well for diverse populations and working environments (Holland & Gottfredson, 1992; Prediger & Vansickle, 1992; Swanson, 1992).
Holland’s typology is a widely cited and used framework for assessing occupational interests. Many inventories were built based on or referred to this framework, including Self-Direct Search (Holland, 1958), Campbell Interest and Skill Survey (Campbell, 1971), Career Assessment Inventory (Johansson, 1975), and O*NET Interest Profiler (Lewis & Rivkin, 1999). Some other inventories, like Strong Interest Inventory (Strong, 1935), were established before Holland’s work based on the Minnesota Multiphasic Personality Inventory (Hathaway & Mckinley, 1942), but also adopted Holland’s typology in later revisions (Donnay & Borgen, 1996; Strong, 1964). Among them, the most well-known are Holland’s own Self Direct Search and the Strong Interest Inventory.

The Self-Direct Search is an extension of the RIASEC theoretical framework, aiming to help individuals narrow down their interests in different occupations. A respondent is first required to finish three Likert scales to assess their preferred activities, occupational interests, and competencies. Then every respondent obtains a three-letter code based on their responses to these three scales. This 3-letter code represents the respondent’s three major occupational personality types. Then, the respondent can use the 3-letter code to identify the matched occupations from...
1,156 occupations listed in a booklet (Holland, 1985).

Holland’s theory provides a useful framework for assessing individuals’ vocational interests; however, it relies on the assumption that individuals’ vocational interests are closely related to their personalities. This assumption was challenged by some empirical evidence. For instance, Morrow (1971) grouped students according to Holland’s typology and found that college major satisfaction was significantly related to students’ personality types for students taking a major in math, but not for students taking a major in sociology. Also, Upperman and Church (1995) found a common tendency of being “Realistic” among the soldiers in the Army despite that they actually had very different occupations. These findings pointed out a potential drawback in Holland’s theory.

Other reported problems of Holland’s approach include overrating and cultural diversity. For instance, Damarin (1998) found students overly self-reported their interest and resulted in a high probability of misidentification. Moreover, Rounds and Tracey (1996) found a lack of cross-culture equivalence in Holland’s hexagon model.

2.2.2 Strong Interest Inventory Assessment

Strong Interest Inventory holds a basic assumption: People working within the same occupation share similar likes and dislikes on both vocational interests and non-vocational activities (Hansen & Campbell, 1985). The original Strong Interest Inventory only contained subscales for ten occupations. The number of occupations increased to 130 after a few revisions (Donnay et al., 2004).
The current Strong Interest Inventory (Herk & Thompson, 2012) has four subsections: General Occupational Themes, Basic Interest Scales, Occupational Scales, and Personal Style Scales. The General Occupational Themes bear on Holland’s theory to distinguish six types of interest and give the broadest classifications among the four subsections in the Inventory. The Basic Interest Scales include 23 scales that assess the respondents’ interest in 23 basic career areas. The Occupational Scales further extend the classification into 207 typical occupations. The Personal Types Scales add four scales in order to measure four additional work-related personalities: Work style, learning environment, leadership style, and risk-taking. (Herk & Thompson, 2012).

As one of the most widely used tools for assessing career interest, Strong Interest Inventory was thoroughly evaluated and showed satisfactory predictive validity, both in the context of assessing occupational and academic interests. Hanna and Rounds (2020) reported that the Strong Interest Inventory had a hit rate of 53.8% in predicting people’s career choices. Miller (2010) reported that the Basic Interest Scales showed a level of 60% accuracy in predicting students’ interests in 13 groupings of college majors.

Despite its current popularity and satisfactory accuracy in prediction, a few drawbacks of the Strong Interest Inventory have been noted. Firstly, it is too complex and time consuming. The entire inventory has more than 300 items, taking at least 40 minutes to finish (Hansen, 2019). Secondly, the necessity of including all four sections was questioned (Fox, 1995; Mueller et al., 2010), especially the section of the Personal Types Scales (Staggs, 2004). Furthermore, the item contents in the General Occupational Themes were criticized for being maladaptive to the rapidly
and ever-changing nature of the occupational world (Day & Rounds, 1997).

2.2.3 Non-theory-based Approaches to Assessing Career Interests

Besides personality type, there are some other approaches recently developed to identify students’ interest in different occupations.

Nye et al. (2019) developed an adaptive vocational interest diagnostic tool. With a military target group, they first analyzed the existing military interest data to determine 20 general important dimensions. Then, items were developed to assess each dimension. A large Army soldiers’ sample was recruited to estimate the items’ parameters, based on which, a fixed and adaptive version of vocational interest diagnostic assessment was assembled. Although this tool was designed for a specific population, its data-driven (non-theory-based) approach to assessment development provided a new and promising way for assessing interest.

Occupational typology theory and its extensions are still the dominant frameworks for assessing career interest despite new approaches that have emerged recently. As useful as they are, theory-based approaches to interest assessment, unfortunately, cannot assist individuals in “directly” finding their preferred occupations. That is, these assessments output a personality profile for each person, but the link between the profile and the available occupations in the job market needs to be further matched (Wille et al., 2014).

Indeed, certain job categories are more suitable for certain occupational personalities, and personality types could predict “broad” occupation categories with about 50-60% accuracy. However, in order to match the occupational personalities to the “specific” real-world
occupations, the occupations in the current job markets need to be classified into “being suitable” for each of the personality types. This task could be extremely onerous and difficult in practice and adds room for misclassification due to personality-occupation mismatch. Unfortunately, identifying specific occupations by matching them to occupational personality theory has its inevitable obstacles, in particular when the occupations in today’s job market are disappearing and emerging faster than ever (Robst, 2007).

2.2.4 Using Occupational Personality to Assess College Major Preference

Currently, the Occupation Information Network (O*NET, 2020) is probably the most well-established online platform for integrating assessments and matching between personalities, college majors, and job categories based on the RIASEC codes. For example, individuals who graduated with a degree in “psychology” are being matched to the job category of clinical psychologists suggested for the ISA personality type (Investigative, Social, Artistic) and a job category of counselling psychologist for the SIA personality type (Social, Investigative, Artist; See O*NET, 2020). To date, no classification reliability and/or prediction accuracy are reported for O*Net efforts in assessment and matching.

Others have reported that occupational personalities showed satisfactory predictive validity for “broad” categories of college majors. Miller (2010) reported that the Basic Interest Scales showed a level of 60% accuracy in predicting students’ interests in 13 “groupings” of college majors. However, a much lower success rate of only about 35% was reported when predicting students’ exact academic majors.
This low match between occupational personality and specific college majors highlights the gap for an alternative assessment program that is designed to directly rank an individuals’ preference for specific college majors(s). CMPA is an assessment program that tries to fill this gap.

CMPA aims to directly assess individuals’ preferences over 51 college majors. To do so effectively and efficiently, it embarks on seven rounds of assessments that take, on average, 15 to 20 minutes to complete. It first uses the Likert scale to screen out the majors that are not preferred and then uses the forced-choice formats in the remaining four rounds to narrow down to the top three preferred majors. The design of employing both the Likert and forced-choice formats is to ensure a high level of efficiency and effectiveness in assessment.

2.3 Likert and Forced-choice in Assessing Attitude and Interest

Likert rating scales and forced-choice questions are two popular methods in measurements. Each has its own strengths and shortcomings. Understanding the characteristics of these two assessment formats helps to take advantage of their strengths and mitigate their shortcomings.

2.3.1 Likert Scales and its Variation

The Likert rating scale was first introduced to measure attitudes (Likert, 1932); however, it has become one of the most widely used formats for assessing all sorts of human mental states (Albaum, 1997). In the original version of the Likert scale, respondents were required to indicate their agreement to a set of statements, on a five-point scale ranging from, say, strongly disagree to
strongly agree that are usually coded as 1 to 5. The sum score across all the items represents the degree of positive/negative attitude towards the measured object (Salkind, 2007).

Several variations of the Likert Scale have been introduced. One of the major modifications was the number of response categories. The extant evidence has shown that an increase in the number of categories on the rating scale made the scale more sensitive. However, too many categories made it hard for the respondents to distinguish between the categories. This leads to a common use of four to seven categories on the rating scale (Chomeya, 2010; Leung, 2011).

The most reported advantages of the Likert scale are that it is relatively easy to develop and the data are easy to handle. The sum score is easy to compute and understand and can be used to compare across individuals. In general, responses from Likert scales are reliable (Chang, 1994; Matell & Jacoby, 1971; Lozano et al., 2008).

Despite its popularity, research has documented the drawbacks of Likert-type scales. For a Likert scale with an odd number of categories, the respondents may choose a middle point for a bunch of unclear and undesired reasons: the statements are unclear or not applicable, the respondents are undecided, and/or have a tendency towards a neutral point (Raaijmakers et al., 2000). In addition, some systemic ways of responding to Likert scales, such as acquiescence, social desirability, and extreme response styles, have also been reported (Clarke III, 2006; Ross & Mirowsky, 1984; Welkenhuysen-Gybel et al., 2003). This tendency to answer questions in a systematic manner that is unrelated to their content is referred to as response set, according to APA Dictionary of Psychology (VandenBos (Ed.)., 2007). These drawbacks of Likert-type scales are often ignored or tolerated because of their simplicity in test development, user experience, and
2.3.2 Forced to Choose from Multiple Dimensions

Another popular format of assessing preference over multiple objects is multi-dimensional (a.k.a. multiple objects) forced-choice questions where the respondents are forced to choose from or rank a block of two more phrases or statements. Each of the phrases in the block represents one of the dimensions being assessed (e.g., college majors). A typical item of a multidimensional forced-choice assessment is shown in Table 2.1, where the four statements represent the dimensions of Compliance, Steadiness, Influence, and Dominance respectively.

In the original version of the multi-dimensional forced-choice test, respondents needed to choose a statement that is “most like me” and “least like me.” In the pairwise version, the respondents need to compare the two statements in an item and choose the statement that best describes their inclination. When there are more than two statements in an item, the respondents are required to rank the statements.

Different forced-choice formats have different ways of scoring to reflect the respondents’ preference for or attitude towards the dimensions. In the original version of the multidimensional forced-choice test, the two most extreme statements are assigned with a highest/lowest score comparing all other items (2 and 0), and the remaining items are assigned a moderate score (1). In the format of the pairwise comparison, if a participant chooses statement-A instead of statement-B, then statement-A is assigned 1 point and statement-B is assigned 0 points. After all the comparisons, total scores for each statement are calculated for all pairwise comparisons. In the
ranking format, the rank for each statement represents the (reverse) level of the preference.

**Table 2.1**

An Example of Forced-choice Item Formats and the Outcome Scores

<table>
<thead>
<tr>
<th>Original Format</th>
<th>Choose a statement that is “most like me” and “least like me.”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
<td>Statement</td>
</tr>
<tr>
<td>1</td>
<td>I manage to relax easily.</td>
</tr>
<tr>
<td>2</td>
<td>I am careful over details.</td>
</tr>
<tr>
<td>3</td>
<td>I enjoy working with others.</td>
</tr>
<tr>
<td>4</td>
<td>I set high personal standards.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pairwise Format</th>
<th>Choose a statement that is “more like me” between two options.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
<td>Option A</td>
</tr>
<tr>
<td>1</td>
<td>I manage to relax easily.</td>
</tr>
<tr>
<td>2</td>
<td>I manage to relax easily.</td>
</tr>
<tr>
<td>3</td>
<td>I manage to relax easily.</td>
</tr>
<tr>
<td>4</td>
<td>I am careful over details.</td>
</tr>
<tr>
<td>5</td>
<td>I am careful over details.</td>
</tr>
<tr>
<td>6</td>
<td>I enjoy working with others</td>
</tr>
</tbody>
</table>

**Computing the Score**

<table>
<thead>
<tr>
<th>Item1</th>
<th>Item2</th>
<th>Item3</th>
<th>Item4</th>
<th>Item5</th>
<th>Item6</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>-</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

**Ranking Format**

Give a rank for the statements below from “Most like me” = 1 to “Least like me” = 4.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Ranks</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 I manage to relax easily.</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>2 I am careful over details.</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3 I enjoy working with others.</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>4 I set high personal standards.</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

Regardless of formats, the total score across the multiple dimensions will be the same (original version = 4, pairwise = 6, and ranking = 7 for Table 2.1) for all respondents. The different forced-choice formats are simply different ways of distributing the total scores within a
person, hence indicating the relative importance of the dimensions to a person.

A forced-choice format was proposed to mitigate the aforementioned drawbacks of Likert-type scales. This is because when the respondents are “forced” to choose or rank the provided options, it is less likely for them to fake their answers (Ford, 1964). Many researchers have also found that the forced-choice format is more reliable and can eliminate the response set like social desirability (Heggestad et al., 2006; Lee et al., 2019; Salgado & Lado, 2018).

The much-appreciated advantage of the forced-choice method does not come with no price. The first problem resides in the scoring. To explain the drawbacks of forced-choice format scoring, imagine there are five dimensions, A to E, each represented by statement-A to statement-E, respectively. Let’s further assume that all possible pairwise comparisons are included in the scale, resulting in a total of 10 items, each with two statements.

Firstly, it is impossible to rank the individuals by the total scores over the 10 items because the total scores will be the same and equal to 10 for all respondents (Brown & Maydeu-Olivares, 2013). After all, the respondents are forced to choose either statement in each of the 10 items.

Secondly, even when the scoring is done separately for each dimension, say dimension A, the scores on dimension A are not comparable among the respondents. This is because when the two respondents score 2 points on dimension A because both chose statement-A twice from the 10 pair-wise comparisons, the meaning of “A=2” is different for the two respondents. Take an example to illustrate this point. For Mary, statement-A might be chosen twice when it is paired with statement B and statement C, but, for John, statement-A might be chosen twice when it is compared with statement D and statement E. For this reason, it is invalid to rank the individuals
by their dimension scores. This problem is exacerbated when the number of dimensions (statements) is large as well as when the process is repeated multiple times, each with a different new set of statements, for the sake of increasing reliability. The cross-person incomparability is a result of the ipsative (of the self) nature of the forced-choice format, meaning that the dimensions scores are meant to be compared within a person. That is, it is only meaningful to compare the score of one dimension to those of ones’ own scores in other dimensions.

The third problem with forced-choice is that when the number of dimensions is high (>10), the forced-choice format is cognitively difficult and inefficient timewise. Take CMPA for instance, when 51 majors need to be assessed using the original format, it is difficult for the respondents to choose their most- and least-liked statements from a list of 51 statements. In the pairwise format, the number of items will increase drastically to 1,250. In the ranking format, it is cognitively taxing, if even possible, for the respondents to rank 51 statements. The problems of cognitive burden and time inefficiency will be exacerbated because the number of items will be increased by k times if the process is to be repeated k times to ensure reliability.

The fourth problem is concerned with the conditional dependence (correlation) between the responses. Take an example of four statements in a block. The process of ranking the four statements will be interdependent. That is, if Mary ranks statement-2 as her top, she can only continue to rank the remaining three statements as the 2nd, 3rd, or 4th. If she continues to rank statement-4 as the 2nd, she only needs to pick one from the remaining two as the 3rd, and the unchosen statement must be the last. In this sense, the forced-choice ranking responses are dependent. In contrast, Likert-type responses are not dependent. If each of the four statements is
laid out as a Likert-type item on a scale, an individual’s response to Item-1 is independent of those of items-2, -3, and -4, given their true inclination towards each dimension. That is, a respondent’s answers to one Likert item will not influence his/her/their answers to the others.

Item response dependence is a major violation of the assumption of the mainstream test theory, including not only the classical test theory but also factor analysis, item response theory, and latent class analysis, rendering these statistical measurement models unsuitable for forced-choice data. In Chapter 3, I will explain how I resolve this problem by modeling the outcomes at the end of seven rounds (rather than the outcomes of the items responses) and using a latent trait model in the structural equation modeling framework tailored to CMPA’s design.

Despite their strengths and limitations, the Likert or forced-choice formats did not seem to influence the internal structure of the relationship between the factors (latent dimensions) and the observed indicators (statements). Buxton (1966) compared forced-choice and Likert formats that assessed the same construct while controlling the item contents and the respondent sample. For both cognitive and non-cognitive assessments, his results showed that the type of response data shared a similar factor structure with similar item-factor relationships. His findings have been supported by some recent studies (e.g., Geldhof et al., 2015; Miller, et al., 2018).

2.4 Validation Framework

Kane’s (1992) theory of argument-based validation was adopted for the present study. Kane attested that validity is not a property of the tests themselves, but test score interpretations and uses. The process of validation, in Kane’s view, includes three steps. It starts with an
interpretation and use argument (IUA) according to the purpose of an assessment. Then, the assumptions and warrants for how an assessment fulfills the IUA are articulated. Finally, evidence is collected to evaluate the argument and make conclusions about the validity for the (Kane, 2001; Kane, 2016). The validity arguments can be assessed by prior knowledge such as known theories or established practices and/or by empirical evidence (Kane, 2013).

Compared with other validity theories, Kane’s argument-based approach provides a very general and flexible framework for validation and does not prescribe any recipe for how validation should be done (Jacobson & Svetina, 2019; Sireci, 2013). An argument-based approach to validation is most suitable when the IUA for the test purpose can be clearly articulated (Lavery et al., 2020).

CMPA was created to assist individuals in effectively and efficiently identifying their three favorite college majors from 51 college majors that are commonly offered in Northern American colleges and universities. The interpretation of the results is ipsative (i.e. of the self) and the results should be used for personal reference. According to this purpose, its IUA is shown in Table 2.2 for your reference. This study will mainly focus on Warrant D, which I will detail in Chapter 3 and explain how I use empirical data to evaluate it.
Table 2.2

IUA: CMPA

<table>
<thead>
<tr>
<th>Warrants</th>
<th>Assumptions</th>
<th>Sources for backing/justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. CMPA directly and authentically assesses an individual’s preference over 51 specific college majors.</td>
<td>It is possible to directly assess an individual’s preference over 51 specific college majors.</td>
<td>-The contents of CMPA are based on a corpus of natural text data that contain the authentic knowledge and practices that are well established and documented by over 40 universities and colleges. -The phrases used to create items are identified by finding the most unique words for each major using the TF-IDF measure. -The procedure is designed to directly assess an individual’s preference for 51 majors.</td>
</tr>
<tr>
<td>B. CMPA can provide results for ipsative interpretation and use.</td>
<td>It is possible to design an individually tailored assessment to compare a person’s preference over 51 majors.</td>
<td>-Round-3 to Round-7 assessments are tailored to each respondent by the three self-referenced screening criteria. -The final report only ranks an individual’s top choices and does not provide any norm referenced scores such as percentile or Z-score.</td>
</tr>
<tr>
<td>C. CMPA is efficient for ranking an individual’s favorite majors.</td>
<td>It is possible to design an assessment that can identify a person’s three favorites from 51 college majors within the time length that is reasonable to most respondents.</td>
<td>-With the three screening steps and a combination of Likert and pairwise forced-choice formats, a person only has to answer approximately 109 items, that take about 18.16 minutes to rate over 51 college majors and identify the three favorites. -If only Likert items were used and no screenings were conducted, the number of items will rise to 51×7= 357 and will take approximately 59.5 minutes to rate over 51 college majors and identify the three favorites. -If only pairwise forced-choice items were used and no screenings were conducted, the number of items will rise to (51×50)/2 ×7 = 8925 and will take approximately 148.78 minutes to rate over 51 college majors and identify the three favorites.</td>
</tr>
<tr>
<td>D. CMPA is effective for ranking an individual’s favorite majors.</td>
<td>The combination of Likert and forced-choice formats, as well as the three self-referenced screening steps, can achieve a high level of effectiveness in ranking an individual’s favorite majors.</td>
<td>-CMPA can provide reliable rankings because each respondent answers approximately109 tailored items in the 7-round procedure. -CMPA can provide valid rankings because: a. The item contents are authentic, specific, and unique to each of the 51 majors. b. The Likert rounds and the first two screening steps serve as a simple yet comprehensive way of screening out &quot;not preferred&quot; majors from 51 majors. c. The four forced-choice rounds and the third screening step serve to prevent possible response styles in answering Likert items and/or response sets because of one’s positive or negative perceptions towards a major. d. The four forced-choice rounds repeatedly and progressively push the respondents to choose between two “preferred” majors.</td>
</tr>
</tbody>
</table>

- In total, there are 51 majors to be assessed by each individual. However, to be exact, only 33 majors are assessed in the Likert rounds. This is because some clusters of specific majors (e.g., Health Science, Kinesiology, and Nursing) are assessed under the big umbrella major of General Health. If a general major sustains after the two screening steps in the three Likert rounds, the specific majors will be assessed in forced-choice rounds.
- Because each individual will have a different number of items, the total number of items is computed based on standard conditions assumed as the following: There are 33 items for the first two Likert rounds (see note a above). Assuming half of the majors pass the first screening, there will be 16 items for the third Likert round. Also, assuming only 16 majors pass the second screening, it will end up with eight pairwise forced-choices items for each of the fourth, fifth, and sixth rounds. Because only three majors will be kept for the final force-choice round after the third screening, there will be three items when all three majors are pairwise compared. Under this reasonably assumed condition, there are 33 + 33 +16 + 8 + 8 + 8 + 3 = 109 items in total over the seven rounds of CMPA. 
- The time needs to complete the assessment is computed based on the assumption that each item takes 10 seconds to complete, and 51 majors are assessed in seven rounds.

Empirical Data
2.5 Latent Trait Model in Structural Equation Modeling (LTM-SEM)

2.5.1 Latent Trait Models

The latent trait model (LTM), named by Birnbaum (1968), and also known as item response theory (IRT), is a model for identifying the latent trait(s) underlying observed data. It assumes that individuals’ measured variables can be predicted by their unobservable traits (Hambleton & Cook, 1977), so we can use measured variables to indicate the latent trait. The latent trait is usually, but necessarily, linked to the measured variables via a logistic function. The parameter estimates of the latent trait model allow the researchers to evaluate the performance of the measured variables in indicating the latent trait (Thorpe, et al., 2007), as well as the level of information (certainty, reliability) in estimating the latent trait.

The latent trait model is a popular choice for modeling ordinal data like those collected by Likert-type scales. To date, the most established and suitable variations for Likert-type data are the graded response model (Samejima, 1997), generalized graded response model (Roberts & Laughlin, 1996), rating scale model (Rasch, 1961), and generalized rating scale model (Muraki, 1992). The major differences between these models are the assumptions about the distance between the points on the Likert scale.

Different LTMs have been proposed for modeling forced-choice data. Zinnes and Griggs (1974) made an early attempt to use LTM to model pairwise forced-choice items. Over the years, psychometricians continued to suggest new alternative ways for handling the scoring and analyzing for forced-choice data, including Andrich’s IRT model (Andrich, 1989), the Multi-
unidimensional pairwise-preference model (MUPP, Stark et al., 2005), and the Thurstonian IRT model (Brown & Maydeu-Olivares, 2011).

Evidence from these psychometrics showed that LTM is a useful validating technique. For this reason, this study adopts an LTM approach to analyzing CMPA data. However, none of the LTM models cited above were chosen for the present study for three reasons. Firstly, those prescribed models are not suitable for CPMA, which assesses a high number of dimensions (51 college majors). Secondly, those prescribed models can only handle either Likert or forced-choice data. Thirdly, the response dependence in the forced-choice items violates the assumption of most LTM models.

To solve the three problems, I conducted a latent trait model within the broader modeling framework of structural equation modeling (SEM) and referred to it as LTM-SEM hereafter.

2.5.2 Analytical Approach of LTM-SEM

Like a factor analysis, a latent trait model, also known as item response theory, can be embedded in a structural equation model (SEM, Muthen, 1983), tailored to specific research questions for validation. An example is a study by Muthén et al. (1991), where he examined the threat to validity due to the items being sensitive. He specified an item response theory in broader structural equation modeling so that he can include a predictor of opportunity-to-learn to examine its effect on item performance in eighth-grade math.

I chose LTM-SEM because of its flexibility in modeling the dichotomized outcome at the end of each round of assessment. In particular, LTM-SEM will be used to model two latent traits.
The first is the major latent trait of Favorite, which is individuals’ inclination towards a major being their favorite. The Favorite latent trait was indicated by the outcomes of all seven rounds (outcome = 1 if response = Agree and Strongly Agree for the Likert rounds, and if a major is chosen for the forced-choice rounds). The second is a minor latent trait of Perception, indicated only by the Liker-round results. In addition, LTM-SEM allows me to specify the dependent relationships in the Likert rounds. The method of LTM-SEM will be detailed in Chapter 3.

In summary, CMPA is designed specifically to assist individuals in assessing their preference for college majors. The inclusion of both Likert and forced-choice formats is geared to mitigating the negative effects while taking advantage of each format so that a large number of majors can be effectively and efficiently assessed. Kane’s argument-based validation approach will be used for current validation. By way of LTM-SEM, this study will focus on evaluating one specific warrant for the IUA: The combination of Likert and forced-choice formats as well as the three self-referenced screening steps of CMPA can achieve a high level of effectiveness in assisting individuals’ in finding their favorite majors.

In the next Chapter, I will articulate the assumptions for the warrant and the corresponding sources for backing. The design of the validation study and the findings will also be reported.
Chapter 3: A Validation Study on the College Major Preference Assessment (CMPA)

This Chapter was written as a manuscript that is to be submitted to a peer-reviewed journal. Some of its content may be redundant with that in Chapter 1 and 2. The purpose of this chapter is to evaluate the effectiveness of CMPA for ranking an individual’s favorite majors based on real data. A high level of effectiveness (high validity) should be evidenced by a high level of coherence between the empirical data and the results of the statistical model, specified according to the CMPA assessment design. To do so, I will explicate three research questions and my corresponding hypotheses about the answers based on the prior knowledge, the theoretical framework, and CMPA’s assessment design. After that, I will analyze the CMPA data for Education and Psychology, mainly using the LTM-SEM to see whether the results are consistent to my hypotheses. If there is a high level of coherence, it can be concluded that there is a high effectiveness (validity) in the intended interpretation and use for CMPA scores.

3.1 Introduction

Choosing a college major is one of the most important yet challenging decisions in life. The extant literature has reported extrinsic and intrinsic factors or correlates for students’ choice of college majors. Among them, personal interest is found to be a pivotal one. Students who choose their majors mainly out of interest have a high chance of retention in their choice, leading to a higher level of satisfaction in their campus and future career.
However, students found it hard to judge whether they were interested in a major simply by its name (e.g., Strawn, 2014) before actually learning it. According to a survey by Wright (2018), not knowing the learning content of a major was reported by many participants as the main reason for hesitation in choosing or switching to a major. Moreover, the reliability of this approach to assessing interest is questionable because each major is only assessed once.

Another popular approach is to rely on theory-based instruments that were developed for identifying “vocational personalities” types. Examples of such practices are the Strong Interest Inventory (Campbell, 1971) or Minnesota Interest Blank (Johnson et al., 1975). Once the vocational personality type(s) of a person is identified, it is “mapped” or “matched” to the specific occupations that were pre-judged to be suitable for the specific personality type. The problem with this indirect approach via matching is that the possible inaccuracy in matching. Moreover, although being connected, a person’s choice in “vocation” is not necessarily equated with their choice in “college majors”, after all.

Given the undoubted evidence that personal interest has on satisfaction in college life and career success, it is surprising that there had been no tools available that can “directly” assist individuals in identifying their preference for college majors. In response to this gap, the College Major Preference Assessment (CMPA, iKoda, 2017) was developed with a goal to specifically help individuals self-assess their preference for college majors and report their top choices.
3.1.1 The Design of CMPA

The purpose of CMPA is to assist individuals in identifying their three (with ties) favorite college majors effectively and efficiently from 51 college majors that are commonly offered in Northern American colleges and universities. The interpretation of the results is ipsative (of the self), and the results are used for personal reference. The target population is any prospective students who intend to take a post-secondary degree in Northern American colleges and have at least a grade-10 level of English language ability.

For each major, a list of ten phrases was prepared. The phrases were identified as follows. First, a corpus is gathered from the online texts describing the degree programs and course descriptions of over 40 North American colleges and universities. Each major was treated as a document and a term (word) frequency-inverse document frequency (TF-IDF) was computed to identify the terms that best distinguished a major from the others (Fan & Qin, 2018). The 10 words with the highest TF-IDF were then referred back to their original textual contents to find out the phrases that they were originally embedded in. For example, the term “earthquake” had a high TF-IDF for the document Geology and it was embedded in the phrase “how earthquakes occur.” The identified phrases were characterized by the unique nature, activity, work, topic, skills, etc. associated with a major. These phrases in their original texts were used as the contents for creating CMPA items. If a phrase was highly specialist, it was simplified to the language that a grade-10 high school student could understand.

With 10 phrases for each major, the total number of phrases summed to be 510. They are
stored in a database, each labeled with its tagged major so that they can be used to form items to be delivered over the seven rounds of assessments. Each round of assessment uses a different set of phrases for the majors.

The first three rounds are in a Likert-type rating scale format, thus, named as “Likert rounds.” All items have the same stem that says “I want to learn about …,” followed by a phrase for a specific major. An example Likert-round item for Education is given in Figure 3.1.

**Figure 3.1**
*Example of Likert and Forced-choice Items*

<table>
<thead>
<tr>
<th>Likert Item:</th>
<th>I want to learn about…</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>designing teaching activities.</strong></td>
</tr>
<tr>
<td></td>
<td>○ Strongly Disagree</td>
</tr>
<tr>
<td></td>
<td>○ Disagree</td>
</tr>
<tr>
<td></td>
<td>○ Agree</td>
</tr>
<tr>
<td></td>
<td>○ Strongly Agree</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Forced-choice Item:</th>
<th>I would rather spend time learning about …</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>○ <strong>helping children achieve</strong></td>
</tr>
<tr>
<td></td>
<td>○ <strong>how a person's memory works.</strong></td>
</tr>
</tbody>
</table>

All the items in the Likert rounds have four scale points (Strong Disagree = 0, Disagree = 1, Agree = 2, and Strongly Agree = 3). In the first two rounds, a person goes through all the items for all majors. However, in the third round, the majors that have a mean score less than a person’s own mean over all majors, calculated based on the responses to the first two rounds, will be regarded as “not preferred” and excluded from the third Likert-round (and the rest of the assessment) for the individual. This screening method is justifiable because, after all, the purpose
of CMPA is to screen in a person’s “top” choices “efficiently.” After the third Likert round, the same screening method is repeated, but based on the results of all three Likert rounds. The remaining majors will enter the forced-choice rounds.

The next four rounds are in pair-wise forced-choice format, thus, named as “forced-choice rounds.” Each forced-choice item has two phrases (for two majors) from which the respondents must choose one they prefer. See Figure 3.1 for an example forced-choice item.

The operation in the first three forced-choice rounds is the same. That is, two phrases are randomly chosen, without replacement, from the screen-in majors to create a forced-choice item until running out of phrases to pair up\(^2\). The top three majors (with possible ties) screened in by these three random-pairs rounds will enter the last round where all pair-wise comparisons are presented to an individual, and the results are ranked. Figure 3.2 summarizes the seven-round procedure.

---

\(^2\)In total, there are 10 sets of phrases. Three sets are used for the seven rounds of assessments. The remaining three sets are used in the case when an odd number of majors are screened into the forced-choice rounds. In this case, an additional phrase in the fourth, fifth, and six rounds are needed because an even number of phrases are needed in order to randomly pair the phrases into items. For each respondent, the phrases of the major that are closest to the screen-2 cut-off will be used.
Depending on the individual, the entire procedure takes 15 to 20 minutes to complete. The final report only ranks the top three majors (with possible ties) and does not provide any norm-referenced scores such as percentiles or Z-scores.

During the intermittent online pilots between the years 2018 and 2019, more than 13,000 people had voluntarily used CMPA, showing the high demand for a tool of this sort. However, as a newly developed assessment, it is essential to evaluate the effectiveness of CMPA. Thus, the purpose of this study is to evaluate the validity and reliability of CMPA, exploring whether CMPA could fulfill its designed purpose.

Considering the large number of majors assessed by CMPA, yet a reasonable length of a Master’s thesis, the present study will only report the results for two majors, Education and Psychology, which are most relevant for a thesis in the Department of Educational and Counselling Psychology, and Special Education of Faculty of Education.
3.1.2 Validation Approaches: Argument-based Approach

There were many validity theories. Kane’s (1992) approach of argument-based validation was adopted for the present study. Kane attested that validity is not a property of the tests themselves, but test score interpretations and uses. The process of validation, in Kane’s view, included three steps. It starts with a proposed interpretation and use Argument (IUA) according to the purpose of an assessment. Then, one articulates the arguments for how the assessment fulfills the IUA. Finally, one evaluates the arguments and makes a conclusion about the validity of the IUA (Kane, 2001; Kane, 2016). The validity arguments can be evaluated based on the prior knowledge (known theories or established practices, the assessment design), and/or empirical data (Kane, 2013).

Kane’s argument-based approach to validation is most suitable when the IUA for the test purpose can be clearly articulated (Lavery et al., 2020). Compared with other validity theories, the argument-based approach provides a very general and flexible framework for validation. It does not prescribe any recipe for how validation should be done for practitioners or applied researchers (Sireci, 2013; Jacobson & Svetina, 2019). For the purpose of validating CMPA, a tailored validation plan should be made to clarify the IUA and validity arguments. Table 3.1 states the warrant for CMPA to be evaluated in this study and explicates the assumptions to be tested with the empirical data.
Table 3.1
CMPA IUA: Warrants, Assumptions and Sources for Backing

**Warrant:** The combination of Likert and forced-choice formats as well as the three self-referenced screening steps of CMPA can achieve a high level of effectiveness in helping individuals’ find their favorite majors.

<table>
<thead>
<tr>
<th>Assumptions</th>
<th>Sources of Backing</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. The contents are authentic and unique to 51 majors for assessing preference for college majors.</td>
<td>This is achieved by identifying assessment contents that are authentic and unique to each major, taken from the natural texts describing the study of the majors using the TF-IDF method.</td>
</tr>
<tr>
<td>b. The identification of one’s favorite major by the seven rounds assessment is not influenced by individuals’ perception about the major.</td>
<td>The findings from research Question-1: Did the respondents display response set due to Likert-type format of self-assessment? If so, was it associated with the Favorite trait?</td>
</tr>
<tr>
<td>c. The Likert rounds and the first two self-referenced screening steps work as a simple yet effective way of screening out individuals’ “not preferred” majors from the 51 majors.</td>
<td>The findings from research Question-2: Can the Likert rounds (with the first two screening steps) effectively screen out individuals’ “not preferred” majors?</td>
</tr>
<tr>
<td>d. The forced-choice rounds and the third self-referenced screening step work as an effective way of identifying individuals’ favorite major(s).</td>
<td>The findings from research Question-3: Can the forced-choice rounds (with the third screening step) effectively identify individuals’ favorite major?</td>
</tr>
</tbody>
</table>

*Note.* This Table shows only the final argument in Table 2.1 in Chapter 2 because the purpose of the present study is to check the validity of this specific argument.

3.2 Research Questions

At the outset, let me define the term “Favorite” to be the latent trait that CMPA targets to assess, which is individuals’ inclination towards a major being their favorite. I will refer to this targeted trait as Favorite hereafter.

The research questions are proposed with an aim to evaluate the assumptions b. d. and e. stated in Table 3.1.
3.2.1 Question-1: Did the respondents display a response set due to the Likert-type format of self-assessment. If so, was it associated with the Favorite trait?

Response set is the “tendency to answer questions in a systematic manner that is unrelated to their content” (VandenBos, 2015). Despite that CMPA deploys the forced-choice format to reduce the impact of possible response set, I would still anticipate the following:

**H1a.** There existed a latent trait pertaining to response set when answering the three Likert-round items. That is, the response data would not be unidimensional.

**H1b.** If the individuals’ Favorite trait scores had been measured with high quality, their correlation with the response set should be close to zero. This is because if the measured scores for the targeted trait truly reflected individuals’ inclination towards a major being their favorite, the response set should not have any influence on them.

3.2.2 Question-2: Can the Likert rounds (with the first two screening steps) effectively screen out individuals’ “not preferred” majors?

Recall that the main objective of the Likert rounds is to screen out the “not-preferred” majors, as opposed to a person’s favorites to be found by the forced-choice rounds. To evaluate the effectiveness in fulfilling this purpose, the following lists the expected empirical evidence, regarding the level of difficulty and discrimination, for the three Likert rounds, and when pertinent, with comparison to those of the Forced-choice rounds.

**H2a.** First and foremost, all three Likert rounds should be highly discriminating in measuring the Favorite latent trait.
**H2b.** The third Likert round should be relatively more difficult (i.e., fewer people liking the major). This is because a major may not pass the first screening step among many respondents and automatically taken as “disliked” in the third round.

**H2c.** As the starting point to screen out the not-preferred majors, I expected the Likert-rounds to be more reliable (producing less measurement error) for measuring low to moderate level of the Favorite trait.

### 3.2.3 Question-3: Can the forced-choice rounds (with the third screening step) effectively identify individuals’ favorite major?

Recall that the main purpose of the forced-choice rounds is to identify individuals’ favorite majors effectively by way of urging them to choose between two liked majors. To evaluate the effectiveness in fulfilling this purpose, the following lists the expected empirical evidence, regarding the level of difficulty and discrimination, for the four forced-choice rounds, and when pertinent, with comparison to those of the Likert rounds.

**H3a.** First and foremost, all four forced-choice rounds should be highly discriminating for measuring the Favorite trait.

**H3b.** The forced-choice rounds should be relatively more difficult (i.e., fewer people liking the major) than the Likert rounds. This is because a major might not pass the first two screening steps for some respondents and automatically taken as “not preferred” and excluded from the forced-choice rounds. Among them, the last forced-choice round should be the most difficult, because the third screening step would further exclude some majors from the last round.
**H3c.** The forced-choice rounds should be relatively more reliable (less measurement error) in measuring people with a moderate to a high level of the Favorite trait than for people with a lower level. This is because the forced-choice rounds are designed purposefully to assess whether people have a “sufficient enough” level of liking towards a major to be their favorite.

### 3.3 Methods

#### 3.3.1 Participants

A sample of 13,569 participants volunteered to take part in this study online without incentives. A proportion of 77.5% identified themselves as female, 21.7% as male, and 0.5% as other. A large percentage of 42.3% were between 17 or 18 years old, and a noticeable proportion of 35.0% was under 16 years old. Most of the participants were pursuing a high school diploma (62.2%) or a University degree (10.9%). Considering the purpose of CMPA is to assist people to find their favorite majors, it is reasonable that we have a large number of participants between 17 or 18 and pursuing a high school or university diploma. Table 3.2 shows the demographic information of the sample.
Table 3.2
Demographics of the Sample

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Under 16</td>
<td>4748</td>
<td>35.00%</td>
</tr>
<tr>
<td>17 or 18</td>
<td>5739</td>
<td>42.30%</td>
</tr>
<tr>
<td>19 to 22</td>
<td>2122</td>
<td>9.65%</td>
</tr>
<tr>
<td>23 to 29</td>
<td>604</td>
<td>2.70%</td>
</tr>
<tr>
<td>30 and above</td>
<td>354</td>
<td>1.41%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Current Level of Education</strong></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>High school diploma</td>
<td>8432</td>
<td>62.15%</td>
</tr>
<tr>
<td>College-level technical vocational diploma</td>
<td>398</td>
<td>2.93%</td>
</tr>
<tr>
<td>University degree</td>
<td>2634</td>
<td>10.94%</td>
</tr>
<tr>
<td>Other diploma/degree</td>
<td>680</td>
<td>2.51%</td>
</tr>
<tr>
<td>I am not working towards any certificates</td>
<td>1425</td>
<td>5.25%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Gender</strong></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>10516</td>
<td>77.51%</td>
</tr>
<tr>
<td>Male</td>
<td>2968</td>
<td>21.88%</td>
</tr>
<tr>
<td>Other</td>
<td>85</td>
<td>0.53%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Current School</strong></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Public (state) college</td>
<td>2143</td>
<td>15.79%</td>
</tr>
<tr>
<td>Private college</td>
<td>369</td>
<td>2.72%</td>
</tr>
<tr>
<td>Public (state) high school</td>
<td>7920</td>
<td>58.37%</td>
</tr>
<tr>
<td>Private high school</td>
<td>1326</td>
<td>9.77%</td>
</tr>
<tr>
<td>Home school</td>
<td>290</td>
<td>2.14%</td>
</tr>
<tr>
<td>Other</td>
<td>1521</td>
<td>11.21%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Plan to do next</strong></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Find a job</td>
<td>1406</td>
<td>10.37%</td>
</tr>
<tr>
<td>Go to a technical/community college or institute</td>
<td>1585</td>
<td>11.69%</td>
</tr>
<tr>
<td>Go to a university</td>
<td>8553</td>
<td>63.06%</td>
</tr>
<tr>
<td>Go to another kind of school</td>
<td>482</td>
<td>3.51%</td>
</tr>
<tr>
<td>Go to graduate school (Masters or Doctoral)</td>
<td>1543</td>
<td>11.38%</td>
</tr>
</tbody>
</table>

3.3.2 The CMPA Assessment Data for the Current Study

The data were dichotomized into binary outcomes and used as the observed indicator variables. That is, every participant had records of whether each one of the 51 majors was chosen.
in the seven rounds of assessment (Yes =1 or No =0). For the Likert rounds, Agree and Strongly Agree were taken as Yes; Disagree and Strongly Disagree were taken as No. For the forced-choice rounds, the outcome would be classified as Yes as long as a major was chosen once in each round. Otherwise, they were classified as No. The proportions of outcome = Yes in 7 rounds were 41.9%, 57.1%, 22.2%, 18.9%, 21.6%, 24.4%, 10.5%, respectively for Education, and 70.6%, 57.8%, 57.3%, 39.4%, 42.3%, 39.6%, 30.7%, respectively for Psychology.

The data for each was further split into two random halves. All the analyses were first conducted on the first half of the random sample and the second half was used as the cross-validation sample.

3.3.3 Data Analysis: Latent Trait Model in Structural Equation Modeling (LTM-SEM)

The latent trait model (LTM), named by Birnbaum (1968), also known as item response theory (IRT), is a measurement model. It assumes that individuals’ test scores can be predicted by their unobservable traits (Hambleton & Cook, 1977). The latent trait is treated as the independent variable for the multivariate observed indicator/manifest variables, usually via a logistic or normal ogive regression. The regression coefficients (intercepts and slopes) provide information about the level of difficulty and discrimination of the observed indicators for measuring the latent trait (Thorpe et al., 2007). Moreover, the errors in estimation of the latent trait variable provide information about the reliability in measuring people with different levels of the latent trait.

The widely used unidimensional 2-parameter logistic model for binary data (Hambleton & Swaminathan, 2013) is formulated to model the probabilities of the observed indicator $U_{ij}$ by the
two parameters of difficulty \((b_i)\) and discrimination \((a_i)\) and the person latent random variable \(\theta_j\) given as

\[
P_{ij}(U_{ij} = 1)|\theta_j, b_i, a_i = \frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_i)]} \tag{1}
\]

where \(P_{ij}\) is the probability of the binary observed indicator \(U_{ij}\), \(i = 1, \ldots, I\) of person \(j = 1, \ldots, N\). The \(\exp\) denotes the exponential function with a base of \(e = 2.71828\).

Equation (1) assumes uni-dimensionality, meaning that only one latent trait underlies the multivariate response data. When more than one latent traits were assumed, the Multidimensional Latent Trait Models (MLTM, a.k.a. MIRT) should be considered. One of the most popular MLTMs (Ackerman et al., 2003; Reckase, 2009) is formulated as,

\[
P_{ij}(U_{ij} = 1| a_i, d_i, \theta_j) = \frac{\exp[d_i + a_i \theta_j']}{1 + \exp[d_i + a_i \theta_j']}. \tag{2}
\]

The \(a_i = (a_{i1}, a_{i2}, \ldots, a_{im})\) is a vector of discrimination (slope) parameters of \(\theta_j\) on indicator \(i\), where \(\theta_j = \text{a vector of m latent trait variables} = (\theta_{j1}, \theta_{j2}, \ldots, \theta_{jm})\). The parameter \(d_i\) is the intercept (conditional mean) for each observed indicator \(i\).

An easy way to understand equation (2) is that the logit (log of the odds) of the outcome \(U_{ij} = 1\) is modeled as a linear combination that includes a baseline intercept \(d_i\) for each observed outcome \(i\), plus the weighted sum of the latent traits \(\theta_{jm}\) where the weights are \((a_{i1}, a_{i2}, \ldots, a_{im})\) as shown on the right-hand side of (3) given below,

\[
\text{logit} = \log \frac{P(U_{ij} = 1)}{1 - P(U_{ij} = 1)} = d_i + a_i \theta_j'. \tag{3}
\]

In a case of \(m = 2\) for example,

\[
\text{logit} = d_i + a_{i1} \theta_{j1} + a_{i2} \theta_{j2}. \tag{4}
\]
The d-parameter is negatively related to (but not equitant) to the difficulty parameter $b_i$ in the unidimensional LTM such that $b_i = -\frac{d_i}{\sqrt{\sum_{k=1}^{M} a_{ik}^2}}$, where $M$ indicates total the number of latent trait variables. With two latent traits for example, $b_i = -\frac{d_i}{\sqrt{a_{i1}^2 + a_{i2}^2}}$. Because $b_i$ is intuitively easier to understand as a measure of difficulty (higher value means more difficult). I would report $b_i$ instead $d_i$ in the Results section when referring to difficulty.

Like factor analysis, an LTM can be specified within the structural equation modeling framework (SEM, Muthén, 1983). In-doing so allows the users to specify and evaluate the models that are hypothesized specifically to answer their research questions, as I shall do next for the current study.

Figure 3.3 shows the four alternative models specified and tested on the current data. Results of one of them would be reported to answer the three research questions. Model-uni includes only one single latent variable, indicated by the seven binary outcomes. This latent variable was to represent the targeted trait that I labeled as Favorite – the individuals’ inclination towards a major being their favorite. Model-2dim-a added latent variable on top of Favorite. This latent variable was indicated only by the outcomes of the three Likert rounds. It was introduced to capture any additional common covariation among the Likert rounds, over and beyond that already accounted for by Favorite. As mentioned before, the Likert-type self-rating scale is susceptible to a response set that is irrelevant to the trait to be measured. Such dispositions, when manifested in assessing a person’s liking, can be understood as one’s immediate or intuitive recognition or appreciation about a major, reflected on the Likert-type rating. Hence, this latent trait was labeled as
“Perception” (similar to social desirability when responding to Likert-type items for assessing personality).

The two remaining models were built on the specification of Model-2dim-a. The specification of Model-2dim-b was the same as Model-2dim-a except that the two latent traits were specified to correlate (shown by a double-headed curvy arrow). The specification of Model-2dim-c was also the same as Model-2dim-a except that the conditional dependence of R4, R5, and R6 on R7 was added, as shown by the arrows from R4, R5, and R6 to R7 in Figure 3.3.

My hypothesis was that Model-2dim-c (uncorrelated traits with a dependence of R7 on R4, R5 & R6) would be the best-fitting among the four for the following three reasons. Firstly, as discussed, the Likert-type format can induce systematic responses that are “unrelated” to the targeted trait. This was one of the reasons why the forced-choice format was deployed in CMPA in the first place so as to reduce such score variation. Second, if the scores of the Favorite trait truly reflect individuals’ inclination towards a major being their favorite, its correlation with Perception should be close to zero. That is, the measurement of Favorite should not be influenced by individuals’ subjective perception about the major. Finally, the conditional dependence of R4, R5, and R6 on R7 cohered with the CMPA’s design for the forced-choice rounds such that the majors entered to the last round were determined by the outcomes of the first three Likert rounds (See Figure 3.2 for the assessment procedure).
Figure 3.3
Specification of LTM-SEM for the CMPA Assessment Design and Data
Model-uni: Unidimensional LTM

Model-2dim-a: LTM with two dimensions

Model-2dim-b: LTM with two correlated latent traits

Model-2dim-c: LTM with two uncorrelated latent traits and dependency of R4, R5, R6 on R7
Before fitting the four alternative models, parallel analyses were conducted to explore the
dimensionality (the number of latent traits) underlying the data sets using the *psych* package in *R*
(Revelle & Revelle, 2015). The number of dimensions was determined by the number of principal
factors/components with eigenvalues higher than those of the 95\(^{th}\) percentile of 5,000 random
parallel datasets.

Two types of measures were used to evaluate the fit of the four alternative LTM-SEM
models to the empirical data. The information-based criteria of Akaike information criterion
(AIC), Bayesian information criterion (BIC), and Sample-Size adjusted BIC (SBIC) were used to
evaluate the relative model fit to the likelihood of the raw data. A model with a smaller
information criterion was more likely to be the model that generated the raw data. In addition, the
chi-square difference test was used to compare the fit of the 2-dimensional models. This was
evaluated by the improvement in fit from the 2dim-a model to the more augmented models of
2dim-b and 2dim-c. The test was based on the drop in the chi-squares ($\Delta \chi^2$), evaluated against the
chi-square distribution with degree of freedom = $\Delta df$, the difference between the two models
being compared. A statistically significant result indicates that the augmented model should be
supported.

To cross-validate the trustworthiness of the results of the final model to be chosen for
reporting, the same specifications were fitted to the second half of the sample to evaluate its fit to
another dataset.

All LTM-SEM analyses were conducted in *Mplus* 7.4 (Muthén & Muthén, 2018) using the
maximum likelihood robust (MLR) estimator. This method was suitable for the current data
because it yielded parameters with standard errors and a chi-square test statistic that was robust to non-normal data (like the binary data in the current study) and non-independence of observations (like R4, R4, R6 on R7 in Model-2dim-c).

3.4 Results

In what follows, I will organize the results by the hypotheses I specified for the three research questions, respectively.

3.4.1 Question-1: Did the respondents display response set due to the Likert-type format of self-assessment. If so, was it associated with the Favorite trait?

**Hypothesis H1a.** For both Education and Psychology, parallel analyses suggest that there were two principal dimensions underlying the seven rounds of data. The results are reported in Appendix A. Table 3.3 reports the fit measures of the four alternative LTM-SEM models. First and foremost, the unidimensional model had much higher information criteria than the two-dimensional models, showing their poor fit to the data for both Education and Psychology. These two findings confirmed hypothesis H1a that the data were not unidimensional and possibly two-dimensional.

**Hypothesis H1b.** This hypothesis was evaluated by the fit measures, comparing the three 2-dimensional models in Table 3.3. Clearly, Model 2dim-b (two correlated traits) was not supported by any of the information criteria, neither for Education nor for Psychology data. Even worse, the $p$-value for the chi-square difference test could not be computed because the chi-square was
“inversely increased” by 3.16 for Education and by 36.14 for Psychology after one additional parameter (trait correlation) was added. These unacceptable results suggest that Model 2dim-b was a wrong model. These findings confirmed hypothesis H1b that the Favorite trait and the response set were not correlated, Namely, whether individuals chose Education or Psychology to be their favorite was not influenced by their perception about these two majors.

Table 3.3
Model fit Comparison - Information Criteria and Chi-square Difference Tests

<table>
<thead>
<tr>
<th>Education</th>
<th>AIC</th>
<th>BIC</th>
<th>SBIC</th>
<th>(\chi^2) Test</th>
<th>(\Delta df)</th>
<th>(\Delta \chi^2)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uni</td>
<td>41324.48</td>
<td>41426.82</td>
<td>41379.15</td>
<td>2dim-b vs. 2dim-a</td>
<td>1</td>
<td>+3.16</td>
<td>NA(^a).</td>
</tr>
<tr>
<td>2dim-a</td>
<td>41043.38</td>
<td>41159.35</td>
<td>41105.33</td>
<td>2dim-a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2dim-b</td>
<td>41048.53</td>
<td>41171.33</td>
<td>41114.13</td>
<td>2dim-b vs. 2dim-a</td>
<td>3</td>
<td>-8.39</td>
<td>p = 0.04</td>
</tr>
<tr>
<td>2dim-c</td>
<td>41040.99</td>
<td>41177.43</td>
<td>41113.87</td>
<td>2dim-a</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Psychology</th>
<th>AIC</th>
<th>BIC</th>
<th>SBIC</th>
<th>(\chi^2) Test</th>
<th>(\Delta df)</th>
<th>(\Delta \chi^2)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uni</td>
<td>47658.49</td>
<td>47754.00</td>
<td>47709.51</td>
<td>2dim-b vs. 2dim-a</td>
<td>1</td>
<td>+36.14</td>
<td>NA(^a).</td>
</tr>
<tr>
<td>2dim-a</td>
<td>47383.98</td>
<td>47499.95</td>
<td>47445.93</td>
<td>2dim-a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2dim-b</td>
<td>47422.11</td>
<td>47544.90</td>
<td>47487.70</td>
<td>2dim-b vs. 2dim-a</td>
<td>3</td>
<td>-27.53</td>
<td>p &lt; 0.01</td>
</tr>
<tr>
<td><strong>2dim-c</strong></td>
<td><strong>47362.43</strong></td>
<td><strong>47498.88</strong></td>
<td><strong>47435.32</strong></td>
<td>2dim-a</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. AIC: Akaike information criterion; BIC: Bayesian information criterion; SBIC: Sample-Size Adjusted BIC. The \(\chi^2\) test = Chi-squared Difference Test; \(\Delta df\) = difference in degrees of free between two models; \(\Delta \chi^2\) = difference in chi-square values between two models. The smallest information criteria and statistically significant \(\chi^2\) difference tests were highlighted in boldface.

\(^a\) The p-value is not available because the chi-square increased by 3.16 for Education and by 36.14 for Psychology after one additional parameter (correlation between two traits) was added. These results suggest that 2dim-b was problematic.

At this juncture, I needed to select from the two remaining models, 2dim-a and 2dim-c, so that I can answer the second and third research questions based on their results. Table 3.3 showed that, for Psychology data, the information-based criteria and chi-square difference test unanimously supported model 2dim-c. However, the results for Education were mixed.

Eventually, I chose to report the results of 2dim-c hereafter, because it was confirmed by the Chi-
square difference tests for both majors and was most akin to the original design of CMPA (i.e.,
two uncorrelated traits with possible Likert-round dependence).

3.4.2 Question-2: Can the Likert rounds (with the first two screening steps) effectively
screen out individuals’ “not preferred” majors?

Hypothesis H2a. According to Baker (2001), discrimination parameter value 0.01 – 0.34 is
regarded as very low, 0.35 – 0.64 as low, 0.65 – 1.34 as moderate, 1.35 – 1.69 as high, and > 1.70
as very high. All three Likert rounds had very high levels of discrimination for measuring
Favorite, judging by the discrimination parameter a1. Figure 3.4 reports the estimates of a1 = 2.76,
2.20, and 6.13 for Education and = 2.51, 2.66, and 2.65 for Psychology. All of them were > 1.7,
the standard for being considered as “very high.” This hypothesis was confirmed.

Hypothesis H2b. Judging by the difficulty parameter b reported in Figure 3.4 and Table 3.4,
the third Likert round of Education (0.79) was indeed relatively more difficult than the first two
Likert rounds (0.25 and -0.22). The same was found for Psychology. The third Likert round (-
0.20) was slightly more difficult than the second (-0.21) and noticeably more difficult than the
first (-0.64).

Hypothesis H2c. For Psychology, the first two Likert rounds were relatively more reliable
(less measurement error) for measuring people with a low to moderate level of Favorite than for
measuring people with a high level of Favorite. These findings were evaluated by the information
function, the certainty about the estimation over the continuum of the latent trait Favorite
(denoted as \( \theta_1 \)), illustrated in Figure 3.5. For Psychology (at the bottom half of Figure 3.5), the
Likert rounds provided good information roughly around the region of -1.5 to 1.5 with a peak at $\theta_1 \approx -0.5$. For Education, the same pattern was observed, (see the top half of Figure 3.5) except that the third round was more reliable for a moderately high level around the region of 0.5 to 1.5 with a peak at $\theta_1 \approx 1$. This finding was understandable given that Education may have been screened out for people with low to moderate levels of Favorite and turned out to be more reliable for a moderately high level of Favorite.

### 3.4.3 Question-3: Can the forced-choice rounds (with the third screening step) effectively identify individuals’ favorite major?

**Hypothesis H3a.** All four forced-choice rounds were highly discriminating for measuring the latent trait Favorite as shown by the estimates of the $a_1$ discrimination parameters reported in Figure 3.4. The minimum $a_1$ value among the four rounds was 2.29 for Education and 2.46 Psychology, which was $> 1.7$, the standard for being considered as “very high” by Baker (2001). This hypothesis was confirmed.

**Hypothesis H3b.** As I expected, the forced-choice rounds were relatively more difficult than the Likert rounds for both Education and Psychology. This is shown by the estimates of the $b$ parameters reported in Figure 3.4. This pattern can also be observed by the location of the characteristic curves of the forced-choice rounds being closer to the right end of the $\theta_1$ continuum, showing that a higher level of $\theta_1$ was needed to choose Education or Psychology as the Favorite. In particular, the last forced-choice round had the highest level of difficulty, which
required a minimum level of $\theta_1 = 1.63$ and 0.96 to have a 50% chance of choosing Education and Psychology, respectively, as the favorite.

**Hypothesis H3c.** The information curves in Figure 3.4 show that the forced-choice rounds provided good information for estimating $\theta_1$ roughly in a region between 0 and 2.5 for Education. This confirmed my hypothesis that the forced-choice rounds yielded less measurement error when measuring people with a moderate to a high level of Favorite than for people with a lower level of Favorite. For Psychology, the first three forced-choice rounds provided good information for a wide range from moderately low ($\approx -1$) to moderately high ($\approx 1.5$), with a peak at $\approx 0.5$. The last round was more noted for its high reliability for measuring a moderate to a high level of $\theta_1$ between 0 to 2.

Finally, the results for all hypotheses regarding the three research questions were replicated by the second half of the samples. The replicated results are provided in Appendix B. The replication of the results consolidated the trustworthiness of the findings reported in this study.

**Table 3.4**

<table>
<thead>
<tr>
<th>Round</th>
<th>Discrimination</th>
<th>Difficulty</th>
<th>Location</th>
<th>Discrimination</th>
<th>Difficulty</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Education</td>
<td></td>
<td>Parameter</td>
<td>Psychology</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$a_1$</td>
<td>$b$</td>
<td>$(d)$</td>
<td>$(a_1)$</td>
<td>$b$</td>
<td>$(d)$</td>
</tr>
<tr>
<td>R1(L1)</td>
<td>2.76</td>
<td>0.25</td>
<td>-0.74</td>
<td>2.51</td>
<td>-0.64</td>
<td>1.69</td>
</tr>
<tr>
<td>R2(L2)</td>
<td>2.20</td>
<td>-0.22</td>
<td>0.49</td>
<td>2.66</td>
<td>-0.21</td>
<td>0.58</td>
</tr>
<tr>
<td>R3(L3)</td>
<td>6.13</td>
<td>0.79</td>
<td>-6.85</td>
<td>2.65</td>
<td>-0.20</td>
<td>0.74</td>
</tr>
<tr>
<td>R4(F4)</td>
<td>2.29</td>
<td>1.12</td>
<td>-2.56</td>
<td>2.46</td>
<td>0.36</td>
<td>-0.88</td>
</tr>
<tr>
<td>R5(F5)</td>
<td>2.44</td>
<td>0.98</td>
<td>-2.39</td>
<td>3.10</td>
<td>0.26</td>
<td>-0.80</td>
</tr>
<tr>
<td>R6(F6)</td>
<td>2.59</td>
<td>0.86</td>
<td>-2.22</td>
<td>2.70</td>
<td>0.34</td>
<td>-0.93</td>
</tr>
<tr>
<td>R7(F7)</td>
<td>2.90</td>
<td>1.63</td>
<td>-4.73</td>
<td>3.42</td>
<td>0.96</td>
<td>-3.28</td>
</tr>
</tbody>
</table>

*Note. $a_1$ is the discrimination parameter of $\theta_1$. The higher the value, the more discriminating the round is. The difficulty parameter $b$, same as the difficulty parameter in Unidimensional IRT, has a relationship with $d$ in Equation (4). The higher the value $b$, the more difficult the round is. The $d$ is the location parameter in Equation (3).*
Figure 3.4
Characteristic Curves - the probability of the major being chosen in each round

Education

Round 1 (Likert 1)
a_1 = 2.76, b = 0.25.

Round 2 (Likert 2)
a_1 = 2.20, b = -0.22.

Round 3 (Likert 3)
a_1 = 6.13, b = 0.79.

Round 4 (Forced-choice 1)
a_1 = 2.29, b = 1.12.

Round 5 (Forced-choice 2)
a_1 = 2.44, b = 0.98.

Round 6 (Forced-choice 3)
a_1 = 2.59, b = 0.86.

Round 7 (Forced-choice 4)
a_1 = 2.90, b = 1.63.
**Psychology**

Round 1 (Likert 1)
\[ a_1 = 2.51, b = -0.64. \]

Round 2 (Likert 2)
\[ a_1 = 2.66, b = -0.21. \]

Round 3 (Likert 3)
\[ a_1 = 2.65, b = -0.20. \]

Round 4 (Forced-choice 1)
\[ a_1 = 2.46, b = 0.36. \]

Round 5 (Forced-choice 2)
\[ a_1 = 3.10, b = 0.26. \]

Round 6 (Forced-choice 3)
\[ a_1 = 2.70, b = 0.34. \]

Round 7 (Forced-choice 4)
\[ a_1 = 3.42, b = 0.96. \]

**Note.** Parameter \( a_1 \) is the discrimination parameter, and \( b \) is the difficulty parameter for Favorite (see Equation 5 for more details).

There are two curves for the first three Likert rounds. The second curve (in dashed line) shows how Perception \( (\theta_2) \) influenced the probability of a major being chosen in the first three Likert rounds (see Equation 5 and Figure 3.3). The solid dot on the curve shows the probability of choosing Education/Psychology in a typical round when Favorite \( (\theta_1) = 0 \).
Figure 3.5
Information Function for the Seven Rounds, Given Perception ($\theta_2 = 0$

**Education**

Round-1 (Likert-1)

Round-2 (Likert-2)

Round-3 (Likert-3)

Round-4 (Forced-choice-1)

Round-5 (Forced-choice-2)

Round-6 (Forced-choice-3)

Round-7 (Forced-choice-4)
Psychology

Note. For the purpose of comparing across different rounds, the scale for Y (information) was set to 0 to 4 for all 7 rounds. The information for the third round of Education is exceptionally high, \( = 18.90 \) at \( \theta_i = 1.1 \), which was higher than the maximum scale point set for the Y-axis.
3.5 Summary

The purpose of this chapter was to provide empirical evidence for the intended interpretation and use of CMPA rankings. To do so, I adopted Kane’s validation approach by way of explicating the warrant and their assumptions about CMPA, which were then tested by a large representative sample data for Education and Psychology. Specifically, I posed three research questions, each with my hypotheses about the findings from LTM-SEM, a statistical framework precisely chosen to suit the current data.

Almost all my hypothesis were supported by the findings from LTM-SEM and replicated by another sample data, providing evidence for that warrant – The combination of Likert and forced-choice formats, as well as the three self-referenced screening steps of CMPA, can achieve a high level of effectiveness on helping individuals find their favorite majors.
Chapter 4: Conclusion & Discussions

This chapter recapitulates the subject matter of this thesis. I will first summarize the findings from Chapter 3, and then address the limitations and contributions of the current study in light of the literature. In closing, I will discuss the directions, in which I can delve for my future research.

4.1 Conclusion

This thesis, in essence, is a part of a validation study for CMPA, taking Kane’s (1992) argument-based approach. CMPA is a relatively new tool for assisting prospective students in finding their preferred college majors. In order to “directly” finding individuals’ favorites, CMPA took a pragmatic strategy (vs. theory-based) to its design and development. The assessment contents were taken from the online natural authentic texts, featuring each college major. To achieve a high level of efficiency and effectiveness, the assessment formats included Likert-type rating and forced-choice format so that the strengths of each can be utilized, and the drawbacks can be mitigated.

The validation study included three steps: 1) proposing the interpretation and use argument (IUA) according to the purpose of CMPA; 2) articulating the warrants for how CMPA fulfills the IUA; 3) verifying the warrants with evidence with prior knowledge and/or empirical evidence. In this study, the warrant to be verified is: CMPA is effective for ranking an individual’s favorite majors. To justify this warrant, I made four assumptions about CMPA. The first assumption was

3 The other warrants are listed in Chapter 2, but they are not the focus of this validation study.
not evaluated in this study. It was justified by the authenticity and uniqueness of the contents for assessing the 51 majors. The assessment contents were taken from authentic text information and extracted by the method of natural language process using TF-IDF technique. According to a previous study by Fan and Qin (2018), this method can retrieve unique contents for each category (college majors).

The remaining three assumptions were evaluated according to my hypotheses about the expected findings for the three research questions: 1) Did the respondents display response set due to the Likert-type format of self-assessment? If so, was it associated with the Favorite trait? (2) Can the Likert rounds (with the first two screening steps) effectively screen out individuals’ “not preferred” majors? (3) Can the forced-choice rounds (with the third screening step) effectively identify individuals’ favorite majors? The analyses were conducted for four large samples, two for Education and two for Psychology, based on the binary outcomes of the seven rounds of assessment.

For the first research question, parallel analyses suggested that data were two-dimensional. Model fit indices of LTM-SEM confirmed this finding, showing that the two-dimensional model, model-2dim-c, fitted the data the best. Model-2dim-c also confirmed my hypothesis that the assessment of the Favorite was not associated with the secondary latent trait Perception, i.e., the response set due to the Likert rating format.

The second and third research questions were evaluated by the discriminating power, difficult level, and information (certainty) in assessing the target latent trait Favorite, comparing the seven rounds of assessment. For discriminating power, all seven rounds were “very highly”
discriminating according to Baker’s (2001) criteria. Despite all being very discriminating, the forced-choice rounds were even more discriminating than the Likert-rounds in finding a person’s favorite. These findings confirmed my hypotheses.

As for difficulty level, the forced-choice rounds were more difficult than the Likert rounds. The last forced-choice round was the most difficult. These findings were consistent with my hypotheses, showing the expected progression in difficulty over the course of the assessment.

Information curves inspected the certainty (indication of reliability for LTM) as a function of the trait level of Favorite. In general, Likert rounds were more reliable for measuring people with a low-to-moderate level of Favorite. In contrast, forced-choice rounds were more reliable for measuring people with a moderate-to-high level of Favorite. This finding confirmed my hypothesis that the Likert rounds would be more reliable for screening out not-preferred majors, and the forced-choice rounds would be more reliable for screening in favorite majors.

In a nutshell, the findings based on Education and Psychology data, by and large, supported the hypotheses I made for the three research questions. Table 4 summarizes the main findings (and the associated assumptions, research questions, and hypotheses). Therefore, the following statement was warranted: “The combination of Likert and forced-choice formats as well as the three self-referenced screening steps of CMPA can achieve a high level of effectiveness on helping individuals find their favorite majors.”
Table 4.1
**CMPA IUA: Warrants, Assumptions, Research Questions and Main Findings**

<table>
<thead>
<tr>
<th>Assumptions</th>
<th>Research Questions</th>
<th>Hypotheses</th>
<th>Main Findings (Evidences)</th>
</tr>
</thead>
<tbody>
<tr>
<td>b. The identification of one’s favorite major by the seven rounds assessment is not influenced by individuals’ perception about the major.</td>
<td>Question-1: Did the respondents display a response set due to the Likert-type format of self-assessment. If so, was it associated with the Favorite trait?</td>
<td>H1a. The response data would not be unidimensional and only had one single trait underlay them.</td>
<td>Parallel analysis; Model fit indices comparison.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>H1b. If the individuals’ Favorite trait scores had been measured with high quality, their correlation with the perception should be close to zero.</td>
<td>Model fit indices comparison suggests model-2dim-c is the most fitted.</td>
</tr>
<tr>
<td>c. The Likert rounds and the first two self-referenced screening steps work as a simple yet effective way of screening out individuals’ “not preferred” majors from the 51 majors.</td>
<td>Question-2: Can the Likert rounds (with the first two screening steps) effectively screen out individuals’ “not preferred” majors?</td>
<td>H2a. All three Likert rounds should be highly discriminating in measuring the Favorite latent trait.</td>
<td>Characteristics curves. (discrimination parameter (a))</td>
</tr>
<tr>
<td></td>
<td></td>
<td>H2b. The third Likert round should be relatively more difficult (i.e., fewer people liking the major).</td>
<td>Characteristics curves. (difficulty parameter (b))</td>
</tr>
<tr>
<td></td>
<td></td>
<td>H2c. The Likert-rounds to be more reliable for measuring low to moderate level of the Favorite trait.</td>
<td>Information curves.</td>
</tr>
<tr>
<td>d. The forced-choice rounds and the third self-referenced screening step work as an effective way of identifying individuals’ favorite major(s).</td>
<td>Question-3: Can the forced-choice rounds (with the third screening step) effectively identify individuals’ favorite major?</td>
<td>H3a. All four forced-choice rounds should be highly discriminating for measuring the Favorite trait.</td>
<td>Characteristics curves. (discrimination parameter (a))</td>
</tr>
<tr>
<td></td>
<td></td>
<td>H3b. The forced-choice rounds should be relatively more difficult than the Likert rounds.</td>
<td>Characteristics curves. (difficulty parameter (b))</td>
</tr>
<tr>
<td></td>
<td></td>
<td>H3c. The forced-choice rounds should be relatively more reliable in measuring people with a moderate to a high level of the Favorite trait than for people with a low.</td>
<td>Information curves.</td>
</tr>
</tbody>
</table>

*Note. This Table shows only the assumptions b. c. and d. in Table 3.1 in Chapter 3. The assumption a. is already justified by prior knowledge of test design (TF-IDF method).*
4.2 Contributions

This thesis made several contributions to both practice and theory. First, it provides an example of using an argument-based approach to validation. One of the criticisms of Kane’s validation approach is that it is not easy to implement in practice (Brennan, 2013). There is no clear guidance on how. There is only a dearth of studies over the 30 years since Kane’s first publication on this subject in 1992 (e.g., Addey et al., 2020; Kane, 2004). Moreover, the few existing examples are restricted to performance assessment, which requires examinees to perform a task rather than select an answer from a ready-made list as most psychological assessments do. It was hard to find an example of a validation study on psychological assessment using Kane’s approach. This study, to my best knowledge, is the first attempt to use Kane’s approach to validate a psychological tool for assessing individuals’ inclination such as interest, preference, and attitude.

Second, this study showed that the combination of Likert and forced-choice formats, with timely and appropriate screenings, can effectively assess preference over a large number of options. This is evidenced by the coherence between the psychometric findings and the CMPA design, for both Education and Psychology data. Such coherence was cross-validated by a different set of samples for both majors, hence consolidated the credibility of the found coherence. This finding resonates with Reid’s (2014, p. 598) conclusion that “a two-step strategy using Likert-type and ipsative/forced-choice formats in sequence, appears to be useful in a situation where a clear preference is required from among a large number of choices.”
Third, this study unveiled the possibility of response set when responding to Likert-type items for assessing preference. According to previous works of literature (Böckenholt, 2017; Rennie, 1982) and the empirical results from this study, I believed that the response set is an artifact of Likert rating format, reflecting individuals’ perception about, unrelated to their true preference for, a major. Indeed, a secondary latent trait was found to be indicated by the Likert rounds alone (Perception) and was uncorrelated to the targeted trait (Favorite) that was indicated by all rounds. This finding had implications on practice. It is important to consider the possible effect of perception when assessing preference using a Likert-type rating scale. Including forced-choice format in the assessment design can mitigate such effect and allows an opportunity for investigating the effect of perception.

4.3 Limitation

The limitation of this study was the online voluntary sample. There was a noticeable unbalance in the gender distribution (77.51% female, 21.88% male, and 0.53% other). To be representative of the target users of CMPA, it was expected to see an equal percentage for male and female respondents. The unbalanced gender distribution is usually observed in online voluntary samples (e.g., Carnes et al., 2012). In this study, one reason for such an unbalanced distribution might be that females were more in need of assistance in finding a college major than males. However, this is just my speculation, and the issue of gender imbalance needs to be further explore.
A drawback of the present study was that I only modeled the final outcomes of forced-choice rounds and was unable to model the data at the item level. Because of the self-referenced adaptive design of CMPA, the numbers and what forced-choice items were administered were different for different individuals. Even more complicated, given that 51 majors were assessed, the two majors being paired to form the forced-choice items differed drastically for different individuals. Modeling such messy data was an extremely arduous task if not possible. Even so, it was a pity that the granular information at the item level was left unattended.

4.4. Future Studies

For the sake of the appropriateness and limited space of this thesis, this study only worked on the two majors of Education and Psychology. Nonetheless, other majors are also worth studying, in particular, those that are of much attention due to their realistic plight. For instance, front-line health care workers (e.g., nurses) were reported to have a long-run shortage (Kabel et al., 2014) and computer science has a very skewed gender representation (Cohoon, 2001). Future studies could focus on these spotlighted majors.

I am also interested in identifying majors that individuals tend to prefer in clusters. For instance, it is not unusual to observe that individuals who choose biology will also tend to choose chemistry as their candidate major. I would like to find out whether the clusters of majors, identified based on a practiced-based assessment like CMPA, will be consistent to the typological theories underpinning the popular career assessment tools such as Holland’s RIASEC Model and Myers-Briggs Type Indicator® (Holland & Gottfredson, 1992; Kennedy & Kennedy, 2004).
Another direction for research could be exploring population heterogeneity. According to previous studies, males and females had quite different motivations for taking a major (Daymont & Andrisani, 1984; Leppel et al., 2001). It would be informative to compare gender differences in the drop rates of computer science in the course of CMPA assessment to those of the majors that are more equal in gender distribution such as psychology. Besides, means difference in the latent trait scores of Perception will provide insights into stereotypes that different genders hold about different majors. All these works can help detect how deep the gender gap is. When digging into the item contents, it would help to reveal the reasons behind the gaps.
Bibliography


Chomeya, R. (2010). Quality of psychology test between Likert scale 5 and 6 points. *Journal of Social Sciences, 6*(3), 399-403.


https://www.ajpe.org/content/ajpe/74/4/75.full.pdf.


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2865412/


Rennie, L. J. (1982). Detecting a response set to Likert-style attitude items with the rating model. *Education Research and Perspectives, 9*(1), 114-118.


https://scholarworks.wm.edu/cgi/viewcontent.cgi?article=6754&context=etd


Appendices

Appendix A. The Parallel Analysis Result for Education and Psychology Dataset
## Appendix B. Replicated Results for the Second Half of the Sample

### Model fit Comparison - Information criteria and Chi-square difference tests.

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
<th>BIC</th>
<th>SBIC</th>
<th>(\chi^2\Delta) Test</th>
<th>(\Delta df)</th>
<th>(\Delta \chi^2)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uni</td>
<td>42136.10</td>
<td>42238.44</td>
<td>42190.77</td>
<td>2dim-b vs. 2dim-a</td>
<td>1</td>
<td>+2.59</td>
<td>NA</td>
</tr>
<tr>
<td>2dim-a</td>
<td>41976.73</td>
<td>41976.73</td>
<td>41875.75</td>
<td>2dim-b vs. 2dim-a</td>
<td>1</td>
<td>+2.59</td>
<td>NA</td>
</tr>
<tr>
<td>2dim-b</td>
<td>41813.21</td>
<td>41936.01</td>
<td>41878.81</td>
<td>2dim-c vs. 2dim-a</td>
<td>3</td>
<td>-4.16</td>
<td>p = 0.24</td>
</tr>
<tr>
<td>2dim-c</td>
<td>41812.64</td>
<td>41952.08</td>
<td>41888.53</td>
<td>2dim-c vs. 2dim-a</td>
<td>3</td>
<td>-26.01</td>
<td>p &lt; 0.01</td>
</tr>
</tbody>
</table>

Note. AIC: Akaike information criterion; BIC: Bayesian information criterion; SBIC: Sample-Size Adjusted BIC. The \(\chi^2\Delta\) test = Chi-squared Difference Test; \(\Delta df\) = difference in degrees of free between two models; \(\Delta \chi^2\) = difference in chi-square values between two models.

Note. For the Education validation dataset, although there is a non-significant p-value in chi-squared difference tests, considering the relatively low AIC and a decrease in \(\Delta \chi^2\), I still choose Model-2dim-c as the final model.

### Discrimination, Difficulty and Steepest Slope Direction for the Seven Rounds

<table>
<thead>
<tr>
<th>Education</th>
<th>Discrimination ((a_1))</th>
<th>Difficulty ((b))</th>
<th>Location Parameter ((d))</th>
<th>Psychology</th>
<th>Discrimination ((a_1))</th>
<th>Difficulty ((b))</th>
<th>Location Parameter ((d))</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1(L1)</td>
<td>2.67</td>
<td>0.08</td>
<td>-0.64</td>
<td>R1(L1)</td>
<td>2.38</td>
<td>-0.27</td>
<td>1.69</td>
</tr>
<tr>
<td>R2(L2)</td>
<td>2.15</td>
<td>-0.13</td>
<td>0.63</td>
<td>R2(L2)</td>
<td>2.63</td>
<td>-0.07</td>
<td>0.55</td>
</tr>
<tr>
<td>R3(L3)</td>
<td>6.40</td>
<td>0.08</td>
<td>-7.06</td>
<td>R3(L3)</td>
<td>2.56</td>
<td>-0.05</td>
<td>0.62</td>
</tr>
<tr>
<td>R4(F4)</td>
<td>2.25</td>
<td>0.46</td>
<td>-2.32</td>
<td>R4(F4)</td>
<td>2.30</td>
<td>0.16</td>
<td>-0.83</td>
</tr>
<tr>
<td>R5(F5)</td>
<td>2.30</td>
<td>0.42</td>
<td>-2.23</td>
<td>R5(F5)</td>
<td>2.84</td>
<td>0.09</td>
<td>-0.74</td>
</tr>
<tr>
<td>R6(F6)</td>
<td>2.71</td>
<td>0.30</td>
<td>-2.22</td>
<td>R6(F6)</td>
<td>2.76</td>
<td>0.13</td>
<td>-0.96</td>
</tr>
<tr>
<td>R7(F7)</td>
<td>3.37</td>
<td>0.44</td>
<td>-4.98</td>
<td>R7(F7)</td>
<td>3.37</td>
<td>0.28</td>
<td>-3.23</td>
</tr>
</tbody>
</table>

Note. \(a_1\) is the discrimination parameter of \(\theta_i\). The higher the value, the more discriminating the item is. The difficulty parameter \(b\), same as the difficulty parameter in Unidimensional IRT, has a relationship with \(d\) in Equation (4). The higher the value \(b\), the more difficult the item is. \(d\) is the location parameter in Equation(3).
Appendix C. An Example of Items about Education

Round 1 (Likert-round 1):
I want to learn about …
  
  teaching.
  ○ Strongly Disagree
  ○ Disagree
  ○ Agree
  ○ Strongly Agree

----------

Round 2 (Likert-round 2):
I want to learn about …
  
  helping children achieve.
  ○ Strongly Disagree
  ○ Disagree
  ○ Agree
  ○ Strongly Agree

----------

Round 3 (Likert-round 3):
I want to learn about …
  
  making teaching materials.
  ○ Strongly Disagree
  ○ Disagree
  ○ Agree
  ○ Strongly Agree
Round 4 (Forced-choice-round 1):

I would rather spend time learning about …

- designing teaching activities.
- how a person's memory works.

(Round 5 and 6 repeated the previous procedure)

Round 7 (Forced-choice-round 4):

I would rather spend time learning about …

- planning lessons.
- how a person's memory works.

I would rather spend time learning about …

- guiding the behavior of youth.
- the organizational structure of society.

I would rather spend time learning about …

- how the mind controls behavior.
- social problems.

Note: This example only shows items which is related to Education in a typical CMPA.
a. This is an example for Round 4. In this case, the respondent compares education with psychology.
b. Round 5 and 6 repeated the procedure in round 4, using a different item of education to compare with other majors.
c. This is an example for Round 7. In this case, the respondent compares education with psychology and sociology, using the all-against-all method and thus constructing three items for this round. The respondent will have a ranking on the three majors in the end. For example, the result could be: education is this respondent’s top favorite, followed by psychology, and sociology is the third preferred major.