ON THE IMPACT OF NEGATIVELY KEYED ITEMS ON THE ASSESSMENT OF THE UNIDIMENSIONALITY OF PSYCHOLOGICAL TESTS AND MEASURES

by

Yue Chen

B.Sc., Beijing Normal University, 2008

M.Sc., Chinese Academy of Sciences, 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Measurement, Evaluation, and Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

October 2017

Abstract

Evidence of test dimensionality supports test scoring, and it is essential to construct validity. Yet many issues remain unclear in assessing dimensionality, especially when the response data are collected through self-report Likert-type tests that include negatively keyed items. The emergence of additional factors that can be attributed to the mixed-keyed format is an issue that draws much attention in investigating the dimensionality of tests that are designed to be unidimensional. Common methods for assessing dimensionality can be categorized into two types: exploratory and confirmatory. Exploratory studies use many rules and criteria, such as the eigenvalues-greater-than-one (K-G) rule and parallel analysis (PA), along with exploratory factor analysis (EFA) to help researchers determine the number of factors (i.e., dimensions). Confirmatory factor analysis (CFA), on the other hand, is often employed to examine the fit of a hypothesized measurement model. A large number of fit indices, including the Chi-square test, the comparative fit index (CFI), and the root mean square error of approximation (RMSEA), have been proposed to evaluate a model's overall fit.

This dissertation investigated, via computer simulation, how these various procedures performed in assessing the dimensionality of item response data collected using tests with negatively keyed items. Factors in the simulation experiment included psychometric models (i.e., the simulation methods) of negatively keyed items, the number of negatively keyed items, the magnitude of item communality, the distribution of observed item response, the scoring methods of negatively keyed items, and the methods and rules used for the statistical judgment of dimensionality.

This dissertation adopts the threshold model of item responses, which assumes a monotonic relationship among the latent variable, item thresholds, and observed item responses. The results indicate that the dimensionality of tests with mixed-keyed items is always correctly identified when the observed item response distribution is symmetric. When it is asymmetric, however, the methods and decision rules used in dimensionality assessment affect the statistical judgment of test dimensionality. The results highlight the benefit of using categorical data analytic methods in dealing with item responses obtained through Likert-type rating scales. Guidelines are provided to inform researchers when assessing the dimensionality of mixed-keyed tests.

Lay Summary

Tests are widely used in the social, behavioural, and health sciences. This dissertation focuses on tests that have both positively and negatively keyed items. For example, in a test/measure of well-being, these items could include negatively keyed statements like “I am sad” and “I am not happy,” along with positively keyed statements like “I feel happy.” This dissertation seeks to better understand how the inclusion of negatively keyed items affects the statistical judgment of the test scoring and interpretation. Four closely related computer simulation studies were conducted. The results highlight the benefit of using categorical data analysis in studying item responses obtained through Likert-type rating scales. Guidelines are presented to assist researchers and practitioners to make informed decisions when analyzing their data.

Preface

This thesis is the original, unpublished, and independent work of the author, Yue (Michelle) Chen, with the guidance of her research supervisor and input from the research committee.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I could not have completed this dissertation without the great support I have received from many people over the years.

I would like to express my deepest appreciation to my supervisor, Dr. Bruno D. Zumbo, for his excellent guidance, knowledge, caring, and patience. I would like to thank Dr. Amery D. Wu for her unfailing support for my professional growth. I am also grateful to Dr. Anita M. Hubley for her insightful comments, enthusiasm, and encouragement. Without their support, it would not have been possible to complete this research.

I would like to thank my labmates, classmates, and professors from the department of ECPS and other departments at UBC. It would have been a lonely campus without them. I would also like to thank my friends, who stood by me through good times and bad.

Finally, I would like to thank my parents. They may not always understand the details of what I am doing, but have always supported and encouraged me unconditionally.

CHAPTER ONE:
INTRODUCTION

This chapter provides a general introduction to the problem investigated in this dissertation. In doing so, it explores the context of the research questions and highlights the purpose of the study. The structure of the dissertation is described at the end of this chapter.

Context of the Study

Tests, measures, questionnaires, and other measurement instruments are commonly used data collection methods in the social, behavioural, and health sciences. Scholars have proposed numerous types of such measures to assess individuals' perceptions, personalities, attitudes, and other affective characteristics, and use varied terminology to describe these measurement instruments. This dissertation adopts the following nomenclature to capture the wide diversity of usage in the vast research literature.

- The terms "test," "measure," "questionnaire," and "scale" are used interchangeably to denote a multi-item instrument that results in one or more composite scores. Please note that the word "scale" will be employed sparingly, and its meaning will be evident in the context of use, because the term has several meanings in the literature.
- Multi-item instruments are made up of items that are comprised of a stem (the statement or question) and a response scale.
- The phrase "response scale" or "rating scale" will be used to modify the term "scale." In addition, "scale response" or "item response" denotes the choice that the test taker or respondent makes when confronted with the item stem.
- The phrases "item response distribution" and "response distribution" are utilized interchangeably to describe the statistical frequency or density of responses to an item calculated for a group of respondents.
- A composite score is a sum or weighted sum of the item scores. The composite scores computed from each respondent's (or test taker's) item responses are called "test scores," "scale scores," "total scores," "test-level scores," or "factor scores," depending on the context.

The most common response scale format is a rating scale, which is also widely referred to as a Likert-type response scale or summative/summated rating scale. Tests that consist of items

with a Likert-type response format or summative rating scale, wherein one computes a composite test score, are referred to as Likert-type tests or summative tests (see Figure 1). One other central feature of these tests is that they are based on self-reporting, although an alternative does exist in which raters judge others' behaviours or characteristics.

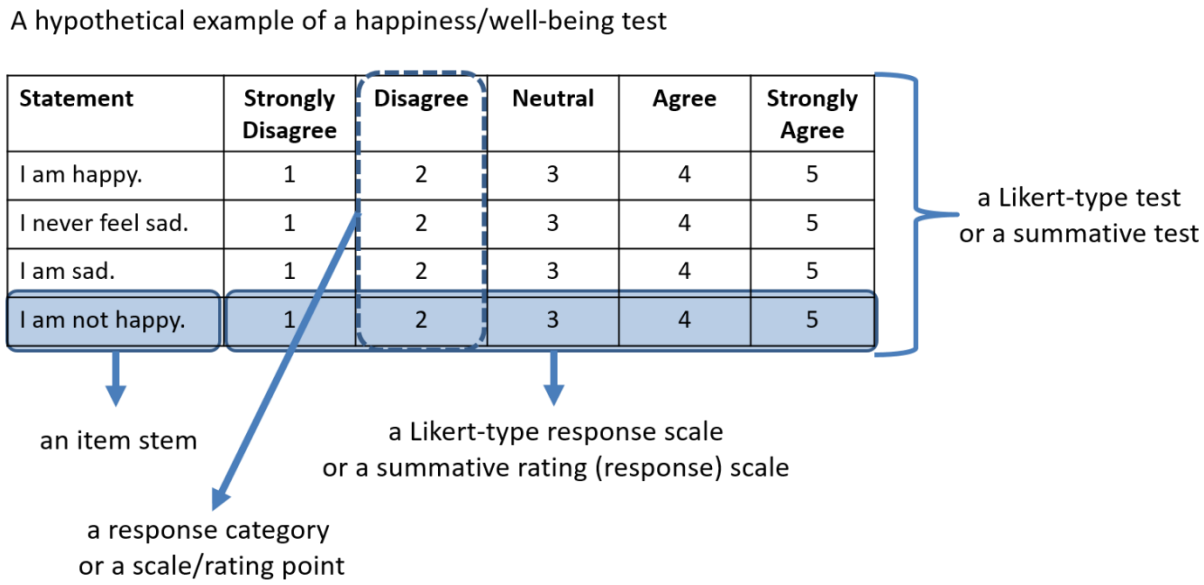


Figure 1.
Graphic depiction of the terms describing a Likert-type test

This dissertation focuses on unidimensional psychological tests. These tests are often short compared to the long test batteries in multidimensional psychological tests, and consist of Likert-type response scales. Typical examples of self-report measures in psychological research include the Rosenberg Self-Esteem (RSE) scale (Rosenberg, 1965), the Penn State Worry Questionnaire (PSWQ; Meyer, Miller, Metzger, & Borkovec, 1990), and Ryff's Psychological Well-Being Scales (PWB; Ryff & Keyes, 1995). Each of these is a multi-item measure that contains stand-alone (disjointed) items that are not dependent on the responses to other items. Each item is followed by a rating response scale that can range from two to nine points, with a common number being five. The descriptors attached to each scale often indicate different agreement levels (e.g., from "strongly disagree" to "strongly agree") with the item stem. Typically, tests that use a Likert-type response scale will have the same number of response options for all of the items because this simplifies computing the total test score.

Composite test scores from item responses are frequently used to quantify the construct of interest. Evidence regarding the dimensionality of a test is therefore essential to supporting the use of its scores. More importantly, data collected from these measurement instruments, and especially the test-level scores, are usually used to make decisions and inferences. Thus, evaluating the validity of the interpretation of the test results is critical. According to the *Standards for Educational and Psychological Testing* (*Standards;* American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), validity refers to "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). It should also be noted that the scoring strategies applied to obtain test-level scores implicitly make assumptions regarding the factor structure of a test. For example, whenever one uses a total or an average score of item responses from all the items in a test, one assumes that the test is unidimensional.

Researchers view a test score as one of the indicators of the measured construct (Hubley & Zumbo, 1996). To adequately evaluate the validity of interpretations or inferences made from these scores, one needs some information to confirm their intended meaning (Guion, 1977; Messick, 1975). One way, among many, to do this is to focus on the theoretical dimensions of the construct that the test is designed to measure. Construct validity, sometimes also referred to as factorial validity (Thompson & Daniel, 1996), seeks agreement between a particular measurement and a theoretically conceptualized construct. Investigating the dimensionality and factor structure of a test is an essential component of construct validity and test scoring. Messick (1995) identifies six aspects of construct validity: content, substantive, structural, generalizable, external, and consequential. The factor structure of a test falls into the structural category. The *Standards* (AERA et al., 2014) also recognizes evidence based on factor structure as one of the five sources of validity evidence.

It should be mentioned that both "dimensionality" and "factor structure" are used to refer to the structure of a phenomenon. The assessment of test dimensionality and factor structure are interrelated methods, and they are not always clearly distinguished in the research literature, especially when it comes to unidimensional tests. In short, the determination of the number of dimensions is a precursor to the interpretation of the factor structure, but the former is done on its own when investigating if a test is unidimensional (usually for the purposes of supporting test

scoring). In this dissertation, the terms "dimensionality" and "factor structure" are used with a subtle distinction. When the focus is on determining the number of factors (i.e., dimensions) underlying a test, the term "dimensionality" is utilized. When the factor loadings and factor interpretation (i.e., interpretation of dimensionality) are of interest, the phrase "factor structure" is used.

Construct-irrelevant variance or construct under-representation can lead to major concerns about the validity of inferences, such as unanticipated or negative consequences of score interpretation (Messick, 1998). If a measurement instrument includes factors that are not part of the construct, or if it excludes factors that are essential to the construct, the inferences made from the test scores may result in undesirable consequences. Response style, or response bias, is a frequently concerned source of construct-irrelevant variance. To offset an unwanted response style, a common practice in constructing measurement instruments is to include items keyed in different directions. Indeed, negatively keyed items are employed primarily to attenuate response bias (Idaszak & Drasgow, 1987). In particular, it is believed that if an equal number of positively and negatively keyed items are included in a test, the effects of response style, such as acquiescence response and extreme response, will be cancelled out (Nunnally, 1978).

In the literature, items that are keyed negatively are often described by ambiguous terms, such as "negative items," or by the interchangeable use of the phrases "negatively worded" and "negatively keyed." This dissertation uses "positively keyed" and "negatively keyed" to refer to items whose responses are keyed in different directions to derive a meaningful test score. For example, the ten-item RSE scale (Rosenberg, 1965) uses a four-point response scale (1-2-3-4), with a larger number meaning a higher level of agreement with the stem. One item stem states that "I take a positive attitude toward myself." A higher level of agreement on this item (i.e., a higher response score) corresponds to greater self-esteem. Another item in the RSE scale has the stem: "All in all, I am inclined to feel that I am a failure." A higher level of agreement on this item (i.e., a higher response score) shows the opposite. The responses to such items are usually re-scored in a reversed way, with a smaller numerical value assigned to those indicating a higher level of agreement—that is, the item responses are recoded such that 1 = 4, 2 = 3, 3 = 2, and 4 = 1. A detailed discussion of the terms used to describe item keying and wording, and the reverse scoring process of item responses to negatively keyed items, will be presented in the next chapter.

Although negatively keyed items are ubiquitous in educational and psychological tests, many researchers question their use (e.g., Barnette, 2000; Lai, 1994; Marsh, 1986; Motl, Conroy, & Horan, 2000; Pilotte & Gable, 1990; Schmitt & Stultz, 1985; Schriesheim & Hill, 1981). These items have been shown to reduce the validity of responses by introducing construct-irrelevant variance and systematic error (Jackson, Wall, Martin, & Davids, 1993; Schriesheim & Hill, 1981). Researchers have demonstrated that including items keyed in different directions in one test may result in an additional factor consisting of all negatively keyed items (Harvey, Billings, & Nilan, 1985; Schmitt & Stultz, 1985). The received view in the research literature is that a simple one-factor model may not represent a mixed-keyed test (i.e., a test consisting of both positively and negatively keyed items) as well as more complex models that allow for the item keying effect(s) (e.g., Carmines & Zeller, 1979; DiStefano & Motl, 2006; Marsh, 1996; Tomás & Oliver, 1999). The presumed item keying and/or wording effect has been found in various measures and populations of respondents, but its manifestation seems sample and context dependent (e.g., Barnette, 2000; Lai, 1994; Marsh, 1986, 1996; Motl et al., 2000; Pilotte & Gable, 1990; Schriesheim & Hill, 1981; Tomás & Oliver, 1999). The issues related to tests with mixed-keyed items include lower reliability and unexpected factor structure. Chapter Two will review these problems, with a focus on the emergence of additional factors in tests that were originally designed to be unidimensional.

The threat that negatively keyed items may potentially pose to the assessment of the factor structure of a test has drawn extensive attention from researchers. This is mainly because the keying-related effect observed in mixed-keyed tests may not only confuse the understanding of the factor structure, but also influences the interpretation of the subsequent statistical analysis of the test scores. In this case, it is important to comprehend the reason for the emergence of the additional unexpected factor(s).

One way to explain the appearance of unexpected factors formed by item keying direction is that items function differently when they are keyed positively or negatively. The differential functioning of so-called “negative” items is usually ascribed to the wording rather than to the keying effect (e.g., Ahlawat, 1985; Marsh, 1986, 1996). It has been suggested that the cognitive and linguistic processing demands inherent in negatively worded items are different from those needed to process positively worded ones (Ahlawat, 1985; Marsh, 1986, 1996).

Marsh (1986, 1996) has shown that respondents' verbal ability is related to their response patterns to negatively worded items.

Admittedly, in many measures, negatively keyed items are more likely to be grammatically negatively worded, while positively keyed items are usually expressed positively. However, keying direction and wording direction are not always consistent. For example, the ten-item RSE (Rosenberg, 1965) contains five positively keyed and five negatively keyed items. Among the negatively keyed items, one item stem is "All in all, I am inclined to feel that I am a failure," which does not contain any grammatical negation. In other words, that item is negatively keyed but positively worded. Therefore, item wording is probably not the sole explanation for the systematic variance among the negatively keyed items. Moreover, other confounding factors exist, such as the content area under study (i.e., targeted construct), the characteristics of scoring (e.g., reverse scoring or not), and scaling methods (e.g., observed total score or factor score).

Since the mechanisms that drive the differential functioning of positively and negatively keyed items on self-report measurement instruments are not completely understood, studies that seek additional factors to explain the occurrence of this systematic variance associated with item keying are eagerly anticipated. Empirical research based on various populations and measures has repeatedly reported issues related to the dimensionality and factor structure of unidimensional tests consisting of both positively and negatively keyed items (e.g., Marsh, 1996; Motl et al., 2000; Pilotte & Gable, 1990; Tomás & Oliver, 1999). However, this research is often unclear about whether such issues can be attributed to the wording effect alone, or whether they are better explained by the keying effect or by the interaction of both effects.

It is extremely difficult (some would say impossible) to disentangle the keying effect from other item properties and/or respondent characteristics, thus complicating (or impeding) its study in empirical work with respondents. Computer simulation studies are better suited to separating effects from different sources, including those derived from item keying and wording. To the best of my knowledge, however, no simulation studies have been conducted to systematically examine the keying effect on test dimensionality and factor structure. Three simulation studies have explored some aspects of the effect of item characteristics, including item keying and item wording, as well as respondent characteristics on the factor structure of a test (Schmitt & Stults, 1985; Spector, Van Katwyk, Brannick, & Chen, 1997; Woods, 2006).

Each of these simulation studies had a slightly divergent focus and employed different designs and procedures. These earlier studies point to the promise of using computer simulation in this context, but more work is still needed to understand the keying effect.

Purpose

This dissertation will explore the performance of different statistical methods of assessing test dimensionality and factor structure in the presence of negatively keyed items (i.e., mixed-keyed tests). An investigation of the factor structure of a test begins with either the *a priori* specification or the empirical determination of the number of factors—the former denoted as confirmatory and the latter as exploratory analyses. These statistical methods treat the item responses as either continuous or ordinal data. These various methods, and how the item response data are handled, may provide an alternative explanation for the unexpected keying effect arising in unidimensional mixed-keyed tests. Despite the prevalence of negatively keyed items in measurement instruments, and the importance of assessing dimensionality and factor structure, there remains a lack of clarity as to how researchers should proceed. For instance, it is often advised that responses to negatively keyed items should be reverse scored before conducting any analysis, despite the fact that little empirical evidence justifies this process.

Indeed, whether or not negatively keyed items were reverse scored is a particularly thorny issue when investigating what researchers report in the literature, as many authors remain silent on the matter. To add further complication, little is known about whether and how the presence of negatively keyed items affects exploratory or confirmatory factor analysis. This combination of events leaves the reader uncertain about what to conclude when reviewing a study reporting on the dimensionality and factor structure of a test comprised of both positively and negatively keyed items (i.e., mixed-keyed test).

Tests are classified as either multidimensional (such as the 44-item Big-Five Inventory and Costa and McCrae's [1992] 60-item NEO Five-Factor Inventory) or unidimensional (such as the RSE scale and the PSWQ). This dissertation concentrates on the widely used unidimensional variety of tests, which result in one test score. In this case, the matter of factor structure largely comes down to the investigation of unidimensionality to support the test scoring and interpretation. This dissertation reports on several computer simulation studies that document how exploratory or confirmatory factor analysis (treating the item responses as either continuous

or ordinal) influences the statistical judgment of dimensionality for mixed-keyed tests. The results are used to provide guidelines for researchers.

Structure

The remainder of this dissertation is divided into three chapters. Chapter Two reviews the relevant literature and provides the psychometric background for the four simulation studies reported in this dissertation. The chapter begins with a further clarification of the terminology used to describe item keying and wording and the issues associated with negatively keyed items. In reviewing the extant simulation studies, it became evident that there are two ways of conceptualizing (and hence simulating) item responses to a mixed-keyed test—negative factor loadings and negatively keyed item thresholds—that have not been previously investigated. Given that this dissertation aims to inform day-to-day research practice with short Likert-type tests, it provides a brief review of the current reporting practices of common psychometric analyses to examine the dimensionality and factor structure of tests with mixed-keyed items. Chapter Two ends with a summary of the gaps in the research literature and an overview of the dissertation's four simulation studies.

Chapter Three starts with a short introduction that summarizes the main messages from chapters one and two. It then goes on to discuss the methods and results of four simulation studies. These studies are organized into two sections according to the conceptualization and simulation of item responses for mixed-keyed tests. Within each section, studies using both exploratory and confirmatory methods are presented. Chapter Four summarizes the findings of the four simulation studies by revisiting the research questions listed towards the end of Chapter Two. This final chapter also describes the novel contributions and limitations of this dissertation research, along with its implications for applied researchers and future directions.

CHAPTER TWO:
LITERATURE REVIEW

This chapter describes the issues related to and the methods used in the assessment of the dimensionality and factor structure of short self-report Likert-type tests (i.e., summative tests) with negatively keyed items. While this dissertation focuses on determining unidimensionality to support test scoring and interpretation, the literature review explores a slightly broader range of subjects, covering various studies that report on unidimensional psychological tests with mixed-keyed items. The findings from studies on multidimensional tests with Likert-type response scales are also included because the results regarding negatively keyed items and their impact on factor structure may be informative for short unidimensional Likert-type tests.

As was noted earlier, the assessment of dimensionality and the evaluation of factor structure are interrelated, and the distinction between them is subtle. The essential difference is that the former focuses on determining the number of dimensions (or factors), whereas the latter focuses on their interpretation. It should be noted that the psychometric literature does not always distinguish between these terms very well. In empirical studies, it is rare for researchers to assess test dimensionality without interpreting the resultant dimensions (i.e., factors). In fact, standard best practices recommend that one consider interpretation (i.e., factor structure) when addressing dimensionality (e.g., Flora & Flake, 2017). Thus, the phrase "assessment of factor structure" is often used together with or instead of "assessment of test dimensionality."

This chapter begins by defining negatively keyed items and distinguishing them from negatively worded items. It then reviews the literature investigating issues related to evaluating the factor structure of mixed-keyed tests. This review is organized into two subsections: (a) empirical studies that report on the effect of negatively keyed items on the internal structure of a test, and (b) simulation studies that explore possible explanations for the impact of these items on the assessment of test dimensionality. In light of the data simulation strategies of negatively keyed items used in the simulation studies, two psychometric models and thus two possible ways to conceptualize the item response process for negatively keyed items are discussed.

Given that the remainder of this dissertation will require some understanding of the common statistical methods used in dimensionality and factor structure assessment, a brief introduction to these common methods is provided in this chapter. For a newly developed

unidimensional test or for an existing test that requires re-evaluation, its dimensionality is often ascertained either by (a) conducting an exploratory analysis of the factor structure, which has the enumeration of the factors as its first step; or (b) performing a direct analysis to determine if a test is unidimensional through confirmatory methods. This chapter describes both of these approaches. It then connects the gaps in the research literature to the purposes of the current study. Finally, it presents the research questions, along with an overview of the study design.

Negatively Keyed Items

Self-report measurement instruments suffer from various sources of error, which can potentially be introduced in every aspect of the measurement process. Item characteristics (e.g., item keying and wording), respondent characteristics (e.g., social desirability), and the setting of test administration, just to name a few, are all possible pitfalls. Many of these potential sources of error fall under the general heading of the "method effect," which refers to the systematic variance in the responses (i.e., item scores) that is not explained by the construct of the measurement, but rather is due to the measurement method (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). The meaning of this effect can vary in different studies depending on the sources to which it is attributed. One much-discussed category of the method effect has been referred to as response sets, response styles, and response biases. Some response styles that are often discussed include acquiescence response, extreme response, and socially desirable response. A substantial amount of research pertains to each of these subcategories. A review of the extensive research literature on the method effect as a whole strays from the topic of this dissertation and thus is beyond its scope.

This dissertation draws attention to the item keying effect. Although this effect is often ignored and can be a source of the method effect, the intention of this dissertation is not to demonstrate how the former can contribute to the latter, nor to compare its impact with other sources of the method effect. Rather, as an exploratory study, this dissertation attempts to disentangle the keying effect from the method effect, which is often attributed to item wording. The keying effect and wording effect are defined, and their relationship is discussed in this section. Recall that the goals of this dissertation are to (a) document how the inclusion of negatively keyed items may affect the statistical judgment of test dimensionality, and (b) provide guidelines to assist researchers in decision making when dealing with negatively keyed items.

The remainder of this section focuses on reviewing studies on the dimensionality assessment of tests with negatively keyed items, followed by a discussion of two possible psychometric models for these items.

Definition

The topic of this dissertation is negatively keyed items and their implications for assessing test dimensionality. To facilitate further examination of this subject, a clarification of the terminology is necessary, especially considering that the semantics used to refer to item keying, wording, and social-psychological meaning are inconsistent and intertwined in the literature. Indeed, many terms have been employed to describe items or item stems, including negatively keyed, reverse scored, and reverse coded; positively keyed and directly scored; negatively worded, reverse worded, negative items, and reversed items; and positively worded and positive items (e.g., Colosi, 2005; Curry, Wakefielf, Price, & Mueller, 1986; Ibrahim, 2001; Schriesheim & Hill, 1981; Sliter & Zickar, 2014; van Sonderen, Sanderman, & Coyne, 2013). The above is a list of only those terms that appear often in the literature. Others such as regular items (Wang, Minor, & Wei, 2011), connotatively consistent, and connotatively inconsistent (Chang, 1995a, 1995b) have also been used, though much less frequently. In many instances, these terms are used without a clear definition. Most notably, various studies do not distinguish between item "keying" and "wording" properly (e.g., Colosi, 2005; van Sonderen et al., 2013). In some cases, items that are stated positively but scored negatively are referred to as "negatively worded" items (e.g., Colosi, 2005; van Sonderen et al., 2013) while, in other cases, vague terms, such as "negative items" or "reversed items," are used without differentiating wording from keying.

Throughout this dissertation, keying direction and wording direction are viewed as two distinct item features. Item keying refers to how the responses to an item should be scored or interpreted. Item wording describes the grammatical features of how an item is stated. Following this rationale, the direction of item keying is related to the test score interpretation, but not necessarily to the wording of an item. Sometimes, negatively keyed items are designated as items that are opposite in meaning compared with the majority of the items in a test (e.g., Schmitt & Stults, 1985). This definition is narrow, and it only holds when there are fewer negatively keyed items than positively keyed ones in a test. More generally, a *negatively keyed* item can be defined

as an item whose meaning is opposite to the polarity of the construct being measured. Items are negatively keyed because a negative response to them (i.e., some level of denial or disagreement) indicates a higher level of the measured construct. On the contrary, items are defined as *positively keyed* because a positive response (i.e., some level of agreement) represents a higher level of the construct being measured. Items are keyed in different directions to reflect their meaning in relation to the construct. In this way, the scored item responses can be aggregated to make a meaningful test score. The keying direction of an item depends on the meaning of the item relative to the meaning of the test score, and thus, it cannot be judged in isolation. For ease of interpretation, the conventional practice is to use a higher test score to reflect a higher standing on the construct. Only if the standard or the meaning of a test score is predetermined can the keying direction be distinguished.

Item keying difference can be observed in the item correlation matrix. If all the item responses are left in their original format (i.e., without applying reverse scoring), the item correlation matrix will show a clear pattern that distinguishes items keyed in different directions. The correlations are positive within items keyed in the same direction, but negative between positively and negatively keyed items. An operationalization of negatively keyed items can be that negatively keyed items are those correlate negatively with the total test score and with other positively keyed items in the same test.

By contrast, the wording direction of an item can be judged independently from other items in the same test. This is because it is essentially a grammatical issue that can be evaluated relatively unambiguously. An item can be stated either as an affirmation (i.e., positively worded) or as a denial or disaffirmation of something (i.e., negatively worded) (Horn, 1989). A *positively worded* item refers to one that is grammatically affirmative and contains no negative syntactic markers. Meanwhile, a *negatively worded* item is one that possesses grammatically negative markers that negate or reverse the meaning the sentence would otherwise convey (Holden, Fekken, & Jackson, 1985; Horn, 1989). Negatively worded items can be identified by looking for negation markers in the stems. Some common negation markers include (a) syntactic negations using “not,” (b) syntactic negations using adverbs such as “never,” and (c) reversals accomplished via words containing certain prefixes (e.g., “in-,” “un-,” and “im-”) or suffixes (e.g., “-less”) (Holden et al., 1985).

By adopting the definitions described above, items can be distributed into four categories by cross-classifying the keying and wording directions. These categories are (a) positively worded and positively keyed, (b) positively worded and negatively keyed, (c) negatively worded and positively keyed, and (d) negatively worded and negatively keyed (e.g., Bentler, Jackson, & Messick, 1971; Coleman, 2013; Schriesheim, Eisenbach, & Hill, 1991).

As shown in Figure 2, negatively keyed items may or may not be grammatically negative (i.e., negatively worded), and negatively worded items may or may not be negatively keyed. As an example, imagine that we have an instrument measuring an individual's happiness, which consists of the four types of items. For the purpose of demonstrating some possible items with different combinations of wording and keying directions, the adjectives "sad" and "happy" are considered to be polar opposites. Assume that all the statements (i.e., item stems) are rated on a five-point Likert-type response scale, from strongly disagree (1) to strongly agree (5). For ease of score interpretation, we decide that a higher score on this measure represents a higher level of happiness. In this case, items such as "I am sad" and "I am not happy" can be identified as negatively keyed. This is because disagreement with these statements indicates a greater level of happiness. When examining these two negatively keyed items closely, however, it is evident that they differ in their wording directions. "I am sad" is a positively worded item (from a purely grammatical point of view), while "I am not happy" is negatively worded. Likewise, agreement with positively keyed items such as "I am happy" and "I never feel sad" demonstrates a higher level of happiness. The former item is positively worded, while the latter is negatively worded (see Figure 2). This example shows that keying and wording can be independent item features. Negatively keyed items can be either positively or negatively worded and vice versa. Studies that use ambiguous terms such as "negative" items therefore fail to provide clear information regarding which item feature is being investigated, potentially leading to the misinterpretation of the results.

		Item **Keying**	
		Positive (PK)	Negative (NK)
Item **Wording**	Positive (PW)	PW & PK (e.g., I am happy.)	PW & NK (e.g., I am sad.)
	Negative (NW)	NW & PK (e.g., I never feel sad.)	NW & NK (e.g., I am not happy.)

Note: A possible response scale for the example items may be: 1) strongly disagree, 2) disagree, 3) neutral, 4) agree, and 5) strongly agree. The keying direction of the example items is selected so that higher scores represent higher levels of happiness.

Figure 2.

Cross-classification of items by wording and keying directions

As stated at the beginning of this chapter, this dissertation investigates the effect of item keying on the assessment of test dimensionality and factor structure. Negatively keyed items, that is, the ones presented in the last column on the right side of the table in Figure 2, are of primary interest. The following discussion of the findings in the literature will focus mainly on tests with negatively keyed items. Bear in mind that the terms used in the literature to describe keying direction and wording direction are not always well defined and are usually employed inconsistently and interchangeably. Sometimes it is unclear whether the researchers studied negatively worded items, negatively keyed items, or a combination of both. When reporting the literature, items will be referred to using the terms defined in this chapter when sufficient information is provided. For example, when the items of the test under investigation are presented verbatim, their keying and wording directions can be identified. In such cases, it can be clarified which types of item are under study. When referring to articles in which the correctness of the terminology cannot be fully judged, the same wording used in that paper will also be used here to avoid misinterpretation.

Literature related to negatively keyed items: Empirical studies

A common practice in educational and psychological measurements is to include items keyed in different directions in one measurement instrument. This is often recommended to guard against an individual respondent's test scores being distorted by his/her response styles (e.g., Crocker & Algina, 1986; Nunnally, 1978; Nunnally & Berstein, 1994; Robinson, Shaver, & Wrightsman, 1991; Spector, 1992). For example, acquiescent respondents have a systematic tendency to overuse one side of a response scale (e.g., the agreement side), regardless of the content of the items (Couch & Keniston, 1960; Hui & Triandis, 1985), which leads to inflated or deflated test scores.

Two interrelated arguments have been raised in support of using negatively keyed items to minimize the impact of acquiescence or extreme response styles (e.g., Cloud & Vaugh, 1970; Couch & Keniston, 1960; Martin, 1964; Nunnally, 1978; Wong, Rindfleisch, & Burroughs, 2003). On the one hand, some researchers suggest that respondents' acquiescence to negatively keyed items will offset their acquiescence to positively keyed ones. The belief is that respondents will be forced to consider each item carefully when items are keyed differently. To respond to these items consistently, individuals have to agree with some of the items and disagree with others. Therefore, negatively keyed items serve as cognitive "speed bumps" and consequently control for acquiescence (Kieruj & Moors, 2013). On the other hand, other researchers argue that while negatively keyed items do not eliminate respondents' acquiescence tendency, they reduce its effect by creating test scores that are less extreme (Nunnally, 1978).

Despite the prevalence of the practice, evidence from a preponderance of studies contests including items keyed in different directions in one test (e.g., Barnette, 2000; Lai, 1994; Marsh, 1986; Motl et al., 2000; Pilotte & Gable, 1990; Schriesheim & Hill, 1981). It has been found that the reliability, and in particular the internal consistency, of a test can be adversely affected when positively and negatively keyed items are scored as a single bipolar scale (Barnette, 1997; Wong et al., 2003). Prior studies show that test versions with a single keying direction (i.e., all positively or all negatively keyed) yield higher internal consistency than do the same tests with mixed-keyed items (e.g., Barnette, 2000; Pilotte & Gable, 1990), although this conclusion is not always reached (e.g., Borgers, Hox, & Sikkel, 2004; Finney, 2001; Sauro & Lewis, 2011). This inconsistency in the research findings may be partially attributed to the ambiguity as to whether item wording, item keying, or both effects was studied. As discussed above, negatively keyed

items can be either negatively worded (e.g., “I am not happy”) or positively worded (e.g., “I am sad”), and similarly, positively keyed items can be either negatively worded (e.g., “I never feel sad”) or positively worded (e.g., “I am sad”). Even though negatively worded items have been found to show a tendency to exhibit lower internal consistency than their positively worded counterparts (e.g., Chang, 1995b; Pilotte & Gable, 1990), the internal consistency for negatively keyed items is not necessarily lower than for positively keyed ones (e.g., Schriesheim et al., 1991).

Besides the issue of internal consistency, more critically, the unexpected effect of the item keying direction raises questions regarding the validity of using test scores of mixed-keyed tests either for research purposes or other decisions. Studies in which item wording has been systematically varied across parallel versions of the same measure have generally found that versions featuring a consistent wording direction (i.e., all positively or all negatively worded items), which also resulted in a uniform keying direction (i.e., all positively or all negatively keyed items), yield better-fitting results to the one-factor models than do their mixed-keyed counterparts (e.g., Benson & Hocevar, 1985; Greenberger, Cheng, Dmitrieva, & Farruggia, 2003; Pilotte & Gable, 1990). However, some exceptions have also been noted (e.g., Finney, 2001).

The dominant theme reported among factor-analytic studies of mixed-keyed tests is the emergence of two factors that are essentially separated by item keying direction (positively versus negatively keyed), although these tests are designed to be unidimensional. In research taking an exploratory approach to determining test dimensionality, principal component analysis (PCA) and exploratory factor analysis (EFA) have often been used. When data from a mixed-keyed test that is supposed to measure one construct is analyzed through PCA or EFA, the results frequently suggest a two-component or two-factor solution differentiating items by their keying direction (e.g., Bieling, Antony, & Swinson, 1998; Carmines & Zeller, 1979; Hensley & Roberts, 1976; Steed, 2001).

In studies using confirmatory factor analysis (CFA) to investigate the factor structure of mixed-keyed tests designed to measure a unidimensional construct, a two-factor solution or a one-factor model with a specified method effect arising from item keying direction has been found to show a better fit than does a simple one-factor model (e.g., Corwyn, 2000; Marsh, 1986; Pohl & Steyer, 2010; Tomás & Oliver, 1999). Published research based on various target

populations and measures has repeated this finding (e.g., Barnette, 2000; Corwyn, 2000; Motl et al., 2000; Pilotte & Gable, 1990; Schriesheim & Hill, 1981).

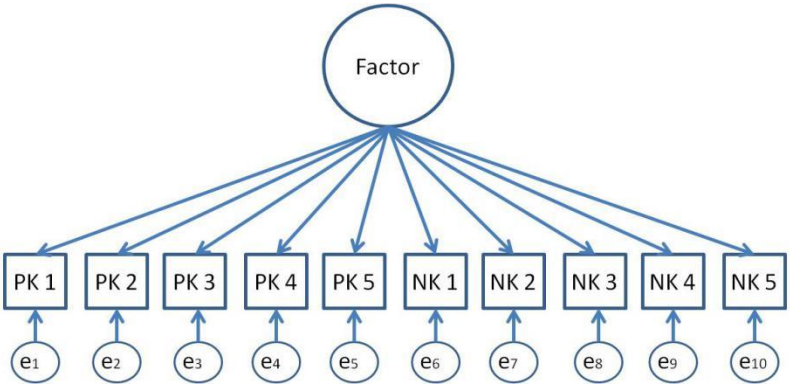
There are many variations in how these alternative factor models are proposed and utilized to account for the item keying effect. However, it is still confusing for scholars to decide which one is appropriate for their data and research purposes. Many studies with different measures and samples have compared these models. Researchers have generally found that the factor structure of their measures improves when they load items on different factors based on their keying direction or use models that account for the covariances among items with the same keying direction (e.g., Carmines & Zeller, 1979; DiStefano & Motl, 2006; Tomás & Oliver, 1999). However, the results regarding which model can best represent the factor structure of mixed-keyed tests are inconsistent (e.g., DiStefano & Motl, 2006; Marsh, 1996; Tomás & Oliver, 1999). This is partly because the fit statistics of these competing models follow closely with each other and often are not directly comparable.

Much of the research examining the use of mixed-keyed tests has centered on self-esteem using the Rosenberg Self-Esteem (RSE) scale. Therefore, the findings on the RSE scale were chosen as examples to demonstrate the alternatives to the one-factor model that have often been suggested in the literature. By reviewing these findings, we can get a better sense of some of the challenges and confusions researchers may face in their day-to-day use of tests with items keyed in both directions.

The RSE scale (Rosenberg, 1965) is widely used in assessing individuals' self-esteem. The original version of this test contains ten items, half of which are positively keyed and the remainder of which are negatively keyed. The initial version uses a four-point Likert-type response scale ranging from "strongly disagree" to "strongly agree," although other researchers have employed scales with different rating points and formats. Despite the differences in the presentation of the response scales, a summated score based on the responses to all ten items is used to quantify an individual's level of self-esteem.

The one-factor model (Model 1, see Figure 3), which is the basis for relying on this scoring method and on a summated total score, is often not supported by the statistical assessment of dimensionality. To better describe the factor structure of the RSE scale, at least eight alternative models have been proposed (see Tomás & Oliver, 1999). One of these models is

unique to the RSE scale and the study of self-esteem, while the others have been commonly suggested to handle tests with mixed-keyed items.

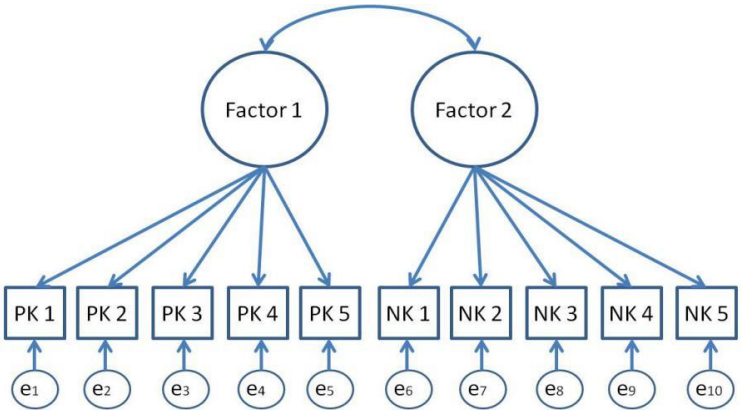


Note: PK stands for positively keyed, and NK stands for negatively keyed.

Figure 3.

Model 1: One-factor model

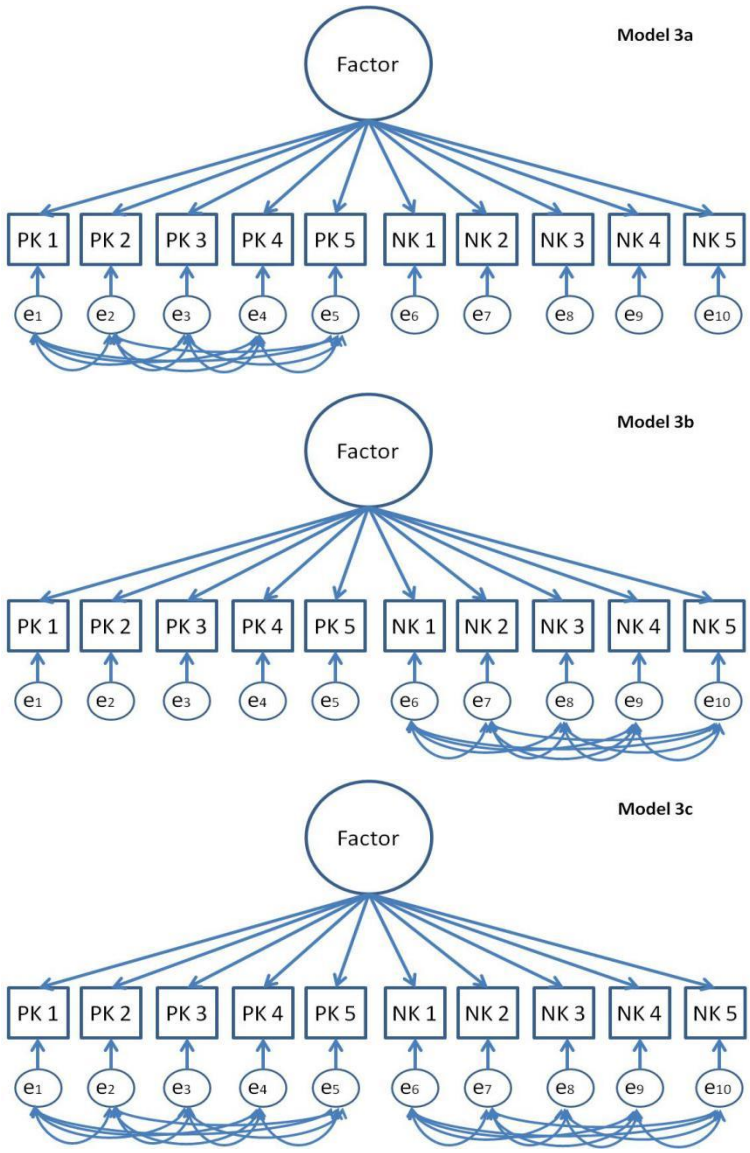
The seven models that have often been used as alternatives to describe unidimensional tests with negatively keyed items are classified into three categories: (a) a two-factor model whose factors are defined by keying directions (Model 2, see Figure 4); (b) three variations of the correlated uniqueness model (models 3a-3c, see Figure 5), including a one-factor model with correlated error terms among positively keyed items (Model 3a), a one-factor model with correlated error terms among negatively keyed items (Model 3b), and a one-factor model that allows for correlated error terms within both positively and negatively keyed items (Model 3c); and (c) three types of bi-factor models (models 4a-4c, see Figure 6), which mimic the three variations of the correlated uniqueness model. In these bi-factor models, the factors related to item keying directions are incorporated as distinct factors, and the correlations between substantive factors and keying factors are assumed to be zero. The factor loadings of a keying factor are typically allowed to differ, implying that each item can be influenced by the "keying effect" factor to a varying degree. The eighth model is somewhat unique to the ten-item RSE scale. It is a variation of the two-factor model, in which only two of the negatively keyed items form the second factor while all the other items load on the first (Model K, see Figure 7; proposed by Kaufman et al., 1991, as cited in Tomás & Oliver, 1999).



Note: PK stands for positively keyed, and NK stands for negatively keyed.

Figure 4.

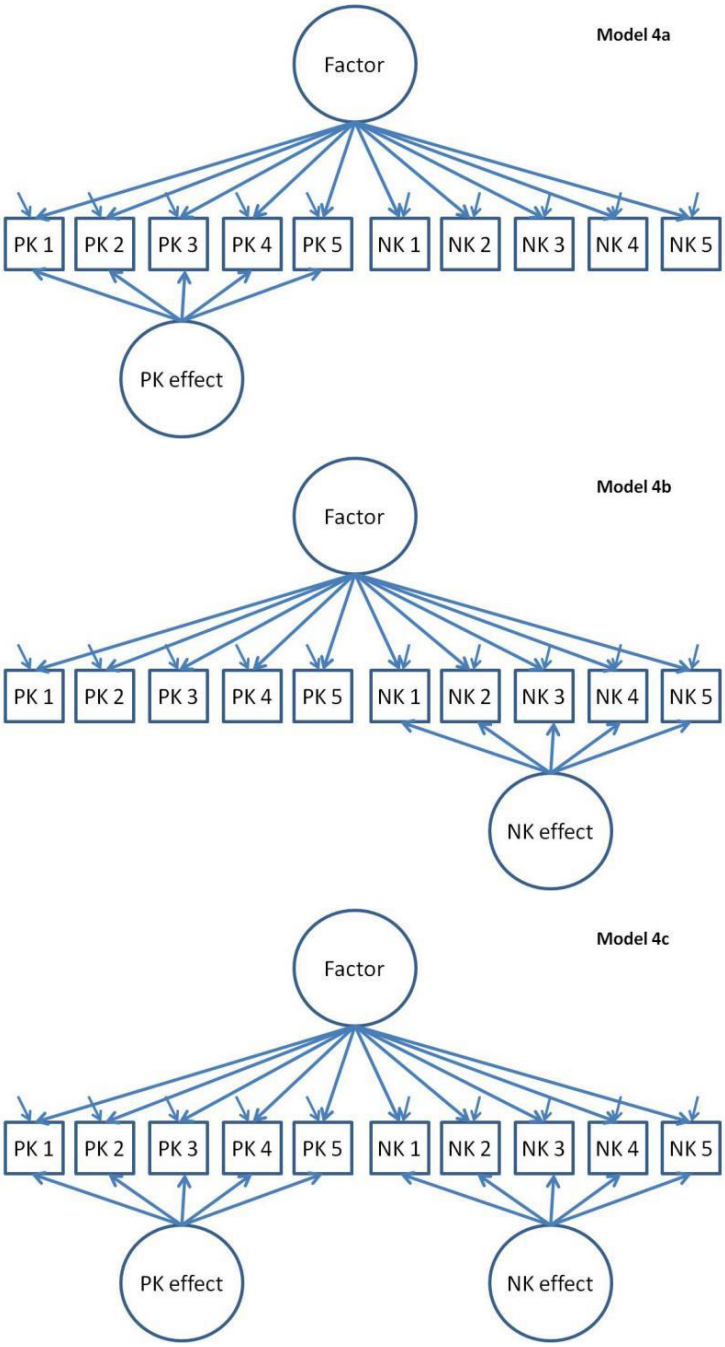
Model 2: Two-factor model



Note: PK stands for positively keyed, and NK stands for negatively keyed.

Figure 5.

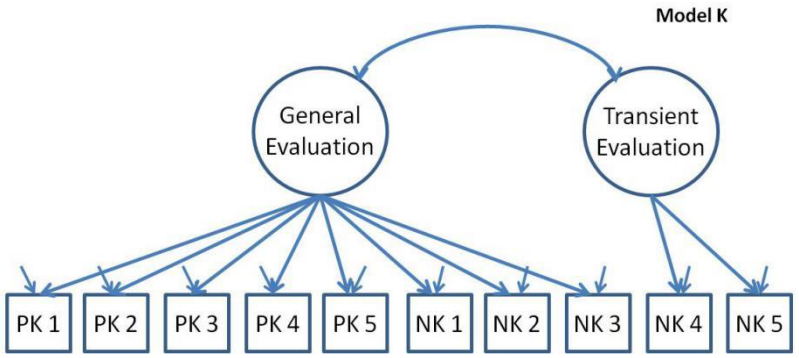
Model 3: Correlated uniqueness model



Note: PK stands for positively keyed, and NK stands for negatively keyed.

Figure 6.

Model 4: Bi-factor model



Note: PK stands for positively keyed, and NK stands for negatively keyed.

Figure 7.

Model K: Model proposed by Kaufman et al., 1991 (cited in Tomás & Oliver, 1999)

The earlier studies tend to focus on a limited number of competing models, including a one-factor model (see Figure 3) and a two-factor solution based on item keying directions (see Figure 4; e.g., Ebesutani et al., 2012; Marsh, 1996; Motl et al., 2000). Later studies compare a wider range of models. These recent models can be placed into two broad categories: correlated uniqueness models and bi-factor solutions (e.g., Ebesutani et al., 2012; Wang, Chen, & Jin, 2014). Three common variations of the correlated uniqueness model (models 3a-3c) are illustrated in Figure 5. These models all have one factor representing the construct intended to be measured, while the error terms of items with the same keying direction are correlated to represent the keying effect.

In DiStefano and Motl's (2006) study, the factor structure of the original ten-item RSE scale (Rosenberg, 1965) was compared via six models. They include a single-factor model (Model 1; see Figure 3); a two-factor model, with separate factors for the positively and negatively keyed items (Model 2; see Figure 4); and models with a self-esteem factor that accounted for the keying effect (i.e., models 3a, 3b, 4a, and 4b; see Figure 5 and Figure 6). They concluded that the one-factor model with correlated error terms among negatively keyed items (Model 3b) was the best fitting model. All the other models, except for the single-factor model (Model 1), achieved similar levels of model fit, as evidenced by almost identical fit statistics.

Tomás and Oliver (1999) further expanded the number of possible alternative models. They compared nine different models to examine the factor structure of a Spanish version of the same ten-item RSE scale (Rosenberg, 1965). The nine models tested included the eight

alternative models described in the earlier paragraphs and the one-factor model (see Figures 3-7). The results showed that models that included the method effects among both positively and negatively keyed items (models 3c and 4c) were the best-fitting models. Both the correlated uniqueness model and bi-factor model exhibited an excellent and nearly equal fit to the data.

Similarly, six models were compared in Marsh's (1996) study, which utilized a seven-item version of the RSE scale with three negatively keyed and four positively keyed items. The six models considered were (a) a single substantive self-esteem factor model (Model 1; see Figure 3 as an example), (b) a model with two factors representing either positively or negatively keyed items (Model 2; see Figure 4 as an example), (c) three variations of the correlated uniqueness models to account for the method effect (models 3a, 3b and 3c'; see Figure 5 as an example), and (d) a model with two factors representing general and transient self-evaluations (Model K, see Figure 7). Based on the fit statistics, it was concluded that the seven-item RSE scale was best described by a one-factor model with correlated uniqueness among negatively keyed items and selected positively keyed items (Model 3c', a variation of Model 3c). Unfortunately, neither the rationale nor the procedure was clear about which items among the positively keyed ones should be chosen to allow for correlated error terms.

The above examples of alternative factor models are taken from studies on the RSE scale. They contain the factor models recommended in the literature to describe the factor structure of mixed-keyed tests with CFA approaches. These factor models, including the two-factor model, correlated uniqueness models, and bi-factor models, are selected to represent the most typical models used in describing the structure of tests with negatively keyed items. Similar findings regarding the dimensionality and factor structure of tests with negatively keyed items have been observed across various measures. Most of these studies focus on measures that are purported to be unidimensional. Examples include the RSE scale (e.g., Carmines & Zeller, 1979; Marsh, 1996; Tomás & Oliver, 1999; Whiteside-Mansell & Corwyn, 2003), the revised Life Orientation Test (LOT–R; Scheier, Carver, & Bridges, 1994), the Social Physique Anxiety Scale (SPAS; Motl & Conroy, 2000; Motl et al., 2000), the Positive and Negative Affect Scale (PANAS; Bagozzi, 1993), the Penn State Worry Questionnaire (PSWQ; Hazlett-Stevens, Ullman, & Craske, 2004), and so on.

In summary, the empirical studies suggest that the more complex models, rather than the single-factor model, more adequately capture the factor structure underlying the data of mixed-

keyed tests. Unfortunately, the interpretation of the alternative models may not be straightforward and can pose challenges to the understanding of the construct and the subsequent analyses of the test scores. Two common interpretations have been proposed for these alternative factor structures. One treats the additional factors in the model as substantive factors that have their own meanings. The other treats the emergence of a second factor or the appearance of correlations between some items as an artifact.

The two-factor solution is a common alternative model that researchers may consider when a one-factor model is rejected. This solution treats positively and negatively keyed items as measuring two distinct but correlated latent constructs. When researchers choose this two-factor model over others, they are likely to give the factors substantive meaning. In the two-factor model, researchers interpret the second factor or the one defined by negatively keyed items as the polar opposite of the underlying construct. Take the revised Life Orientation Test (LOT-R; Scheier et al., 1994) as an example. The LOT-R was developed to assess the construct of dispositional optimism, which is defined as positive outcome expectancies (Scheier et al., 1994). Although dispositional optimism was conceptualized as a unidimensional concept at the beginning of the scale's development (Scheier & Carver, 1985, 1987; Scheier et al., 1994), the response data in a number of studies indicated that the positively and negatively keyed items split into two factors (e.g., Creed, Patton, & Bartum, 2002; Herzberg, Glaesmer, & Hoyer, 2006; Lai & Yue, 2000). These factors observed with the LOT-R have been interpreted as two independent constructs, named as optimism and pessimism, rather than as a single trait (e.g., Herzberg et al., 2006). When researchers give factors substantive meanings, the implied theoretical stance is that they are distinct constructs or dimensions rather than the opposite ends of a continuum. Interpreting the results as two substantive factors/dimensions changes the conceptual definition of the construct that is originally hypothesized to be unidimensional; therefore, researchers must be cautious in making such an interpretation.

The other common way to construe these alternative factor models is to treat the second factor (usually the one defined by negatively keyed items) as a method effect. Researchers who do this conclude that the two-factor structure underlying the response data is a result of a single meaningful dimension that is contaminated by a method effect or artifact (e.g., the keying and/or wording effect) (e.g., Carmines & Zeller, 1979). This method effect interpretation is widely employed to understand the results from the correlated uniqueness models (e.g., Bachman &

O'Malley, 1986) and bi-factor models (e.g., Innamorati et al., 2014). Limited work has been done to understand the mechanisms that produce the artifactual factors. Reasons that have been suggested for the emergence of the second factor include (a) a lack of reading ability to process negatively worded items (e.g., Cordery & Sevastor, 1993); (b) careless responses to negatively keyed items (e.g., Schmitt & Stults, 1985); and (c) response style, such as acquiescence responses (e.g., Savalei & Falk, 2014).

Correlated uniqueness models are usually used to control the method effect associated with item keying direction, while bi-factor models are often used in studies where the method effect is considered an attribute of interest (e.g., DiStefano & Motl, 2006; Motl et al., 2000). For example, the method effect represented in bi-factor models has been conceptualized in association with response style (e.g., DiStefano & Motl, 2006), which is defined as a personality trait that is consistent over time and across measures (Bentler et al., 1971). Regarding the method effect interpretation in either correlated uniqueness or bi-factor models, fundamental questions arise about the nature of the variance associated with some but not all of the items in a test. For example, it is unclear whether this variance reflects the effect of item keying or is a general cognitive/psycholinguistic phenomenon associated with item wording. It is also unclear to what extent such variance might be viewed as a function of item features, individual characteristics, or the interaction between the two, and whether this variance is stable or transient. These possibilities do not seem to be mutually exclusive (e.g., Holden et al., 1991; Tourangeau & Rasinski, 1988). Not understanding the nature of the method effect renders the interpretability of the factors extracted from a supposedly unidimensional mixed-keyed test ambiguous.

Besides the confusion surrounding factor meanings, researchers may be unsure which factor model is appropriate for their data. None of these complex models seems to be clearly superior to the others in terms of model fit and interpretability. Although the test is designed to follow a one-factor structure, a two-factor model or the other one-factor models controlling for the method effect can still be possible alternatives a researcher might consider in advance, or turn to if the one-factor model fits the data poorly. If the positively and negatively keyed items genuinely belong to two factors with different substantive meanings, the one-factor model would seem inappropriate when used to investigate the factor's relationship with other external variables.

In summary, empirical studies focused on the internal structure of a test are often insufficient to confirm the meaning of the factors that emerge in the dimensionality assessment. Also, they are unable to disentangle the effects from different sources on test dimensionality assessment. Without knowing the "true" structure of a test, it is hard to judge which model is the right choice and how the covariance among the mixed-keyed items may change the statistical conclusions regarding dimensionality. For this reason, the following subsection reviews simulation studies on negatively keyed items and their effect on assessing test dimensionality.

Literature related to negatively keyed items: Simulation studies

Empirical studies suggest that negatively keyed items are suspected to introduce construct irrelevant variance and may lead to the emergence of a second factor when the test is designed to be unidimensional. The literature has explored various reasons for this phenomenon (e.g., Cordery & Sevastor, 1993; Schmitt & Stults, 1985; Spector et al., 1997; Woods, 2006). The factors that can potentially affect the assessment of test dimensionality can be attributed to three primary sources: (a) the features of a test, (b) respondents' characteristics, and (c) the statistical methods used to assess dimensionality. This lack of transparency regarding the interplay among these sources makes it challenging to draw conclusions based on findings from empirical research. Simulation studies are better suited to separating the effects from different sources, including item keying and wording.

Unfortunately, to the best of my knowledge, no simulation study has been conducted to systematically examine the keying effect on the statistical judgment of test dimensionality as informed by various analytic methods. A few such studies have explored some aspects of the effect of item characteristics, including item keying and wording, as well as respondent characteristics on the factor structure of a test (Schmitt & Stults, 1985; Spector et al., 1997; Woods, 2006). These simulation studies differ in their focus, study designs and procedures. They show that simulation is a promising method for exploring item keying effect, but such studies are rare and many issues remain unresolved.

Schmitt and Stults (1985) explored a situation in which a subpopulation misread the negatively keyed items and responded to them as if they were positively keyed. They examined how such a situation may affect the factor structure of a test. Their study used factor loading matrices, based on item correlation matrices from real datasets, to generate item responses. They

employed three correlation matrices to replicate a variety of tests that are different in item intercorrelations. One of the correlation matrices represents the item intercorrelations of a unidimensional test, and the other two represent the intercorrelations of multidimensional tests. Negatively keyed items were indicated by negative factor loadings in the data generation model. The same thresholds were applied to the latent response distribution to obtain the observed response categories. The item responses were generated on a seven-point Likert-type scale with an asymmetric distribution. Before analyzing the data, all the responses to the negatively keyed items were reverse scored, as this is what applied researchers normally do. A group of careless respondents was created during this recoding process by leaving their responses unrecoded. The researchers subsequently analyzed the generated data, treating the ordinal item responses as continuous. Exploratory principal component analysis (PCA) was performed, and the findings suggest that a factor defined by negative keying direction can emerge with only 10% of the respondents misresponding to the negatively keyed items.

Woods (2006) also concentrated on the effect of careless responses on the assessment of model fit via confirmatory factor analysis (CFA). This study utilized a unidimensional two-parameter logistic item response theory (2PL IRT) model to generate dichotomous item responses to a unidimensional test. Ten items out of twenty-three (about 43%) were negatively keyed. Unlike Schmitt and Stults (1985), however, Woods (2006) simulated careless respondents by switching their responding categories (i.e., 0-1 and 1-0) on negatively keyed items. The item discrimination parameters used to generate the response data for both positively and negatively keyed items were all positive in the IRT models. Although not stated directly, the responses to negatively keyed items were probably generated through the reversed relationship between thresholds and observed response categories. The subsequent CFA analysis was carried out with weighted least squares means and variance adjusted (WLSMV). Consistently with Schmitt and Stults (1985), Woods (2006) concluded that, with relatively few individuals (10% of the total sample) misresponding to ten negatively keyed items in a 23-item test, a one-factor model could be wrongly rejected by the CFA model fit statistics.

Although both of these studies focused on careless responses to items that are keyed or worded differently from the majority, Schmitt and Stults (1985) referred to such items as negatively keyed while Woods (2006) called them reverse worded. Both studies defined a careless respondent as someone who read a few items in a test and inferred that all the rest were

stated in the same direction, causing him/her to respond all the items in a similar manner (Schmitt & Stults, 1985; Woods, 2006). In light of this definition, it seems more appropriate to attribute the results to the misreporting of negatively worded and negatively keyed items (e.g., "I'm not happy" in a test measuring happiness). It is possible that a person might ignore syntactic negation markers, such as "not," that negate the meaning of a statement and responds as if that item was positively worded. However, the plausibility is low for a careless respondent to misread an item that is positively worded but negatively keyed (e.g., "I am sad" in a test measuring happiness) and respond in an opposite manner than they should.

Schmitt and Stults (1985) and Woods (2006) examined the impact on test dimensionality of having a subpopulation that consistently provides misresponses to negatively worded and negatively keyed items. Unlike these researchers, Spector and colleagues (1997) investigated the effect of item extremity on the dimensionality assessment of mixed-keyed tests. Spector and colleagues (1997) also relied on a different response process model for data generation. Although Schmitt and Stults (1985) used factor models and Woods (2006) used IRT models in their data simulation, both assumed that the observed item responses could be attributed to an underlying response model with continuous and normally distributed latent response functions. The latent response distribution underlying each item could be dichotomized at each threshold representing a response category. If persons and item response categories are put on a continuum reflecting the construct of interest, the person will tend to endorse a response category when his/her standing on the latent trait is higher than the response category. The literature sometimes refers to this as the dominance response process (Coombs, 1964; Likert, 1932). In contrast, Spector and colleagues (1997) used the ideal point principle to guide their data simulation. The ideal point process assumes that a person endorses an item or a response category only when he/she is near to the standing of the response category on the continuum of the attribute under investigation (Cliff, Collins, Zatkin, Gallipeau, & McCormick, 1988; Thurstone, 1928). In other words, in the ideal point process, people do not pick an option if they are much lower or much higher on the latent trait continuum relative to that response category.

Spector and colleagues (1997), assumed that the latent response distribution underlying all the items was the same, but that the relationships between the latent trait and observed response categories varied depending on item extremity and keying direction. They defined extreme items as those that respondents found difficult to endorse. Item extremity was

manipulated through different transformation functions from latent response distribution to observed response categories. The authors argued that, while the relationship between the latent trait and observed response categories was linear for moderate items, it was curvilinear for extreme items, with a linear part close to one end of the continuum but mostly flat in other parts.

Figure 8 shows the theoretical item characteristic curves for the moderate and extreme positively keyed items employed by Spector and colleagues (1997). For extreme items, respondents use only part of the response scale, which means their responses are clustered on one side. For positively keyed items that are extreme, the responses fall on the disagreement side of the response scale, while for negatively keyed extreme items, the responses are skewed towards the agreement side. The observed responses were simulated on a six-point response scale, and both CFA and EFA were conducted to explore the factor structure of the simulated datasets. Spector and colleagues (1997) concluded that the emergence of factors formed by item keying direction could be artifactually produced by different responding patterns associated with item extremity.

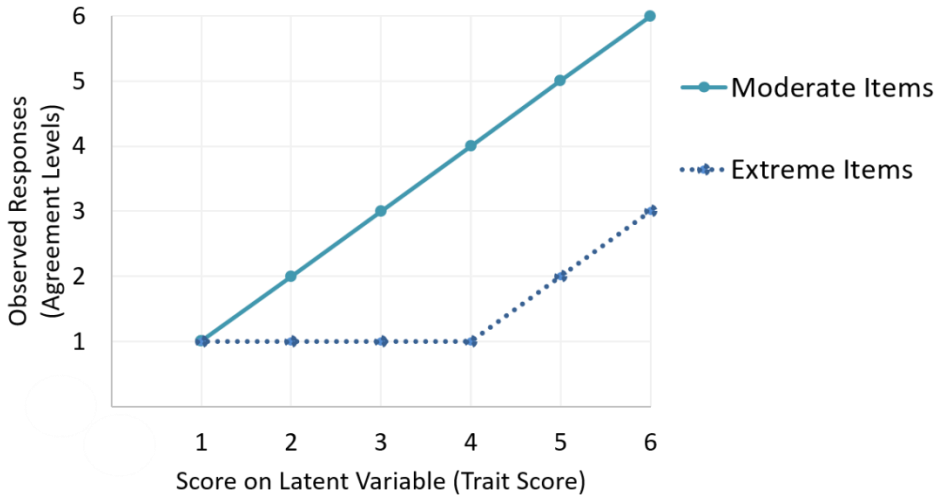


Figure 8.
Characteristic curves for (moderate vs. extreme) positively keyed items

Taken together, these three simulation studies suggest that when respondents react inconsistently to items that are keyed in different directions, either due to “carelessness” or to item extremity (e.g., floor effect or ceiling effect), the expected factor structure of a test may not be supported by the statistics from factor analysis.

A couple of other simulation studies that may seem relevant are not included in the above review. These studies also simulated tests with negatively keyed items, but are not discussed in detail here because (a) they did not study the keying effect, but rather, some type of response style that may coexist with negatively keyed items (e.g., Savalei & Falk, 2014); or (b) they assumed that the inclusion of negatively keyed items would lead to a method effect, and studied the consequences of ignoring it (e.g., Gu, Wen, & Fan, in press).

The findings of these simulation studies are important for applied researchers to consider when interpreting the results of a dimensionality assessment. However, studies of this sort are scarce, and have investigated only a limited number of factors that may affect test dimensionality. The lack of understanding of how the presence of negatively keyed items may affect the assessment of test dimensionality, combined with the dearth of research, calls for a simulation study that systematically investigates the effect of item keying on the evaluation of dimensionality and factor structure.

Two psychometric models for generating responses to negatively keyed items

As mentioned in the review of the simulation studies, different models have been used to generate data that represents item responses (see Gu et al., in press; Savalei & Falk, 2014; Schmitt & Stults, 1985; Spector et al., 1997; Woods, 2006;). In what follows, I first briefly describe the fundamentals of threshold models, which are often used to scale item responses in psychological testing. Models based on item response theory (IRT) and item factor analysis are often described as part of this category. Next, I discuss two possible psychometric models for negatively keyed items under the framework of factor analysis. This discussion will serve as the background and rationale for the simulation methodology used in this dissertation.

Threshold models, also known as latent response variable models with categorical data, are commonly used mathematical ways to statistically model item responses (e.g., Muthén, 1983; Muthén & Asparouhov, 2002; Schmitt & Stults, 1985; Zumbo, Gadermann, & Zeisser, 2007). They adopt the dominance item response process, which follows the assumption of Likert-type rating scales that higher agreement with positively keyed items indicates a higher level of the measured attribute (Likert, 1932). Threshold models assume that each item has a latent response distribution (y^*) and an observed response distribution (y). The observed response (y) comes from a latent response distribution (y^*) underlying that item. Although the observed responses

are usually within a limited number of categories, the latent response distribution is continuous and, for computational convenience, is usually assumed to follow a normal distribution. The observed response distribution (y) is determined by the latent response distribution (y^*) and the thresholds.

Thresholds reflect the position on the underlying continuous and normally distributed variable (y^*) that distinguishes a category of the observed variable (y). The relationship between a latent response distribution y^* and an observed ordinal distribution y can be expressed as

$$y = c, \text{ if } \tau_c < y* < \tau_{c+1},$$

with thresholds τ_c as parameters defining the categories $c = 1, 2, \ldots, C - 1$, where $\tau_0 = -\infty$ and $\tau_c = +\infty$ (Liu, Wu, & Zumbo 2010; Muthén, 1983). The value for an observed ordinal response (y) changes when the latent response variable y^* exceeds a threshold τ_c value. It assumes a perfect correspondence between the observed item response and the construct being measured.

In threshold models, the construct shapes the latent response distribution underlying each item that assesses it. For each item, the observed responses are a manifestation of its latent response distribution. Following the threshold model, responses to negatively and positively keyed items can vary either due to different latent response distributions (y*) or different thresholds. The two psychometric models for negatively keyed items also reflect different ways of conceptualizing the item response process to these items. As Hubley, Wu, Liu, and Zumbo (2017) argue, parameters in a measurement model carry information or assumptions about the item response process. Thus, the following discussion attempts not only to describe the psychometric model of data-generation for the simulation study, but also to make connections between the psychometric models and how people respond to items.

Negative factor loading model for negatively keyed items

On the one hand, responses to negatively keyed items can be generated through negative factor loadings in the latent response model. That is, the latent response distributions (y^*) have different relationships with the construct for items that are keyed in different directions. Some simulation studies have used this strategy (e.g., Savalei & Falk, 2014; Schmitt & Stults, 1985). For positively keyed items, the factor loadings in the latent response model are positive, indicating a higher standing on the construct associated with a higher standing in the latent

response distribution (y_p*). For negatively keyed items, the factor loadings are negative in the latent response model. A higher standing on the construct leads to a lower value on the latent response distributions (y_n*). However, in the transformation stage from latent response *y** to observed response categories *y*, *y* is coded so that its relationship with the latent response (*y**) will be the same for both positively and negatively keyed items. In other words, lower values of the latent response (*y**) always correspond to lower values of the observed response (*y*).

One way to link this psychometric model to the items and tests we use in day-to-day research is to relate it to positively worded but negatively keyed items. Following the rationale of this model, an item is keyed negatively because it assesses a negative aspect or the polar opposite of the construct. For example, in the RSE scale (Rosenberg, 1965), an item states, "All in all, I am inclined to feel that I am a failure." It is reasonable to argue that respondents may draw on their experiences of "lack of self-esteem" that are prompted (or cued) by the key word "failure" in the item stem. Thus, this item can be seen as measuring "lack of self-esteem," and its underlying latent response distribution is negatively correlated to the distribution of the measured construct of interest. A higher standing on the latent response distribution of "lack of self-esteem" (y_n*) leads to a higher agreement with this statement. In turn, a higher standing on the latent response distribution (*y**) corresponds to a higher agreement category (*y*), and this positive relationship between *y** and *y* is consistent for both positively and negatively keyed items.

Reversed threshold model for negatively keyed items

On the other hand, responses to negatively keyed items can come from a reverse response process arising from a latent response distribution (*y**) to the observed response categories (*y*). Such an item is negatively keyed to accommodate the reversed relationship between the latent and the observed response. This simulation strategy has also been used in the research literature (e.g., Woods, 2006). This model for negatively keyed items applies thresholds with reversed correspondence relationships to the observed responses of positively and negatively keyed items. Indeed, these items are indistinguishable in the latent response model. The latent response distributions of all the items are assumed to be positively correlated with the construct (i.e., positive factor loadings in the latent response model). Negatively keyed items are separated from positively keyed ones by specifying the relationship between thresholds and the corresponding observed response categories differently. This means that, for positively keyed items, a higher

value on the latent response (y^*) corresponds to a higher agreement response category; for negatively keyed items, a higher value on the latent response (y^*) corresponds to a lower one.

If we take an item from the RSE scale (Rosenberg, 1965) as an example, a possible interpretation of this psychometric model in terms of its implications on the response process to negatively keyed items can be described as follows. One of the negatively keyed items in the RSE scale states, "I feel I do not have much to be proud of." Respondents need to pick an answer from a four-point Likert-type response scale ranging from "strongly disagree" to "strongly agree." Instead of assuming this item is assessing a "lack of self-esteem" as the first psychometric model does, this model assumes that this item is measuring pride or high self-esteem, just like other positively keyed items. Because the word "proud" in the stem connotes a possession of self-esteem, it is reasonable to assume that respondents will rely on their experience of "high self-esteem" to answer this item. Thus, the underlying response process is better characterized by a latent response distribution of y^* representing "high self-esteem." That is, a higher value on the y^* scale represents a higher level of self-esteem.

Since the relationship between the latent response distribution (y^*) of this item and the construct of self-esteem is specified to be positive, the factor loading between this item and the construct (i.e., factor) is also positive in the latent response model. However, in this item, the negation marker "not" changes the relationship between the latent response (y^*) and the observed response (y). Without this negation marker (i.e., "I feel I do have much to be proud of"), the relationship between the latent response (y^*) and the observed response (y) should be positive. That is, a higher value on the latent response distribution would correspond to a higher level of agreement with the statement. By adding the negation marker, the relationship between the latent and the observed responses will be reversed. A higher standing on the latent response distribution of this item will now be associated with a higher level of disagreement. Namely, an individual with a higher standing on the latent response distribution (i.e., high self-esteem) will have a lower level of agreement with this item and thus a lower value on the observed response scale (y).

These two psychometric models potentially depict two distinct conceptualizations of responses to negatively keyed items, and they may be helpful in differentiating between two types of these negatively keyed items: positively worded or negatively worded. Both models have been used in simulation studies to generate item responses to tests with negatively keyed

items, but the potential difference between these models and their implications for the item response process have not been discussed. It is important to acknowledge that these psychometric models are mathematical abstractions of the phenomena of interest. Existing evidence is insufficient to directly link these models and the real cognitive process that individuals employ in their response to negatively keyed items. The actual response process is more complicated and can vary depending on item properties, individual characteristics, and the context of item responding. The examples provided above are meant to be used as a way of understanding the process of responding to negatively keyed items under the two psychometric models (i.e., simulation strategies). Without additional evidence from the cognitive processes underlying item responses, it cannot be concluded that the two types of negatively keyed items that differ in their wording direction are each represented by one of these two psychometric models. It is noteworthy that the direction of reasoning is important in this context. The psychometric models may imply a type of item responding but the item responses (on their own) do not necessarily imply a psychometric model exclusively. As a general problem in psychometrics, the data alone do not dictate the model; psychometric modeling is an interplay of theory, model, and data.

Common Methods in Assessing the Dimensionality of Tests with Negatively Keyed Items in Validation Practice

Investigating the dimensionality or factor structure of data collected from a test is *de rigueur* in day-to-day research. The factor structure of item response data is composed of a certain number of factors and the relationships among them. The factor structure of a test is usually reported as evidence to support the use or interpretation of the score (e.g., Thompson & Daniel, 1996; Zimprich, Kliegel, & Rast, 2011). To determine the factor structure of a test, researchers typically use factor analysis, which is usually conducted either through exploratory or confirmatory approaches, depending on the study design and purpose.

This section provides a summary of the literature that describes different methods to determine factor structure. It is not meant to be a comprehensive review or a complete step-by-step tutorial, but is rather an outline of some important decisions researchers must make when assessing factor structure. It begins with a summary of a general scoring method for negatively keyed items. The rest of the section is organized according to two themes in the investigation of

test dimensionality and factor structure, namely, exploratory approaches and confirmatory approaches. First, it explains the rules and criteria used to inform the decision on the number of factors. This is followed by a brief discussion of exploratory factor analysis (EFA). Finally, it presents a summary of confirmatory factor analysis (CFA) approaches for evaluating a specified factor structure.

Scoring methods applied to negatively keyed items

One goal of a measurement instrument, such as self-report Likert-type tests, is to quantify the construct of interest. To do this, the original responses obtained from a mixture of positively and negatively keyed items must be scored appropriately.

Scoring the responses is usually the first step researchers need to perform before conducting any other analyses. The most common way of handling a mixed-keyed test is to code the responses to negatively keyed items in reverse order and then treat them in the same way as the responses to positively keyed items (e.g., DiStefano & Motl, 2006; Greenberger et al., 2003; Horan, DiStefano, & Motl, 2003). After reverse scoring the responses to these negatively keyed items, for all the items, a relatively large value of a response represents a high level of the construct being measured. Then the test score can be computed as a total, an average, or a factor score based on the scored responses to all the items.

An example is provided here to demonstrate the reverse scoring or recoding process. If all the items are answered on a five-point Likert-type response scale, then for negatively keyed items an original answer of five (strongly agree) is recoded to a score of one (strongly disagree), a four (agree) is recoded to a two (disagree), a three (neutral) remains the same, a two (disagree) is recoded to a four (agree), and a one (strongly disagree) is recoded to a five (strongly agree). A simple mathematical rule can be used to summarize the reverse scoring process:

$$\text{Reverse score } (y) = \max(y)+1-y,$$

where y is the original response score of a negatively keyed item, and max(y) is the maximum possible value for y (i.e., the total number of response categories on the response scale). In the above example of a five-point rating response scale, max(y) is five because the responding scale only goes up to five. To reverse score, take 5+1 = 6, and subtract the number indicating the original response. For example, the reversed score for a four (agree) would be two (disagree); the

formula used for this transformation can be written as reverse score (4, agree) = 5+1-4 = 2 (disagree).

Intuitively, this reverse scoring process numerically transforms the original responses to the negatively keyed items so that relatively large values, regardless of item keying directions, represent high levels of the construct being measured (Furr & Bacharach, 2013). This procedure assumes that when the negatively keyed items are reverse scored, all items should be psychometrically indistinguishable or interchangeable. However, the literature offers insufficient empirical evidence to support this crucial assumption. If the scoring method itself introduces construct-irrelevant variance or covariances, and does not reflect the true variance and covariance structure of the responses under investigation, minor factor(s) may appear when in fact they do not reflect the structure of the construct. For example, reverse scoring changes the item response distributions of some of the items in a test (e.g., after reverse scoring, a positively skewed distribution becomes negatively skewed), which may affect the subsequent analysis results. Another possibility is that reverse scoring of the negatively keyed items may not always be necessary. For example, the factor loadings in factor analysis can be either positive or negative, which can reflect the keying difference of items. If an item yields a negative factor loading, that item is negatively related to the factor, and it is considered to be negatively keyed relative to the direction of the factor score.

Exploratory methods to determine the number of factors

Exploratory approaches are often used when there is no *a priori* theory about the dimensionality of the measured construct or about which items should load on each of the factors. EFA aims to account for the shared variance of a set of observed variables (e.g., items) by a small number of common factors. When used for validation purposes, EFA is often conducted as an initial assessment of the factor structure of a test in its development or revising stage. EFA is a complex, multi-step process. Although this dissertation mainly concentrates on the decision of the number of factors, other analytical decisions, such as estimation methods and rotation methods, must be made to conduct an EFA. Choices made on these key issues have an impact on the number of factors to be extracted and the interpretation of the results (Armstrong & Soelberg, 1968; Comrey, 1978; MacCallum, 1983; Weiss, 1976). Estimation methods will be discussed in the review of CFA, which is the subsection following this one. The current subsection will

briefly describe the methods used to inform the decision of the number of factors and some commonly used rotation methods in EFA.

Choosing how many factors to retain is essential in exploring the factor structure of a test, and it is usually the first decision a researcher must make in this exploratory process. The decision of the number of factors accounts for the relationships among the items, and more importantly, it explains the structure of the measured construct. Either under- or over-factoring can largely impact the interpretation of the analysis results (Fava & Velicer, 1992, 1996). Under-factoring results in a loss of information and the inability to portray the true factor structure (Gorsuch, 1983). Over-factoring creates trivial factors that might seem important, but their over-interpretation threatens the proper understanding of the construct being measured and its factor structure (Dingman, Miller, & Eyman, 1964).

Many rules and indices have been proposed to determine the correct number of factors to retain when assessing dimensionality (see Hattie, 1985; Hayton, Allen, & Scarpello, 2004; Russell, 2001; Zwick & Velicer, 1986). Unfortunately, the various rules of thumb often lead to different solutions (Humphreys & Ilgen, 1969; Humphreys & Montanelli, 1974), and no one method has been found to be accurate under all conditions (e.g., De Ayala & Hertzog, 1991; Warne & Larsen, 2014). In practice, researchers must determine the number of factors on a case-by-case basis by applying one or more of these methods and decision rules.

Given that a large number of rules and indices have been utilized to inform the decision of the number of factors, the following paragraphs focus only on describing the two methods used in this dissertation. The Kaiser-Guttman (K-G) rule was chosen due to its popularity and simplicity, along with parallel analysis (PA) for its strong empirical support. Both of these methods make use of the eigenvalues from PCA. PCA-based approaches, or eigenvalue-based approaches, serve as pointers to the number of factors to retain (Liu, Zumbo, & Wu, 2012). Unlike other methods relying on the assessment of model fit, these approaches ignore this characteristic and consider only the magnitude of the variance accounted for by each component.

The eigenvalues-greater-than-one rule (Kaiser, 1960) is a prominent method (Thompson & Daniel, 1996). It is also known as the Kaiser-Guttman (K-G) rule, K1 rule, and Guttman rule. It will be referred to as the K-G rule in this dissertation. As one of the most widely used criteria to decide the number of factors (Thompson & Daniel, 1996; Warne & Larsen, 2014), it is the default option in some popular statistical software packages, such as SPSS. An eigenvalue is an

estimate of variance explained by a factor in a dataset (Ferguson & Cox, 1993), and an eigenvalue larger than one indicates it is greater than the average variance. The K-G rule has been found to be most effective with large sample sizes, fewer than 40 items, and item-to-factor ratios ranging from three to five (Gorsuch, 1983, 1997). The validity of using the K-G rule to determine the number of factors in a population matrix has been demonstrated in the literature (Cliff, 1988; Guttman, 1954), but, when applied to sample data, it has been reported to lead to over-extraction in many cases (e.g., Fabrigar, Wegener, MacCallum, & Strahan, 1999; Zwick & Velicer, 1982, 1986). Despite the poor performance of the K-G rule, it is still widely used in practice (Hoyle & Duvall, 2004).

Parallel analysis (PA) is a method based on the idea that a meaningful component underlying a dataset should possess a larger eigenvalue than the corresponding eigenvalue obtained through generated random variables (Horn, 1965). PA requires the generation of a series of random datasets with the same number of items and respondents (i.e., same size) as the original data matrix. The number of factors is indicated by the point where the eigenvalues for the real dataset drop below the average eigenvalues for the random datasets. The number of eigenvalues from the original data that are larger than the average eigenvalues from the random datasets is considered to be the number of factors to retain. Simulation studies show that PA does not depend on the distributional assumptions made on the data (e.g., normal or non-normal distributions; Dinno, 2009; Glorfeld, 1995). To improve the accuracy of PA, researchers recommend comparing the eigenvalues from the real dataset to those corresponding to the 95^{th} percentile of the eigenvalue distribution from the random datasets, rather than to the average eigenvalues (Cota, Longman, Holden, Fekken, & Xinaris, 1993; Glorfeld, 1995). The research literature strongly supports PA (Fabrigar et al., 1999; Thompson & Daniel, 1996) because it has been found to function well under various conditions (Humphreys & Montanelli, 1974; Zwick & Velicer, 1986), and to be more accurate than other decision rules (Dinno, 2009; Glorfeld, 1995; Liu et al., 2012; Velicer, Eaton, & Fava, 2000; Zwick & Velicer, 1986).

Besides using PCA-based procedures to help decide on the number of factors, researchers can also employ the model comparison approach through EFA. In the EFA model comparison approach, a series of models with an increasing number of factors is tested in sequence. It usually starts with a one-factor solution, and adds one factor at a time until the model fits the data adequately. Then researchers can decide on the number of factors. Chi-square test statistics based

on maximum likelihood (ML) estimation can be utilized to compare nested models. The literature notes that using Chi-square test statistics to determine the number of factors tends to result in over-extraction (Gorsuch, 1983; Hakstian, Rogers, & Cattell, 1982; Hayashi et al., 2007; Zwick & Velicer, 1986).

Rather than comparing all possible factor solutions through the EFA, it is common for researchers to use the model comparison approach in conjunction with PCA-based approaches to finalize their decision on test dimensionality. In this case, EFA is conducted as a follow-up once PCA-based rules and criteria have pointed to the approximate number of factors. Only a limited number of possible factor solutions surrounding the area of a "pointer" are tested via EFA. Usually, the ultimate goal of performing a factor analysis is to identify the underlying latent variables. To achieve this, interpretability, or the potential that the results of the factor analysis can be given clear meaning or labels, is important. Thus, when choosing the number of factors, researchers may regard not only model fit, but also the pattern and the interpretability of the factor structure observed in an EFA.

Moreover, factor rotation is often considered because without it, clusters of variables are unlikely to be identified by the initial factor extraction methods (Gorsuch, 1983). Rotation cannot improve the basic aspects of the EFA results, such as the model fit and the total amount of variance extracted from the items. Rather, it is used to improve the interpretability, reliability, and reproducibility of factors (Weiss, 1976). The goal of rotation is to simplify and clarify the data structure. Because the number of positions for the factor axes is unlimited, a unique solution to the rotation problem is not possible (Comrey, 1978). A simple structure (Thurstone, 1947), which has served as the principal criterion for rotation, is achieved by rotating factors until each is maximally collinear, with a distinct cluster of vectors (Rummel, 1970).

Researchers select rotation methods mainly based on the absence or presence of inter-factor correlations. Orthogonal rotation produces statistically uncorrelated factors, while oblique rotation allows them to be correlated. Orthogonal rotation is recommended because of its conceptual clarity, computational simplicity, and accessibility to be incorporated into the subsequent analysis (e.g., Nunnally, 1978). Varimax (Kaiser, 1958), Quartimax (Carroll, 1953), and Equamax are readily available orthogonal methods of rotation. Varimax rotation is a common choice (Fabrigar et al., 1999; Henson & Roberts, 2006; Kline, 1994). However, assuming factors are uncorrelated is often impractical because most psychological and

educational factors are correlated. Using the Varimax rotation therefore produces unrealistic or less useful factor structures.

Although oblique rotation adds statistical complexity, it more accurately represents the complex nature of the examined variables because constructs in the real world are rarely uncorrelated (Harman, 1976). With the development of computers and statistical software, the computational simplicity of orthogonal rotation methods became less compelling to researchers. More recent works have advised that oblique rotation methods should be used regardless of the assumptions about inter-factor correlation (Osborne, 2015; Schmitt, 2011). This is because oblique rotations allow a weak or zero correlation between factors, and their results are comparable to those from an orthogonal rotation when the inter-factor correlation is negligible (Schmitt, 2011).

There are a variety of oblique rotation methods. Some common options include Promax (Hendrickson & White, 1964), Quartimin, Direct Oblimin (Jennrich & Sampson, 1966), and Geomin (Yates, 1987). Promax rotation is conceptually simple and is available in many widely used statistical software packages, including *SPSS*, *Stata*, and *Mplus*. It has been recommended for large datasets because its computations can be performed quickly. It is a multi-step method, starting from a Varimax rotation and then relaxing the constraint of no inter-factor correlations to allow the factors to be correlated. To apply Promax rotation, a power parameter must be specified. This parameter must be greater than one, but should usually not exceed the value of four. The default power parameter is four in SPSS and three in Stata. The choice of this power parameter affects the factor solution (Browne, 2001). The higher it is set, the more likely the researcher is to obtain factor structures that are low in cross-loadings but high in inter-factor correlations. Geomin, which is the default rotation method in Mplus, tends to produce solutions that are easy to interpret, as it focuses on reducing cross-loading magnitudes. Simulation studies show that Geomin rotation is a promising method when the true factor loading structure is unknown (Asparouhov & Muthén, 2009). On the other hand, it has been reported to function poorly with complex factor pattern loading matrices (Asparouhov & Muthén, 2009). Also, since Geomin rotation uses an iterative algorithm, it is possible that multiple solutions are reached because it converges to a local minimum (Asparouhov & Muthén, 2009) or no solution is reached due to it failing to converge. In general, researchers recommend trying different oblique rotations to better describe the factor structure (Kline, 1994; Rummel, 1970; Sass & Schmitt,

2010). Rotation methods can significantly affect the magnitude of inter-factor correlations and cross-loadings (Sass & Schmitt, 2010). Researchers may also wish to consider the potential factor structure complexity when selecting a rotation method.

In summary, the K-G rule and PA are widely used as pointers to help researchers decide on the region of the possible number of factors. A researcher may start with these methods, and then rely on EFA to compare competing models with different numbers of factors, as pointed by the K-G rule or PA results. In the model selection process, the fit and interpretability of different factor solutions should be considered. Rotation is often applied because it helps clarify and simplify the EFA results. Recall that the goal of this subsection is not to provide a complete description of EFA but to prepare readers with basic background knowledge about the commonly used methods so that the methodology and the results of this dissertation can be easily understood. For a comprehensive introduction, see Brown (2006) and Thompson (2004).

Confirmatory methods to assess the factor structure of mixed-keyed tests

Researchers rarely collect and analyze data without an *a priori* idea of how the variables are related (Floyd & Widaman, 1995). When employed for validation purposes, CFA is often applied to test a theory or several competing theories of the construct. Researchers frequently draw on evidence from CFA to support the use of test-level scores (Zimprich et al., 2011), when they already have hypotheses about the structure of a test, including the number of factors, the relationships among items and factors, and the associations among factors. For example, when a total score of all the items is used to quantify the construct, it implies that this test follows a unidimensional structure. A one-factor model should be tested through CFA to investigate whether this unidimensional assumption is supported by the data. Additionally, CFA has been employed to investigate relationships among different variables or tests in validation studies (Thompson & Daniel, 1996). This subsection focuses on the use of CFA in the investigation of test dimensionality, and briefly describes three common estimation methods and some frequently reported fit indices.

As described in the review of empirical studies on negatively keyed items, many alternatives to the one-factor model have been proposed to describe the structure of a presumed unidimensional test with mixed-keyed items (see Figure 3 to Figure 6). Often, one or two of these alternatives are selected as competing models in addition to the one-factor model. In such

cases, researchers determine the final factor structure of a test by comparing the fit statistics obtained through CFA.

When conducting CFA, the maximum likelihood (ML) procedure is one of the most commonly used statistical methods for parameter estimation. The literature suggests that ML performs best with a sufficient sample size, proper model specification, and multivariate normality (Schmitt, 2011). The assumption of multivariate normality can be violated when the (observed) response data are collected through ordinal rating response scales with a small number of options. When the data distributions do not match the assumptions of the ML estimation method, it can result in biased estimates of parameters and standard errors (Beauducel & Herzberg, 2006; Flora & Curran, 2004; Rhemtulla, Brosseau-Liard, & Savalei, 2012).

Alternative estimation methods exist. Among them, robust continuous ML (MLR) estimation and weighted least squares means and variance adjusted (WLMVS) estimation are often recommended in the literature. MLR uses standard Pearson correlations, while weighted least squares (WLS) uses polychoric correlations. Both adjust the Chi-square test statistic and can reach accurate parameter estimates and findings depending on data conditions and model specifications (see Rhemtulla et al., 2012).

To evaluate and compare models, researchers review fit indices, parameter estimates, and sometimes, modification indices. Chi-square tests of model fit and other descriptive fit indices are used in factor analysis to assess the global fit of the model. Among them, Chi-square tests are the only ones based on distributional statistical theory (Hayashi et al., 2007). However, Chi-square tests have been criticized for being too sensitive to reject even trivial model misspecifications with large sample sizes (Hu & Bentler, 1998; Miles & Shevlin, 2007; Saris, Satorra, & van der Veld, 2009). Meanwhile, when the sample size is small, models with substantial misspecifications may not be rejected (Saris et al., 2009). Despite the criticism they have received, the results of Chi-square tests are still routinely reported in studies using CFA. Moreover, the statistics from Chi-square tests are the basis for most other fit indices.

Besides Chi-square tests, researchers have proposed a large number of descriptive fit indices (see Hu & Bentler, 1998). Some researchers have distinguished and categorized them into a few types, such as absolute fit indices, relative fit indices, and parsimony fit indices (e.g., Hu & Bentler, 1999; Kline, 2011). The Chi-square test, the root mean square error of approximation (RMSEA), the goodness-of-fit index (GFI), Akaike's information criterion (AIC),

and the Bayesian information criterion (BIC) are some examples of absolute fit indices. Absolute fit indices are referred to as such because they are not obtained through model comparison but are derived from the fit of the observed and model-implied covariance matrices (Jöreskog & Sörbom, 1993; McDonald & Ho, 2002). By contrast, relative fit indices (McDonald & Ho, 2002), also known as comparative fit indices or incremental fit indices (e.g., Miles & Shevlin, 2007), are based on a comparison between the tested model and a null model. The null model should always have a poor fit (i.e., large Chi-square), as it is often specified as a model with all variables (i.e., items) uncorrelated. In other words, the null model often assumes that no common factor exists underlying the variables (McDonald & Ho, 2002). Relative fit indices include the comparative fit index (CFI; Bentler, 1990), the Tucker-Lewis index (TLI), the Bentler-Bonett normed fit index (NFI; Bentler & Bonett, 1980), and the Bollen's incremental fit index (IFI). Parsimonious fit indices such as the parsimony goodness-of-fit index (PGFI), and the parsimony normed fit index (PNFI; Mulaik et al., 1989), are relative fit indices adjusted to penalize more complex models over the simpler or more parsimonious models. This group of fit indices appears to be less frequently reported in the literature.

Bentler (2007) recommends limiting the number of fit indices reported. The RMSEA, CFI, and TLI (also called the non-normed fit index, or NNFI) are among the most popular ones. All three carry some penalty for model complexity and range from zero to one. The CFI and TLI are highly correlated. For these two fit indices, a value close to one indicates a good model fit; for RMSEA, this value is close to zero. There is no absolute cut-off when using these indices to judge model fit. However, general guidelines suggest that an RMSEA smaller than 0.05 indicates a good model fit, and an RMSEA smaller than 0.08 corresponds to an acceptable fit (MacCallum, Browne, & Sugawara, 1996). CFI values of 0.95 for continuous outcomes (Hu & Bentler, 1999) and 0.96 for ordinal categorical outcomes (Yu, 2002) indicate an acceptable model fit. TLI values of 0.95 or higher suggest that the model fit is adequate (Hu & Bentler, 1999). On the whole, it is recommended that multiple criteria, as well as theoretical reasoning, should be used in deciding which factor model to choose (Thompson & Daniel, 1996).

Evaluating fit indices may lead researchers in one of two directions. When the indices indicate a good fit, they are likely to further examine the model by looking into parameter estimates. Such parameters usually include factor loadings, inter-factor correlations, and error variances. When the indices show a poor model fit, researchers sometimes modify their

hypotheses so that the measurement model is more consistent with the structure of the actual data collected. Fit indices assess the overall model fit, while modification indices can reveal where the deviation occurs by identifying potential modifications to the hypothesized measurement model. A modification index is the expected amount the Chi-square will drop if a parameter is estimated as part of the model rather than being fixed to zero. It therefore represents the potential benefit of revising or freeing the relevant parameter. Usually, a researcher who turns to modification indices in revising the *a priori* hypothesized model wish them to be large enough to be considered meaningful. The modification index value of 3.84, which corresponds to the Chi-square value that should be exceeded at the 0.05 level for one degree of freedom, is a possible cut-off point. Although modification indices provide useful information about how the model-data fit can potentially be improved, it is dangerous for researchers to rely solely on them when revising models. Allowing model modification moves a CFA from its confirmatory mode to an exploratory mode (Flora & Flake, 2017). Post-hoc modifications that are not based on theory can lead to models that fit by chance or to models that are unstable across different respondent samples (MacCallum, Roznowski, & Necowitz, 1992).

This subsection is intended to be a brief, non-technical review of some key aspects in the use of CFA in assessing dimensionality. For a more comprehensive discussion, please refer to Brown (2006), Hoyle (1995), and Thompson (2004).

Gaps in the Literature and the Purpose of This Study

To understand the construct being investigated and to support the scoring of a measurement instrument, it is a common validation practice to assess the dimensionality and factor structure of a test. However, the mixed-keyed tests pose challenges to the assessment and interpretation of test dimensionality. Data collected from mixed-keyed tests that are designed to be unidimensional often turn out to support a more complex structure rather than the one-factor model.

Despite numerous studies on “negative” items, it is not always clear if these researchers examined negatively keyed items, negatively worded items, or items with negative social-psychological value (e.g., socially unacceptable or undesired attitudes and behaviours). This makes it difficult to compare the various findings. Studies properly distinguishing among keying, wording and the social-psychological meaning of an item are necessary to support a better

understanding of problems related to "negative" items. This dissertation defines and separates the terms used to describe item keying and item wording directions.

Additionally, this dissertation attempts to disentangle the effect of item keying and wording on the psychometric properties of a test, and in particular, analyzes the effect of keying on the assessment of test dimensionality. Admittedly, most items are likely to fall into two categories, positively worded and positively keyed items, and negatively worded and negatively keyed items. In other words, the keying and wording direction are likely to be consistent. Therefore, it is difficult to clearly separate the keying effect from the wording effect in studies using data collected through existing measurement instruments. The vague terminology and the close connection between item keying and wording lead to a situation where the issues regarding factor structure have been observed, but their causes remain unknown. It is possible that the problems concerning the factor structure found in mixed-keyed tests are due to the differences in the cognitive and linguistic processing demands associated with item wording. It is also possible that the strategies utilized to score negatively keyed items do not reflect the keying differences properly or that the methods employed to assess the factor structure of mixed-keyed tests do not handle opposing keying directions well. Indeed, all of these possibilities may be at play.

The inconsistent findings from empirical studies suggest that, among all mixed-keyed tests, only those with certain characteristics may tend to have poor psychometric properties. It is possible that only tests with few scale points on the response scale and a large proportion of negatively keyed items will suffer from poor psychometric properties. Without a systematic evaluation of mixed-keyed tests, their psychometric properties, and the proper methods for obtaining these statistics, we can obtain only a partial understanding of the nature of the negative keying effect. For example, tests usually have more positively keyed items than negatively keyed ones, but the proportion of the negatively keyed items can differ. Take the RSE scale (Rosenberg, 1965) and the Penn State Worry Questionnaire (Meyer et al., 1990) as examples. The ten-item RSE scale (Rosenberg, 1965) contains five negatively keyed items, or 50% of the total. Meanwhile, among the sixteen items in the Penn State Worry Questionnaire (Meyer et al., 1990), five are negatively keyed, or about 31% of the total items. However, in most of the studies on the item keying effect, only tests with all positively keyed items, all negatively keyed items, or an equal number of positively and negatively keyed items have been examined. These types of tests are special cases where 0%, 100% or 50% of the items are keyed negatively. Extending the

investigation of the keying effect to a greater variety of conditions will help elucidate whether and how the proportion of the negatively keyed items in a test may influence the test factor structure.

To take this one step further, this dissertation aims to provide some suggestions to researchers on handling negatively keyed items. A major gap in the literature is the lack of documentation on how data collected from mixed-keyed tests should be analyzed. The effect of item keying has been insufficiently studied, even though how to handle such data is a basic decision researchers must make in their day-to-day practice. To address this issue, this dissertation focuses on assessing the dimensionality of tests with negatively keyed items. The performance of common methods used in validation practice to assess dimensionality and factor structure will be investigated and, based on the results, a set of guidelines will be offered.

In summary, the inclusion of both positively and negatively keyed items in one test is a common practice. Although it has been assumed that differently keyed items function in the same way, empirical studies call this assumption into question, as unidimensional tests with mixed-keyed items often result in a factor structure influenced by item keying direction. Given the lack of systematic studies on the effect of negatively keyed items, it is unclear whether the emergence of an unexpected factor structure should be attributed to item wording, item keying, respondent characteristics, data analytic methods, or the combination of these factors. It is also unknown to what extent and under what conditions the number of factors will be inflated in the presence of negatively keyed items.

Research Questions and Study Overview

To address the observed gap in the research literature, I investigate how including negatively keyed items in short unidimensional Likert-type tests affects the statistical conclusions on their dimensionality and factor structure in this dissertation. Primarily, I seek to document the conditions under which the presence of negatively keyed items will and will not lead to the correct judgment of test dimensionality and factor structure as suggested by different statistical methods. Based on the findings, I then propose recommendations for applied researchers regarding how to deal with data obtained from tests with negatively keyed items.

The overall research questions guiding this study are as follows:

Q1: Do the different psychometric models (i.e., simulation models) of negatively keyed items affect the statistical judgment of test dimensionality under different conditions?

Q2: Does the reverse scoring of negatively keyed items affect the statistical judgment of test dimensionality?

Q3: Under which conditions will the K-G rule or PA identify the correct number of factors?

Q4: When EFA is conducted with an inflated number of factors, what will the factor structure look like? Will factors emerge according to the keying direction of items?

Q5: Under which conditions will the one-factor model be rejected either by the Chi-square test or other fit indices?

Q6: When the one-factor model is not supported by fit statistics in CFA, what are the consequences of revising the model using modification indices?

To answer these research questions, I conducted a set of simulation studies to investigate the performance of different rules and indices in judging the dimensionality and factor structure of mixed-keyed tests using factor analytical methods. A flow chart is presented below showing the main factors considered (see Figure 9).

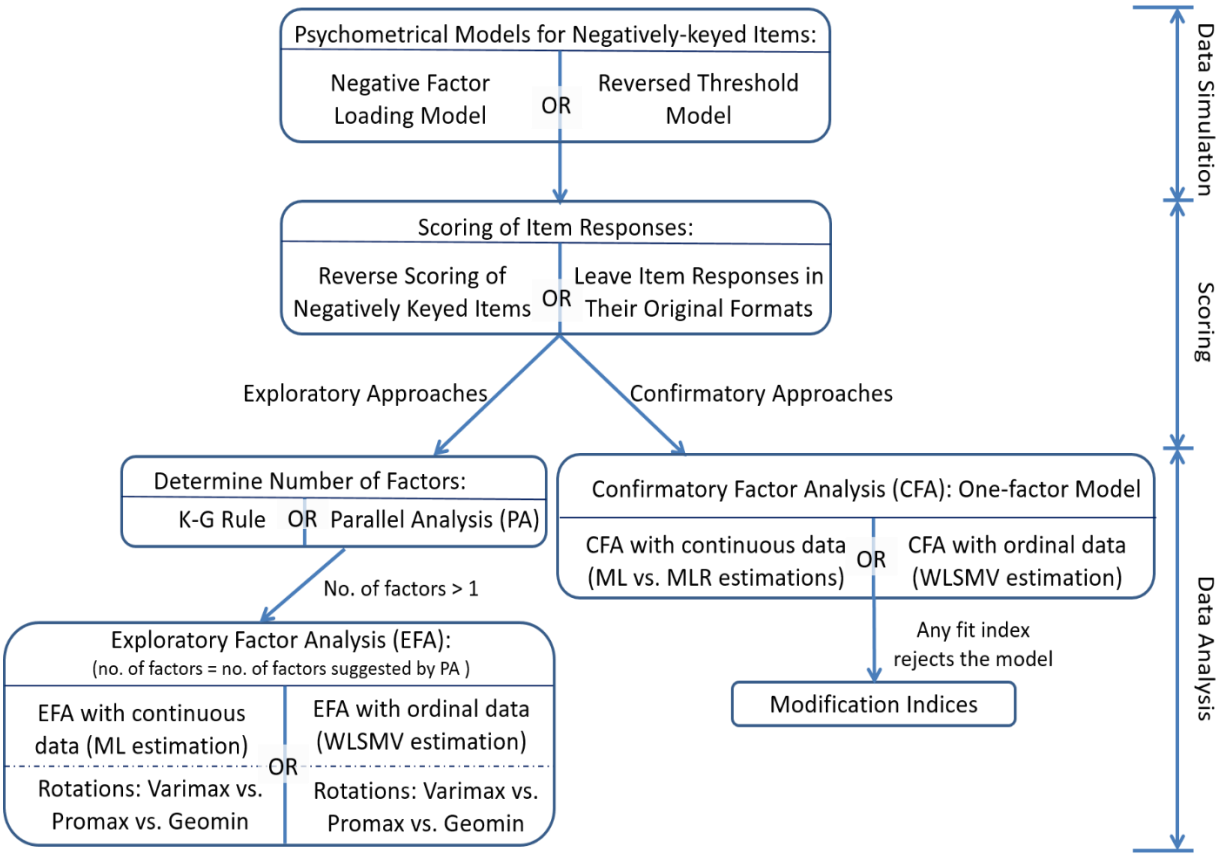


Figure 9.

Factors manipulated at different stages of the simulation research

As Figure 9 shows, each of the four simulation studies is carried out in three stages: data simulation, item response scoring, and data analysis. The variables manipulated in the data simulation stage are: (a) the psychometric model of negatively keyed items (i.e., negative factor loading model or reversed threshold model), (b) the number or proportion of negatively keyed items, (c) the magnitude of communality of the items, and (d) the distribution of the observed item responses. In the data analysis stage, the factors that are considered include (a) scoring method for the negatively keyed items, and (b) statistical methods for assessing test dimensionality. The methods that investigate test dimensionality fall into two broad categories: (a) exploratory approaches, and (b) confirmatory approaches. More specifically, when exploratory approaches are used, the number of factors to retain is first judged via two PCA-based methods, the K-G rule and PA. When PA indicates that more than one factor should be kept, EFA with rotations will be conducted as a follow-up. Two common estimation methods,

ML and WLSMV, will be explored within the EFA framework. In conjunction with these two estimation methods, three rotation methods, Varimax, Promax, and Geomin, are considered when EFA is applied to the dataset. When confirmatory approaches are used to assess test dimensionality, three different estimators are considered, ML, MLR, and WLSMV. The overall model fit is judged by four fit indices including Chi-square test statistics, CFI, TLI and RMSEA. The study design and procedures will be described in detail for each study in Chapter Three.

CHAPTER THREE:
SIMULATION STUDIES

The purpose of this chapter is to present the four simulation studies that investigate the impact of negatively keyed items on the assessment of test dimensionality for validation purposes. As described in Chapter Two, including both positively and negatively keyed items in one test is a common practice for short psychological tests that are purported (designed) to measure one latent variable. It has been assumed that items keyed in different directions function in the same way. However, as Chapter Two clarifies, empirical studies call this assumption into question, as unidimensional tests with mixed-keyed items often have a factor structure that is defined by item keying direction. It is unclear from the empirical studies whether the emergence of this unexpected factor structure can be attributed to item wording, item keying, respondent characteristics, data analytic methods, or a combination of these variables. Among these factors, the item keying effect is one of the most commonly ignored. Although simulation studies are better suited to investigating the impact from different sources separately, they have rarely been done to explore the item keying effect. It is therefore unknown to what extent and under what conditions the number of factors will be inflated in the presence of negatively keyed items. To answer this question, I conducted four interrelated simulation studies. I simulated population data on a twelve-item test using factor models with varying numbers of negatively keyed items.

To present the findings clearly, the four simulation studies are organized into two sections according to the methods used to generate the responses to negatively keyed items. Studies 1 and 2, which are contained in the first section, involve simulations of the item responses from a one-factor model in which negatively keyed items have negative factor loadings in the latent response model. The simulation procedures used for these two studies assume that responses to a negatively keyed item follow a latent response that is negatively correlated with the construct of interest. This psychometric model was described in Chapter Two as the "negative factor loading model for negatively keyed items." In this model, an item is negatively keyed because it measures the polar opposite of the construct. For example, in a job satisfaction measure, an item like "My job is dull" can be seen as measuring the unsatisfying aspects of the work. In this regard, the latent response underlying the item, "My job is dull," has a negative correlation with the job satisfaction factor (i.e., negative factor loading in the latent

response model). In this example, dissatisfaction and satisfaction are viewed as the two ends of one continuum, and a test like this can be referred to as a bipolar measure. To some extent, a mixed-keyed test is a bipolar measure in which the negatively keyed items are measuring the polar opposite of the construct. The details of the data simulation procedures will be described in the method section of Study 1.

The last two studies (i.e., studies 3 and 4) are presented in the second section. In them, responses to the negatively keyed items are generated during the transformation from latent response y^* to observed response y. Chapter Two designates this model as the "reversed threshold model for negatively keyed items." The relationships between latent response y^* and observed response y are reversed for items keyed negatively compared with their positively keyed counterparts. It is assumed that the latent responses underlying both positively and negatively keyed items are the same in their relationships with the construct. Take two items from a job satisfaction test as an example. A negatively keyed item, "I don't enjoy going to work," and a positively keyed item, "I enjoy my job," can be seen as both assessing job satisfaction and having the same latent response distributions. However, the relationship between the latent response distribution and the observed response category is reversed for the negatively keyed item because it is a negation of the positively keyed item. A higher level of job satisfaction on the latent response distribution corresponds to a higher level of agreement on positively keyed items, but a lower level of agreement on negatively keyed ones. The details of the data simulation procedures will be given in the method section of Study 3.

This research focuses on the correctness of the decision of the number of factors. Studies 1 and 3 evaluate how many factors to retain through exploratory approaches, while Studies 2 and 4 assess the model fit via CFA. Given that a purpose of this dissertation is to inform day-to-day applications, the methods used to decide on the number of factors were chosen based on the common research practice. Throughout, data simulation and item response scoring were conducted with *SPSS 21.0* (IBM Corp, 2012), eigenvalues based on polychoric correlations were obtained through "psych" (Revelle, 2014) and "nFactors" (Raiche, & Magis, 2010) packages in *R*, and all the other analyses were conducted via *Mplus 7* (Muthén & Muthén, 1998-2012).

Section One: A Negative Factor Loading Model for Negatively Keyed Items

The two simulation studies in this section document how statistical decisions are made about the number of factors under different conditions, which vary in the number of negatively keyed items, item communality levels, observed response distributions, and the methods and rules used to judge test dimensionality. The simulation model generated item responses to negatively keyed items through negative factor loadings in the latent response model—i.e., the item response thresholds are all the same, but the factor loadings are negative for the negatively keyed items. Chapter Two described this psychometric model as a "negative factor loading model for negatively keyed items."

Study 1: The impact of negatively keyed items on the decision of the number of factors using exploratory approaches

This study focuses on choosing the number of factors to retain through exploratory approaches. The research questions it aims to address are as follows:

Q1.1: Does the reverse scoring of negatively keyed items affect the number of factors identified by the Kaiser-Guttman (K-G; i.e., eigenvalue-greater-than-one) rule and parallel analysis (PA)?

Q1.2: Under what conditions will the K-G rule correctly point to the number of factors?

Q1.3: Under what conditions will PA correctly point to the number of factors?

Q1.4: When more than one factor is suggested for retention, what will the factor structure look like? Will a second factor be formed by negatively keyed items in EFA?

Method

Study design

This study simulated negatively keyed items by specifying negative item loadings in the data generation factor model. The factors manipulated in the data simulation stage are presented in Table 1. As shown in the table, three factors are systematically manipulated and fully crossed during the data simulation process. These factors are: (a) the number or proportion of negatively keyed items, (b) the magnitude of item communality, and (c) the distribution of observed item responses. There are four levels of the number of negatively keyed items (0, 2, 4, 6 items out of 12 items), three levels of communality (0.06, 0.25, and 0.56), and two observed item response

distributions (symmetric and asymmetric distributions; see Table 1). The number of negatively keyed items is manipulated to represent tests with various proportions of negatively keyed items. The levels of communality are manipulated through changing the magnitude of factor loadings. The three levels of communality are chosen to represent a wide range of factor loadings that are commonly seen in psycho-educational tests. These three levels of communality correspond to factor loadings of 0.25, 0.50, and 0.75. Two levels of observed item response distributions are chosen so that both symmetric and skewed conditions are included in this simulation study. Many variables measured in psycho-educational and health research are expected to follow bell-shaped distributions which are unimodal and symmetric. It has also been noted that strongly skewed distributions are often encountered in psychology research (Aron, Coups, & Aron, 2013). Together, these factors manipulated in the data simulation stage cover a relatively wide range of conditions that represent a variety of unidimensional tests.

The length of the test (i.e., the total number of items) is fixed at 12, and the observed responses are on a scale of one to five. The test length is chosen to represent a typical short, unidimensional psychological test. A systematic review of publications in six psycho-educational and health journals for the time period of 1999 to 2004 concludes that the median and average lengths of unidimensional psychological tests are 11 and 18 (Slocum, 2005). The current simulation study selects the test length to be 12 because this number is close to the median test length and it allows for conditions with an equal number of positively and negatively keyed items. Also, the five-point rating scale is chosen because it is a common Likert-type response format (Slocum, 2005). It is also in the grey zone where researchers are often unclear about whether the item responses from such scales should be treated as continuous or ordinal data (Rhemtulla et al., 2012), and thus needs more studies. In addition, the test is fixed to follow a one-factor model. This results in a total of 24 (i.e., $4 \times 3 \times 2$) simulated datasets.

Table 1.

Factors manipulated in the data simulation

Factors Manipulated	Number of Levels	Specification for Each Level
Number of negatively keyed items	4	0 (0.00%) 2 (16.67%) 4 (33.33%) 6 (50.00%)
Communality	3	Communality = 0.06 (factor loadings = 0.25) Communality = 0.25 (factor loadings = 0.50) Communality = 0.56 (factor loadings = 0.75)
Skewness of the observed response distribution	2	Symmetric; skewness = 0 Asymmetric; skewness = -2

Besides the factors that are manipulated in the data simulation process, the scoring method for negatively keyed items is also manipulated after original item responses are simulated. Two scoring methods are investigated: (a) reversely scoring negatively keyed items, and (b) leaving all the responses in their original values. Each scoring method is applied to the negatively keyed items once the response data are simulated. Note that there are six conditions that do not contain any negatively keyed items, and these methods do not apply to them. Utilizing the two scoring methods for only the conditions having negatively keyed items results in a total of 42 (i.e., 4×3×2×2-6) unique datasets.

The primary outcome variable of interest in this study is the decision of the number of factors, which is determined via the K-G rule and parallel analysis (PA). When PA is used, the criterion used to determine the number of factors is that the eigenvalue associated with a factor extracted from the correlation matrix is larger than its expected value at the 95^{th} percentile eigenvalues obtained from random uncorrelated data. The expected eigenvalues are acquired by simulating normal random samples that parallel the observed data in their sample size and the number of variables. The number of factors suggested by the K-G rule and PA will be reported separately. In daily practice, researchers rely extensively on K-G and PA methods when deciding on the region of the possible number of factors. A researcher may start with these methods, and

then use EFA to compare competing models with different numbers of factors, as indicated by the K-G or PA results. When EFA is conducted, the decision of the number of factors and the factor structure is usually based on the fit and interpretability of different EFA models. Following this routine, when the PA results suggest that more than one factor should be retained, a follow-up EFA is performed to explore the pattern of the factor loadings.

To investigate factor structure using EFA, two factors are considered and varied. Firstly, the observed item responses can be treated as either continuous or ordinal. Accordingly, different estimation methods are employed in the extraction procedure. Maximum likelihood (ML) estimation is used for EFA when the observed item responses are treated as continuous, while weighted least squares means and variance adjusted (WLSMV) estimation is used when they are treated as ordinal. Also, three types of rotation are applied to these datasets: (a) Varimax, (b) Promax, and (c) Geomin. Varimax is an orthogonal rotation method that assumes no correlations between factors. Promax is an oblique rotation method that allows correlations between factors. In addition to these two traditional methods, another oblique rotation method, Geomin, is also included. A fully-crossed factorial design (2 estimators × 3 rotation methods) leads to a total of six results for each dataset. Although factor rotation methods are applied, the focus here is not on the substantial interpretation of the factor solutions but on the general factor loading patterns. The primary reason for conducting the EFA is to examine whether the emergence of factors is associated with item keying direction.

Procedures

A flow chart summarizing the major steps in conducting this study is presented in Figure 10. As shown in the chart, the study begins with data simulation and then applies different scoring methods to the simulated item responses. After scoring, the datasets are ready to be analyzed. In the data analysis stage, each dataset (i.e., each combination of factors manipulated before data analysis stage) is examined using PCA to obtain eigenvalues. The number of factors, as pointed by the K-G rule and PA, is reported for each condition. In cases where PA suggests retaining more than one factor, EFA with three types of rotation is applied to explore the factor structures. The details of each step are described in the following paragraphs.

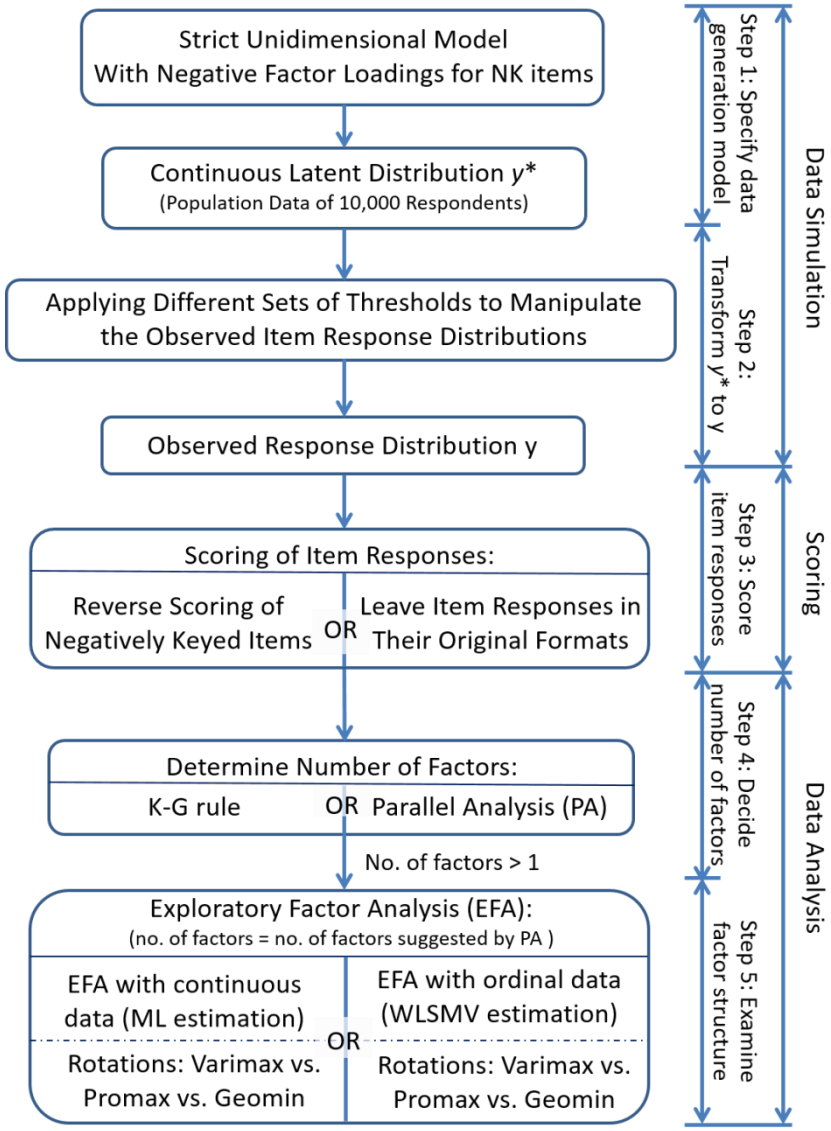


Figure 10.

A flow chart representing the research process of Study 1

Step 1: Specify the data generation model. The simulations are based on the premise that the observed discrete responses (*y*) are a manifestation of an unobserved underlying continuous distribution (*y**). In other words, they assume that each item is designed to measure a theoretically continuous construct, and that the observed responses are discretized realizations of the continuous *y**. The unobserved univariate continuous distribution that generates an observed ordinal distribution is referred to as a latent response distribution *y** (Muthén, 1983, 1984).

This study uses a one-factor model with 12 items to generate the item response data (see Figure 11). This model is chosen because it typifies the CFA model specifications that are commonly encountered in practice. The population data generated from this simulation model reflect a situation where strict unidimensionality is true at the population level. All the *y** values follow normal distributions and are standardized to have a mean of zero and a standard deviation (SD) of one. The continuous *y** distributions follow multivariate normal distributions.

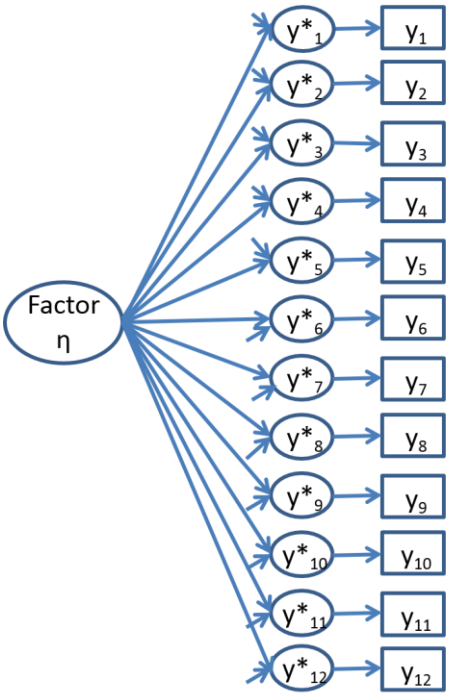


Figure 11.
Specifications for the one-factor model

Step 2: Transform continuous latent response distributions to observed responses. Each of the observed response variables (i.e., *y*) is determined by its own latent response variable (*y**; see Figure 11). The latent response (*y**) underlying each item is continuous, but the observed item responses can be broken down into *K* categories to represent responses obtained through Likert-type response scales. The observed responses can be transformed into either symmetric or skewed distributions by applying different sets of thresholds.

To transform the latent response (*y**) into *K* ordered response categories, a set of (*K*-1) thresholds is needed. In this study, the observed item responses are simulated on a one-to-five

point response scale, which is a representation of a five-point rating response scale. Hence, four thresholds are needed to transform the latent responses (y^*) into the observed responses (y). The observed response distribution is manipulated by using different sets of thresholds. This study investigates two observed response distribution conditions: (a) symmetric, and (b) negatively skewed with a skewness of -2. The correspondence between the latent response (y^*) and observed response categories (y) for each distribution condition is illustrated in Table 2.

Table 2.

Thresholds used in the response transformation

Response Categories	Corresponding y^* Values	
	Symmetric	Skewed (skewness = -2)
1	Lowest thru -1.8000	Lowest thru -1.66429
2	-1.7999 thru -0.6000	-1.66428 thru -1.27956
3	-0.5999 thru 0.6000	-1.27955 thru -1.02406
4	0.6001 thru 1.8000	-1.02405 thru -0.68564
5	Higher than 1.8000	Higher than -0.68564

Step 3: Score the item responses. After the data are simulated, two methods are applied to score the observed responses to negatively keyed items. These methods are (a) reversely scoring negatively keyed items, and (b) leaving the responses in their original format. In both conditions, the responses to the positively keyed items remain as they are.

Step 4: Decide on the number of factors. To determine the number of factors for each dataset, all the datasets undergo principal components analysis (PCA) to obtain eigenvalues. Following the traditional method, PCA is conducted on Pearson correlation matrices, that is, the item responses are treated as if they are on a continuous scale. For each dataset, the number of factors, as identified by the K-G rule and PA, are reported. Finally, a global assessment of the correctness or incorrectness of the number of factors is made.

Step 5: Examine the factor structures. When PA identifies more than one factor, EFA is conducted using the suggested number. As described in the study design, EFA is performed with different estimators and rotation methods. Firstly, when EFA is carried out, item responses can be treated either as continuous or as ordinal variables. When they are treated continuously, ML estimation is used; when they are treated as ordinal variables, WLSMV estimation is used.

Estimated communality levels from the EFA models are presented, along with fit statistics for each. After evaluating the global fit of the EFA models, both orthogonal (i.e., Varimax) and oblique (i.e., Promax and Geomin) rotations are applied to the data to examine the pattern of factor loadings. Factor loadings and factor correlations are reported. When the EFA results do not support the number of factors suggested by PA, possible explanations for this inconsistency are explored.

Checking the simulation method: Descriptive statistics from one of the simulated datasets

As Chapter Two explains, negatively keyed items can be operationalized as ones that, without being reverse scored, are negatively correlated with the test score and other positively keyed items. To serve as a check on the simulation methodology, the descriptive statistics of each dataset are reviewed before conducting any further analyses. Before jumping to the results and conclusions, it may be worthwhile to present some basic statistics that summarize a simulated dataset.

The descriptive statistics, including the mean, skewness, and item correlations, of one simulated dataset are presented. In this dataset, the number of negatively keyed items is six, the communality level is high (0.56), and the observed item response distribution is symmetric. Table 3 presents the mean and skewness of the observed item responses. The first two columns on the left show the item ID and the keying direction. The statistics presented in the table are based on the originally simulated item responses. In other words, responses to the negatively keyed items are not reverse scored. The mean scores of all the item responses are close to 3, which falls in the middle of a five-point rating scale. Also, the skewness values are close to zero for all the items, indicating that the distributions are symmetric.

Table 3.

Mean and skewness of observed item responses (loading = 0.75, symmetric, $N_{NK} = 6$)

Item Id	Keying Direction	Mean	Skewness
I1	PK	3.01	0.00
I2	PK	3.00	0.00
I3	PK	3.01	0.00
I4	PK	3.00	0.01
I5	PK	3.00	0.00
I6	PK	3.01	0.02
I7	NK	3.00	0.00
I8	NK	3.00	-0.01
I9	NK	3.00	-0.04
I10	NK	2.99	-0.03
I11	NK	3.00	-0.01
I12	NK	3.01	0.00

Note: PK denotes positively keyed items and NK denotes negatively keyed items.

Table 4 presents the inter-item correlations. The correlations between positively and negatively keyed items are negative, while all the other correlations are positive. This pattern is consistent with the operationalization of negatively keyed items.

Table 4.

Correlation matrix of observed item responses (loading = 0.75, symmetric, $N_{NK} = 6$)

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11
I1	--										
I2	0.507	--									
I3	0.497	0.506	--								
I4	0.503	0.501	0.505	--							
I5	0.504	0.518	0.511	0.502	--						
I6	0.499	0.502	0.506	0.499	0.512	--					
I7	-0.498	-0.505	-0.500	-0.498	-0.505	-0.494	--				
I8	-0.506	-0.508	-0.510	-0.507	-0.509	-0.509	0.505	--			
I9	-0.507	-0.498	-0.515	-0.502	-0.509	-0.507	0.495	0.507	--		
I10	-0.504	-0.506	-0.514	-0.496	-0.512	-0.504	0.499	0.501	0.500	--	
I11	-0.495	-0.496	-0.495	-0.486	-0.501	-0.507	0.504	0.505	0.489	0.494	--
I12	-0.503	-0.494	-0.505	-0.495	-0.497	-0.503	0.497	0.487	0.500	0.498	0.512

The above tables show that the observed item response distribution and item keying directions have been manipulated according to the simulation design. The similarity between the general pattern of the simulated dataset and the datasets we may see in applied research shows the credibility of the simulation method.

Results and conclusions

Decision on the number of factors

The K-G rule and PA are PCA based procedures. Following the conventional methods to determine the number of factors, PCA is conducted on a Pearson correlation matrix to extract eigenvalues. As expected, both scoring methods for negatively keyed items produced identical eigenvalues. Therefore, the number of factors pointed either by the K-G rule or PA is not affected by the scoring method. Table 5 presents the number of factors suggested either by the K-G rule or PA for (a) different communality levels, and (b) different numbers of negatively keyed items. Both the K-G rule and PA always correctly identify the number of factors as one when the observed item response distributions are symmetric. In other words, none of the other factors (i.e., item communality level, the number of negatively keyed items, and scoring methods for negatively keyed items) affects the identification of the number of factors for a symmetric observed item response distribution. Hence, only conditions under which the observed item distribution is skewed are presented in Table 5. The conditions in which no item is negatively keyed (i.e., all items are positively keyed) serve as baselines for comparison with the conditions where different numbers of negatively keyed items are manipulated.

The far left column of Table 5 lists the number of negatively keyed items, and next to it are the factor loadings specified in the simulation model. The remaining two columns present the number of factors pointed by the K-G rule and PA, respectively. When the observed item response distribution is asymmetric, the number of factors that the K-G rule suggests in most of the conditions is wrong. The K-G rule points to the right number of factors only when the item communality level is high and none of the items is negatively keyed. It is evident that, in the presence of an asymmetric observed item response distribution and negatively keyed items, the K-G rule inflates the number of factors. As for PA, if no negatively keyed items exist, as shown in the first three rows (the rows with the first column labeled “0” under “No. of NK”), PA always point to one factor. In the presence of asymmetric observed response distributions, the number of

factors increases depending on the level of item communality and the number of negatively keyed items. The general pattern shows that with a higher average item communality level and a larger number of negatively keyed items, the decision of the number of factors based on PA is more likely to be inflated.

Taken together, there is, in essence, a four-way interaction among the studied conditions—i.e., observed response distribution, item communality level, number of negatively keyed items, and decision rules (K-G rule or PA). That is, the results suggest that the decision of the number of factors is always correct when the item response distribution is symmetric, regardless of any other factors manipulated in this study. In the presence of skewed item response distributions, the decision on the number of factors depends on the number of negatively keyed items, item communality levels, and the decision methods used. When the observed item response distribution is asymmetric, the number of factors pointed by the K-G rule is inflated in all conditions except one. Compared to the K-G rule, PA is more robust when dealing with negatively keyed items and skewed item response distributions.

Table 5.

Number of factors based on a Pearson correlation matrix

Number of NK items	Item Loadings	Observed Item Response Distribution Asymmetric (Skewness = -2)	
		The K-G rule based on Pearson correlations	PA results based on Pearson correlations
0 (Baseline)	0.25	3	1
	0.50	3	1
	0.75	1	1
2	0.25	3	1
	0.50	3	1
	0.75	2	2
4	0.25	3	1
	0.50	2	1
	0.75	2	2
6	0.25	3	1
	0.50	2	2
	0.75	2	2

Note: NK denotes negatively keyed items.

Exploring factor structures using EFA when more than one factor is identified

Exploring the factor structure of a test is a step that usually follows the identification of the number of factors. Sometimes researchers may also take an iterative approach to inform their decision using the model fit and the pattern of factor structure observed in EFA. Following this practice, EFA is used to explore the factor structure when the number of factors suggested by PA is greater than one. The number of factors to extract in EFA is set based on the results obtained from PA. This is because the literature has criticized the K-G rule for its frequent over-factorization (Hakstian et al., 1982; Zwick & Velicer, 1982, 1986). This inflation has also been observed in the current study (see Table 5). Moreover, this inflation also appears under some of the baseline conditions where there are no negatively keyed items. The primary focus of the following EFA is to examine if items will load on different factors due to their keying direction. The inflation of the number of factors under baseline conditions makes it difficult to interpret the EFA results and the comparisons between these and other conditions.

In total, PA points to an incorrect number of factors under four conditions (see Table 5), where the K-G rule also points to the wrong number of factors. These four conditions are: (a) item loadings equal 0.75, two items are negatively keyed, and the observed item responses are asymmetric (i.e., L0.75_NK2_Asymm); (b) item loadings equal 0.75, four items are negatively keyed, and the observed item responses are asymmetric (i.e., L0.75_NK4_Asymm); (c) item loadings equal 0.75, six items are negatively keyed, and the observed item responses are asymmetric (i.e., L0.75_NK6_Asymm); and (d) item loadings equal 0.50, six items are negatively keyed, and the observed item responses are asymmetric (i.e., L0.50_NK4_Asymm). The numbers of factors identified by PA in all these four conditions are two.

To begin reporting the EFA results, the item communality levels are estimated. Since rotation does not change these item communality levels, for each condition, four sets of item communalities (2 estimators × 2 scoring methods) are estimated. The results show that the communality levels are identical regardless of whether the negatively keyed items are reverse scored or left in their original values. Hence, when reporting these results in Table 6, the scoring methods applied to negatively keyed items are not included. For each condition under investigation, two sets of communality estimates are listed, one obtained with statistics from the EFA model estimated by treating the observed ordinal item responses as continuous with the ML

estimator, and the other using the WLSMV estimator. The communality estimates are obtained as 1-(residual variance).

Table 6.

Estimated communality levels from two-factor EFA models

	L0.75_NK2_Asymm		L0.75_NK4_Asymm		L0.75_NK6_Asymm		L0.50_NK6_Asymm	
	Cont. (ML)	Ordinal (WLSMV)	Cont. (ML)	Ordinal (WLSMV)	Cont. (ML)	Ordinal (WLSMV)	Cont. (ML)	Ordinal (WLSMV)
I1	0.43	0.58	0.43	0.58	0.40	0.55	0.18	0.27
I2	0.41	0.57	0.41	0.57	0.42	0.57	0.15	0.26
I3	0.43	0.58	0.43	0.58	0.43	0.58	0.17	0.26
I4	0.41	0.57	0.41	0.56	0.40	0.55	0.16	0.24
I5	0.42	0.58	0.42	0.58	0.43	0.58	0.17	0.25
I6	0.39	0.55	0.39	0.55	0.40	0.56	0.13	0.23
I7	0.43	0.58	0.43	0.58	**0.41**	**0.56**	**0.17**	**0.25**
I8	0.43	0.59	0.43	0.59	**0.39**	**0.57**	**0.17**	**0.25**
I9	0.41	0.57	**0.43**	**0.57**	**0.41**	**0.56**	**0.15**	**0.26**
I10	0.41	0.56	**0.40**	**0.59**	**0.41**	**0.56**	**0.15**	**0.25**
I11	**0.25**	**0.56**	**0.40**	**0.56**	**0.42**	**0.56**	**0.15**	**0.25**
I12	**0.71**	**0.58**	**0.39**	**0.57**	**0.44**	**0.60**	**0.17**	**0.26**

Note: Communalities for items that are negatively keyed are bolded. Cont. stands for continuous; L# denotes the factor loading specified in the simulation model; NK# denotes the number of negatively keyed items; Asymm indicates observed item responses following asymmetric distributions; ML stands for maximum likelihood estimation; and WLSMV denotes weighted least squares means and variance adjusted estimation. The simulated (expected) communality levels for conditions noted with "L0.75" (first three conditions) and "L0.50" (last condition on the right) are 0.56 and 0.25, respectively.

Recall that for conditions simulated with factor loadings equal to 0.75 (noted as "L0.75" in the column headers of Table 6), the expected communality level is 0.56. As for those simulated with factor loadings equal to 0.50 (noted as "L0.50" in the column headers of Table 6), this value is 0.25. Compared with the communalities estimated using ML when the item responses are assumed to be continuous, those obtained through WLSMV on ordinal data are always higher and closer to the simulation values. Under each condition, the estimated values for the communality levels of positively and negatively keyed items are similar. No evidence suggests that the communality estimates for negatively keyed items are biased with one exception. The estimation of communalities for negatively keyed items is incorrect when ML is

used under the condition where the communality is high, two out of twelve items are negatively keyed, and the observed item responses are asymmetric (i.e., L0.75_NK2_Asymm). Under this condition, one negatively keyed item shows much higher communality level than all the other items, while the other negatively keyed item has a lower communality level (see the last two rows in the second column from the left in Table 6) under this condition.

Table 7 presents the fit statistics of the two-factor EFA solutions for these four conditions. As rotation does not change the fit of the factorial solution to the correlation matrix, the reported results does not include it here. For each condition investigated, four sets of model fit statistics were obtained (2 estimators × 2 scoring methods). As with item communalities, the two scoring methods for negatively keyed items produce the same model fit statistics, and they are thus not presented in Table 7. The fit statistics suggest a good fit for all these two-factor solutions.

Table 7.

Fit statistics for the two-factor EFA models

Conditions	Data Type (Estimator)		Chi-Square Test			RMSEA	SRMR
			Chi-Square	df	*p*		
L0.75_NK2_Asymm	Cont.	(ML)	48.89	43	0.248	0.004	0.004
	Ordinal	(WLSMV)	26.07	43	0.981	0.000	0.006
L0.75_NK4_Asymm	Cont.	(ML)	34.18	43	0.830	0.000	0.004
	Ordinal	(WLSMV)	41.08	43	0.555	0.000	0.008
L0.75_NK6_Asymm	Cont.	(ML)	21.06	43	0.998	0.000	0.003
	Ordinal	(WLSMV)	28.59	43	0.955	0.000	0.007
L0.50_NK6_Asymm	Cont.	(ML)	28.41	43	0.958	0.000	0.005
	Ordinal	(WLSMV)	28.63	43	0.955	0.000	0.009

Note: L# denotes the factor loading specified in the simulation model; NK# denotes the number of negatively keyed items; Asymm indicates observed item responses following asymmetric distributions; Cont. stands for continuous; ML stands for maximum likelihood estimation; and WLSMV denotes weighted least squares means and variance adjusted estimation.

As described above, a fully-crossed factorial design with three factors leads to a total of 12 factor structures (2 estimators × 2 scoring methods × 3 rotation methods) for each condition. Given that one of the purposes of this study is to inform day-to-day research, Varimax, Promax, and Geomin rotation methods were selected due to their popularity. Different rotations create different loadings and inter-factor correlations for a factor and therefore change the sums of the

squares of the loadings for the factor and the factor interpretation. The following tables (Table 8 to Table 15) present the factor loadings and the factor correlations of the two-factor EFA solutions under different conditions. The title of each table briefly describes the condition under investigation. In the table titles, *L#* represents factor loading, *NK#* represents the number of negatively keyed items, and *Asymm* indicates that the observed item response distribution is skewed (i.e., asymmetric). For example, *L0.75_NK2_Asymm with original item responses* (Table 8) means that the dataset fitted to the two-factor EFA solution was simulated from a one-factor model with factor loadings of 0.75 (i.e., *L0.75*), two negatively keyed items (i.e., *NK2*), skewed distributions for all observed item responses (i.e., *Asymm*), and the original item responses being used in the EFA modeling. Two estimation methods and three rotation methods resulted in six factor solutions under each of these conditions. Taking Table 8 as an example, the first three solutions were obtained with item responses being treated as continuous data. ML estimation was used, and Varimax, Promax, and Geomin rotations were applied. The last three solutions were obtained with ordinal item response data. WLSMV was employed for model estimation, and the same three types of rotation were applied. When oblique rotation methods were used, the corresponding factor correlations are presented in the bottom rows of each table.

As shown in tables (Table 8 to Table 15), when item responses were treated as continuous, two-factor solutions with factors defined by item keying directions appear under all the conditions. This pattern is consistent across different rotation methods. However, when item responses were treated as ordinal, the patterns of the factor loadings seem to suggest that the factors may be overly extracted from these datasets. The results show that the factor solutions vary with different rotation methods. When Varimax was applied, items show cross-loadings on both factors in most of the conditions (see tables Table 8 to Table 15). When Geomin was used, items largely load on the first factor and the second factor seems redundant. The patterns of the factor loadings from Promax were inconsistent across different datasets, but the solutions resembled either those obtained from Varimax or those obtained from Geomin.

Table 8.

Factor loadings obtained from two-factor EFA solutions: L0.75_NK2_Asymm with original item responses

	Continuous						Ordinal					
	Orthogonal (Varimax)		Oblique (Promax)		Oblique (Geomin)		Orthogonal (Varimax)		Oblique (Promax)		Oblique (Geomin)	
	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
I1	0.63	0.18	0.65	0.01	0.66	0.01	0.73	0.20	0.76	0.01	0.76	0.12
I2	0.61	0.19	0.63	0.03	0.64	-0.01	0.73	0.20	0.75	0.01	0.75	0.02
I3	0.63	0.18	0.65	0.01	0.66	0.00	0.73	0.20	0.76	0.01	0.76	-0.03
I4	0.62	0.18	0.63	0.02	0.64	0.00	0.73	0.18	0.76	-0.01	0.75	-0.01
I5	0.63	0.18	0.64	0.01	0.65	0.00	0.73	0.20	0.76	0.01	0.76	0.06
I6	0.60	0.18	0.61	0.03	0.62	-0.01	0.71	0.20	0.73	0.01	0.73	0.09
I7	0.63	0.18	0.65	0.01	0.65	0.00	0.74	0.19	0.76	0.00	0.76	-0.05
I8	0.63	0.18	0.65	0.01	0.66	0.00	0.74	0.20	0.76	0.00	0.77	-0.06
I9	0.62	0.18	0.63	0.02	0.64	0.00	0.73	0.19	0.76	-0.01	0.76	-0.03
I10	0.62	0.18	0.64	0.01	0.65	0.00	0.39	1.44*	0.11	1.44*	0.75	0.07
I11	**-0.17**	**-0.61**	**-0.01**	**-0.63**	**0.00**	**0.63**	**-0.72**	**-0.20**	**-0.74**	**-0.01**	**-0.76**	**0.13**
I12	**-0.18**	**-0.61**	**-0.02**	**-0.63**	**0.00**	**0.63**	**-0.73**	**-0.21**	**-0.75**	**-0.01**	**-0.75**	**-0.13**
Factor Correlations												
	--		0.45		-0.53		--		0.44		0.04	

Note: L# denotes the factor loading specified in the simulation model; NK# denotes the number of negatively keyed items; Asymm indicates observed item responses following asymmetric distributions. Factor loadings for negatively keyed items are highlighted in bold. * indicates instances where the residual variance estimated for that item was negative.

Table 9.

Factor loadings obtained from two-factor EFA solutions: L0.75_NK2_Asymm with reverse scored item responses for negatively keyed items

	Continuous						Ordinal					
	Orthogonal (Varimax)		Oblique (Promax)		Oblique (Geomin)		Orthogonal (Varimax)		Oblique (Promax)		Oblique (Geomin)	
	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
I1	0.63	0.18	0.65	0.01	0.66	-0.01	0.61	0.47	0.55	0.25	0.76	0.12
I2	0.61	0.19	0.63	0.03	0.64	0.01	0.54	0.53	0.41	0.39	0.75	0.02
I3	0.63	0.18	0.65	0.01	0.66	0.00	0.51	0.57	0.34	0.46	0.76	-0.03
I4	0.62	0.18	0.63	0.02	0.64	0.00	0.52	0.55	0.36	0.43	0.75	-0.01
I5	0.63	0.18	0.64	0.01	0.65	0.00	0.58	0.50	0.48	0.32	0.76	0.06
I6	0.60	0.18	0.61	0.03	0.62	0.01	0.58	0.47	0.51	0.27	0.73	0.09
I7	0.63	0.18	0.65	0.01	0.65	0.00	0.49	0.59	0.30	0.51	0.76	-0.05
I8	0.63	0.18	0.65	0.01	0.66	0.00	0.50	0.59	0.30	0.51	0.77	-0.06
I9	0.62	0.18	0.63	0.02	0.64	0.00	0.51	0.56	0.34	0.45	0.76	-0.03
I10	0.62	0.18	0.64	0.01	0.65	0.00	0.58	0.49	0.50	0.30	0.75	0.07
I11	**0.17**	**0.61**	**0.01**	**0.63**	**0.00**	**0.63**	**0.44**	**0.63**	**0.19**	**0.60**	**0.76**	**-0.13**
I12	**0.18**	**0.61**	**0.02**	**0.63**	**0.00**	**0.63**	**0.62**	**0.45**	**0.59**	**0.22**	**0.75**	**0.13**
Factor Correlations												
	--		0.50		0.53		--		0.79		0.04	

Note: L# denotes the factor loading specified in the simulation model; NK# denotes the number of negatively keyed items; Asymm indicates observed item responses following asymmetric distributions. Factor loadings for negatively keyed items are highlighted in bold.

Table 10.

Factor loadings obtained from two-factor EFA solutions: L0.75_NK4_Asymm with original item responses

	Continuous						Ordinal					
	Orthogonal (Varimax)		Oblique (Promax)		Oblique (Geomin)		Orthogonal (Varimax)		Oblique (Promax)		Oblique (Geomin)	
	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
I1	0.63	0.18	0.65	0.02	0.65	0.00	0.70	0.32	0.75	0.03	0.76	-0.03
I2	0.61	0.18	0.63	0.03	0.64	-0.01	0.68	0.32	0.73	0.04	0.75	-0.03
I3	0.63	0.18	0.65	0.02	0.65	0.00	0.69	0.33	0.73	0.05	0.77	0.03
I4	0.62	0.17	0.64	0.01	0.65	0.01	0.67	0.33	0.71	0.05	0.74	-0.05
I5	0.63	0.17	0.64	0.01	0.65	0.00	0.70	0.30	0.76	0.01	0.75	-0.06
I6	0.60	0.18	0.61	0.03	0.61	-0.02	0.68	0.30	0.74	0.01	0.73	-0.08
I7	0.63	0.18	0.65	0.01	0.66	0.00	0.69	0.34	0.72	0.06	0.78	0.06
I8	0.63	0.17	0.65	0.01	0.66	0.01	0.42	0.91	0.16	0.89	0.77	0.01
I9	**-0.17**	**-0.63**	**0.00**	**-0.65**	**0.01**	**0.66**	**-0.68**	**-0.32**	**-0.73**	**-0.03**	**-0.75**	**-0.01**
I10	**-0.17**	**-0.61**	**-0.01**	**-0.63**	**0.00**	**0.64**	**-0.70**	**-0.31**	**-0.75**	**-0.03**	**-0.79**	**-0.10**
I11	**-0.18**	**-0.60**	**-0.02**	**-0.62**	**0.00**	**0.63**	**-0.66**	**-0.35**	**-0.69**	**-0.09**	**-0.75**	**-0.03**
I12	**-0.18**	**-0.60**	**-0.03**	**-0.61**	**-0.01**	**0.62**	**-0.68**	**-0.32**	**-0.73**	**-0.05**	**-0.71**	**0.30**
Factor Correlations												
	--		0.50		-0.53		--		0.65		-0.15	

Note: L# denotes the factor loading specified in the simulation model; NK# denotes the number of negatively keyed items; Asymm indicates observed item responses following asymmetric distributions. Factor loadings for negatively keyed items are highlighted in bold.

Table 11.

Factor loadings obtained from two-factor EFA solutions: L0.75_NK4_Asymm with reverse scored item responses for negatively keyed items

	Continuous						Ordinal					
	Orthogonal (Varimax)		Oblique (Promax)		Oblique (Geomin)		Orthogonal (Varimax)		Oblique (Promax)		Oblique (Geomin)	
	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
I1	0.63	0.18	0.65	0.02	0.65	0.00	0.59	0.48	0.54	0.27	0.76	0.03
I2	0.61	0.18	0.63	0.03	0.64	0.01	0.58	0.48	0.53	0.27	0.75	0.03
I3	0.63	0.18	0.65	0.02	0.65	0.00	0.63	0.44	0.62	0.17	0.77	-0.03
I4	0.62	0.17	0.64	0.01	0.65	-0.01	0.57	0.49	0.51	0.28	0.74	0.05
I5	0.63	0.17	0.64	0.01	0.65	0.00	0.57	0.50	0.50	0.30	0.75	0.06
I6	0.60	0.18	0.61	0.03	0.61	0.02	0.54	0.50	0.46	0.33	0.73	0.08
I7	0.63	0.18	0.65	0.01	0.66	0.00	0.65	0.42	0.67	0.13	0.78	-0.06
I8	0.63	0.17	0.65	0.01	0.66	-0.01	0.62	0.46	0.60	0.21	0.77	-0.01
I9	**0.17**	**0.63**	**0.00**	**0.65**	**-0.01**	**0.66**	**0.61**	**0.44**	**0.59**	**0.20**	**0.75**	**-0.01**
I10	**0.17**	**0.61**	**0.01**	**0.63**	**0.00**	**0.64**	**0.68**	**0.39**	**0.73**	**0.06**	**0.79**	**-0.10**
I11	**0.18**	**0.60**	**0.02**	**0.62**	**0.00**	**0.63**	**0.62**	**0.42**	**0.62**	**0.16**	**0.75**	**-0.03**
I12	**0.18**	**0.60**	**0.03**	**0.61**	**0.01**	**0.62**	**0.42**	**0.71**	**0.14**	**0.71**	**0.71**	**0.30**
Factor Correlations												
	--		0.50		0.53		--		0.78		0.15	

Note: L# denotes the factor loading specified in the simulation model; NK# denotes the number of negatively keyed items; Asymm indicates observed item responses following asymmetric distributions. Factor loadings for negatively keyed items are highlighted in bold.

Table 12.

Factor loadings obtained from two-factor EFA solutions: L0.75_NK6_Asymm with original item responses

	Continuous						Ordinal					
	Orthogonal (Varimax)		Oblique (Promax)		Oblique (Geomin)		Orthogonal (Varimax)		Oblique (Promax)		Oblique (Geomin)	
	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
I1	0.61	0.17	0.62	0.02	0.63	0.00	0.72	0.41	0.71	0.16	-0.71	0.11
I2	0.62	0.17	0.64	0.01	0.65	0.00	0.42	0.63	0.15	0.63	-0.74	0.07
I3	0.63	0.17	0.65	0.01	0.66	0.01	0.39	0.66	0.09	0.70	-0.75	0.04
I4	0.61	0.18	0.62	0.02	0.63	-0.01	0.43	0.61	0.18	0.60	-0.71	0.11
I5	0.63	0.17	0.65	0.01	0.66	0.01	0.42	0.64	0.14	0.65	-0.76	0.02
I6	0.61	0.17	0.63	0.02	0.63	-0.01	0.41	0.63	0.14	0.64	-0.75	-0.01
I7	**-0.17**	**-0.62**	**-0.02**	**-0.63**	**0.00**	**0.64**	**-0.44**	**-0.61**	**-0.19**	**-0.59**	**0.75**	**0.01**
I8	**-0.18**	**-0.60**	**-0.03**	**-0.61**	**-0.02**	**0.62**	**-0.48**	**-0.59**	**-0.26**	**-0.55**	**0.68**	**-0.30**
I9	**-0.17**	**-0.62**	**-0.01**	**-0.64**	**0.00**	**0.65**	**-0.35**	**-0.67**	**-0.03**	**-0.74**	**0.76**	**0.03**
I10	**-0.17**	**-0.62**	**-0.02**	**-0.63**	**-0.01**	**0.64**	**-0.38**	**-0.65**	**-0.09**	**-0.68**	**0.75**	**0.02**
I11	**-0.17**	**-0.62**	**-0.01**	**-0.64**	**0.00**	**0.65**	**-0.42**	**-0.62**	**-0.16**	**-0.62**	**0.76**	**0.03**
I12	**-0.17**	**-0.64**	**0.00**	**-0.66**	**0.01**	**0.67**	**-0.44**	**-0.63**	**-0.18**	**-0.63**	**0.78**	**0.03**
Factor Correlations												
	--		0.48		-0.51		--		0.75		-0.22	

Note: L# denotes the factor loading specified in the simulation model; NK# denotes the number of negatively keyed items; Asymm indicates observed item responses following asymmetric distributions. Factor loadings for negatively keyed items are highlighted in bold.

Table 13.

Factor loadings obtained from two-factor EFA solutions: L0.75_NK6_Asymm with reverse scored item responses for negatively keyed items

	Continuous						Ordinal					
	Orthogonal (Varimax)		Oblique (Promax)		Oblique (Geomin)		Orthogonal (Varimax)		Oblique (Promax)		Oblique (Geomin)	
	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
I1	0.61	0.17	0.62	0.02	0.63	0.00	0.55	0.50	0.49	0.30	0.71	0.11
I2	0.62	0.17	0.64	0.01	0.65	0.00	0.58	0.48	0.54	0.25	0.74	0.07
I3	0.63	0.17	0.65	0.01	0.66	-0.01	0.60	0.47	0.57	0.23	0.75	0.04
I4	0.61	0.18	0.62	0.02	0.63	0.01	0.55	0.50	0.48	0.31	0.71	0.11
I5	0.63	0.17	0.65	0.01	0.66	-0.01	0.62	0.45	0.61	0.18	0.76	0.02
I6	0.61	0.17	0.63	0.02	0.63	0.01	0.62	0.42	0.63	0.15	0.75	-0.01
I7	**0.17**	**0.62**	**0.02**	**0.63**	**0.00**	**0.64**	**0.62**	**0.42**	**0.63**	**0.15**	**0.75**	**-0.01**
I8	**0.18**	**0.60**	**0.03**	**0.61**	**0.02**	**0.62**	**0.43**	**0.70**	**0.17**	**0.68**	**0.68**	**0.30**
I9	**0.17**	**0.62**	**0.01**	**0.64**	**0.00**	**0.65**	**0.63**	**0.41**	**0.65**	**0.13**	**0.76**	**-0.03**
I10	**0.17**	**0.62**	**0.02**	**0.63**	**0.01**	**0.64**	**0.63**	**0.41**	**0.64**	**0.13**	**0.75**	**-0.02**
I11	**0.17**	**0.62**	**0.01**	**0.64**	**0.00**	**0.65**	**0.63**	**0.41**	**0.65**	**0.13**	**0.76**	**-0.03**
I12	**0.17**	**0.64**	**0.00**	**0.66**	**-0.01**	**0.67**	**0.65**	**0.42**	**0.67**	**0.13**	**0.78**	**-0.03**
Factor Correlations												
	--		0.48		0.51		--		0.77		0.22	

Note: L# denotes the factor loading specified in the simulation model; NK# denotes the number of negatively keyed items; Asymm indicates the observed item responses followed asymmetric distributions. Factor loadings for negatively keyed items are highlighted in bold.

Table 14.

Factor loadings obtained from two-factor EFA solutions: L0.50_NK6_Asymm with original item responses

	Continuous						Ordinal					
	Orthogonal (Varimax)		Oblique (Promax)		Oblique (Geomin)		Orthogonal (Varimax)		Oblique (Promax)		Oblique (Geomin)	
	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
I1	0.38	0.17	0.40	0.02	0.42	0.00	0.41	0.32	0.38	0.18	0.52	0.00
I2	0.35	0.18	0.35	0.06	0.36	-0.04	0.39	0.32	0.35	0.18	0.51	-0.01
I3	0.38	0.17	0.40	0.03	0.42	0.00	0.47	0.24	0.51	0.02	0.54	0.10
I4	0.36	0.16	0.38	0.03	0.39	0.00	0.38	0.31	0.34	0.18	0.49	-0.01
I5	0.39	0.16	0.41	0.01	0.43	0.02	0.40	0.31	0.38	0.16	0.51	0.01
I6	0.30	0.19	0.28	0.10	0.29	-0.09	0.29	0.40	0.15	0.37	0.45	-0.14
I7	**-0.15**	**-0.38**	**-0.01**	**-0.41**	**0.03**	**0.44**	**-0.25**	**-0.49**	**-0.04**	**-0.52**	**-0.45**	**0.23**
I8	**-0.15**	**-0.38**	**0.00**	**-0.41**	**0.04**	**0.44**	**-0.32**	**-0.38**	**-0.21**	**-0.32**	**-0.47**	**0.10**
I9	**-0.18**	**-0.35**	**-0.05**	**-0.36**	**-0.02**	**0.38**	**-0.43**	**-0.28**	**-0.43**	**-0.10**	**-0.52**	**-0.04**
I10	**-0.18**	**-0.34**	**-0.06**	**-0.34**	**-0.03**	**0.36**	**-0.42**	**-0.28**	**-0.41**	**-0.11**	**-0.51**	**-0.04**
I11	**-0.18**	**-0.34**	**-0.06**	**-0.34**	**-0.03**	**0.36**	**-0.42**	**-0.28**	**-0.41**	**-0.11**	**-0.51**	**-0.04**
I12	**-0.18**	**-0.37**	**-0.04**	**-0.38**	**-0.01**	**0.40**	**-0.33**	**-0.40**	**-0.21**	**-0.34**	**-0.48**	**0.12**
Factor Correlations												
	--		0.67		-0.75		--		0.77		-0.21	

Note: L# denotes the factor loading specified in the simulation model; NK# denotes the number of negatively keyed items; Asymm indicates observed item responses following asymmetric distributions. Factor loadings for negatively keyed items are highlighted in bold.

Table 15.

Factor loadings obtained from two-factor EFA solutions: L0.50_NK6_Asymm with reverse scored item responses for negatively keyed items

	Continuous						Ordinal					
	Orthogonal (Varimax)		Oblique (Promax)		Oblique (Geomin)		Orthogonal (Varimax)		Oblique (Promax)		Oblique (Geomin)	
	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
I1	0.38	0.17	0.40	0.02	0.42	0.00	0.41	0.32	0.38	0.17	0.52	0.00
I2	0.35	0.18	0.35	0.06	0.36	0.04	0.40	0.32	0.36	0.18	0.51	0.01
I3	0.38	0.17	0.40	0.03	0.42	0.00	0.47	0.24	0.51	0.02	0.54	-0.10
I4	0.36	0.16	0.38	0.03	0.39	0.00	0.38	0.31	0.34	0.18	0.49	0.01
I5	0.39	0.16	0.41	0.01	0.43	-0.02	0.40	0.30	0.38	0.15	0.51	-0.01
I6	0.30	0.19	0.28	0.10	0.29	0.09	0.29	0.40	0.16	0.37	0.45	0.14
I7	**0.15**	**0.38**	**0.01**	**0.41**	**-0.03**	**0.44**	**0.25**	**0.49**	**0.04**	**0.52**	**0.45**	**0.23**
I8	**0.15**	**0.38**	**0.00**	**0.41**	**-0.04**	**0.44**	**0.33**	**0.38**	**0.22**	**0.31**	**0.47**	**0.10**
I9	**0.18**	**0.35**	**0.05**	**0.36**	**0.02**	**0.38**	**0.43**	**0.28**	**0.43**	**0.10**	**0.52**	**-0.04**
I10	**0.18**	**0.34**	**0.06**	**0.34**	**0.03**	**0.36**	**0.42**	**0.28**	**0.42**	**0.11**	**0.51**	**-0.04**
I11	**0.18**	**0.34**	**0.06**	**0.34**	**0.03**	**0.36**	**0.42**	**0.28**	**0.41**	**0.11**	**0.51**	**-0.04**
I12	**0.18**	**0.37**	**0.04**	**0.38**	**0.01**	**0.40**	**0.33**	**0.40**	**0.22**	**0.34**	**0.48**	**0.12**
Factor Correlations												
	--		0.67		0.75		--		0.77		0.21	

Note: L# denotes the factor loading specified in the simulation model; NK# denotes the number of negatively keyed items; Asymm indicates observed item responses following asymmetric distributions. Factor loadings for negatively keyed items are highlighted in bold.

As shown by the factor loadings in the above tables (Table 8 to Table 15), when item response data were treated as continuous, regardless of the rotation method used, a two-factor structure was found with factors essentially formed according to item keying direction. When item responses were treated as ordinal, Varimax and Promax rotations produced factor structures with relatively heavy cross-loadings, and only Geomin seemed to recover the true data structure. The two-factor structures produced by Geomin consist of a primary factor and a nuisance factor on which all items load weakly. This data structure is similar to the psychometric model that simulated the data.

When deciding on the number of factors using eigenvalues, it is conventional to rely on PCA based on Pearson correlations. After the eigenvalues are obtained through PCA, they are evaluated via different decision rules, such as the K-G rule and PA. To explore whether using

polychoric correlations in the PCA stage will suggest a different number of factors to retain, the same decision rules are applied to eigenvalues based on polychoric correlation matrices. Table 16 presents the number of factors suggested by the K-G rule. The PA results suggest one-factor solutions under all the conditions, and therefore are not listed in the table. One could accurately conclude that using PA with polychoric correlations in the PCA stage to obtain eigenvalues and Pearson correlations in the data simulation stage produces the correct number of factors under all the simulated conditions. Comparing the results based on polychoric correlations with those based on Pearson correlations suggests that treating item responses as ordinal and calculating eigenvalues based on polychoric correlations helps point to the right number of factors to retain even when the observed response distributions are skewed. When the response distributions are symmetric, the K-G rule and PA both point to the right number of factors with eigenvalues calculated either based on Pearson or polychoric correlations.

Table 16.

Number of factors: K-G rule based on polychoric correlation matrix

Number of NK	Item Loadings	Observed Item Response Distribution Asymmetric (Skewness = -2)
0 (Baseline)	0.25	3
	0.50	1
	0.75	1
2	0.25	3
	0.50	1
	0.75	1
4	0.25	3
	0.50	1
	0.75	1
6	0.25	3
	0.50	1
	0.75	1

Note: NK denotes negatively keyed items.

To summarize the findings from Study 1, responses to each of the research questions are presented as follows.

Q1.1: Does the reverse scoring of negatively keyed items affect the number of factors identified by the K-G rule and PA?

Neither reverse scoring the negatively keyed items nor leaving them in their original format affects the number of factors identified by the K-G rule or PA.

Q1.2: Under what conditions will the K-G rule correctly point to the number of factors?

The K-G rule, when applied to PCA eigenvalues from either Pearson or polychoric correlations, points to the right number of factors when the observed item responses are symmetric. When these response are asymmetric, applying the K-G rule to eigenvalues obtained from polychoric correlations points to the right number of factors when the communality level is median (0.25) or high (0.56). The K-G rule's performance is not altered by the presence of negatively keyed items. In fact, it shows inflation in the suggested number of factors even without the presence of negatively keyed items when the communality level is low.

Q1.3: Under what conditions will PA correctly point to the number of factors?

PA points to the right number of factors under most conditions. Only when the observed item response distribution is asymmetric, the communality level is high, the number of negatively keyed items is relatively large, and the Pearson correlation matrix is used at the PCA stage will PA inflate the number of factors.

Q1.4: When more than one factor is suggested for retention, what will the factor structure look like? Will a second factor be formed by negatively keyed items in EFA?

Under conditions where PA points to more than one factor, the suggested number is two. If the follow-up EFA is conducted with ML estimation (i.e., treating item responses as continuous), a two-factor solution with factors formed by item keying direction will be observed under each of these conditions, regardless of the rotation methods applied. Note that all of the two-factor models are incorrect because the data were simulated through a one-factor model. The correct factor structure is only recovered under conditions where the item response data are treated as ordinal in EFA and rotated by Geomin.

Study 2: The impact of negatively keyed items on the model fit in CFA

The purpose of the second simulation study is to investigate the effect of negatively keyed items on assessing the factor structure of a test using CFA. One of the advantages of CFA is its ability to offer multiple indices to evaluate the quality of the model fit. These fit indices

assist researchers in deciding whether to reject or tentatively retain an *a priori* specified model. Acceptable fit indices usually lead to the conclusion that none of the evidence collected supports the rejection of the model. However, if the indices indicate a potential model misfit, researchers may revise the original model either based on statistics, such as modification indices, or on previous research findings. To demonstrate how modifying models using modification indices may influence the interpretation of the factor structure of mixed-keyed tests, modification indices are presented when the one-factor model is rejected by CFA fit indices.

More specifically, the research questions that this study aims to address are as follows:

Q2.1: Does the reverse scoring of negatively keyed items affect the assessment of model fit?

Q2.2: Under what conditions will the one-factor model be rejected either by the Chi-square test or other fit indices?

Q2.3: When the one-factor model is rejected, which parameters will the modification indices suggest to be freed? Will these additional parameters support an alternative model that accounts for the item keying effect?

Method

Study design

Following the same data simulation design and procedures as Study 1, the present study examines whether model fit statistics from CFA support the one-factor structure in the presence of negatively keyed items. It uses the same datasets as Study 1. Each dataset consists of 12 items and 10,000 respondents. These datasets are simulated by varying three factors in a factorial design: (a) the magnitude of communality in the factor structure, (b) the number of negatively keyed items, and (c) the distribution of observed item responses. The study explores three levels of communality, which are manipulated by varying the factor loadings in the data simulation model. Item loadings of 0.25 represent a low level of communality ($h^2 = 0.06$), those of 0.50 a medium level ($h^2 = 0.25$), and those of 0.75 a high level ($h^2 = 0.56$). The number of negatively keyed items is set at 0, 2, 4 and 6. The distribution of observed item responses is either symmetric with a skewness of 0 or asymmetric with a skewness of -2. The manipulation of the observed item response distributions is done by changing the thresholds used in the transformation from latent response distributions to observed responses. The simulation follows

a 3×4×2 completely crossed design. As in Study 1, two scoring methods are applied to negatively keyed items. The two scoring methods are (a) using item responses in their original format for both positively and negatively keyed items, and (b) reverse scoring the item responses to negatively keyed items while leaving those to positively keyed items untouched. In total, 42 unique datasets are created.

As in Study 1, item responses are treated either as continuous or ordinal data when factor analysis is performed. CFA is conducted with different estimators to accommodate different data types. ML and MLR are used for continuous item responses, and WLSMV is utilized for ordinal data. The global model fit indices are of primary interest in this study. Following the guidelines suggested by Jackson, Gillaspy, and Purc-Stephenson (2009), model fit is assessed using various fit statistics. Four fit indices are used to evaluate the goodness of fit: (a) Chi-square test statistics, (b) CFI, (c) TLI, and (d) RMSEA with associated confidence intervals (CIs).

In cases where these fit indices suggest a poor model fit, the study uses modification indices to identify places where the misfit occurs. Allowing model fit to drive the process of dimensionality assessment deviates from the theory-testing purpose of CFA. The parameters associated with large modification indices are presented. The intention is to demonstrate how alternative models can be derived by allowing modification indices to shape the modeling process. These alternative models, however, are not tested further.

Procedures

Figure 12 is a flow chart showing the major steps in conducting this study. The data simulation design and the first three steps in the procedure are the same as those used in Study 1. The description therefore focuses on the data analysis methods (i.e., steps 4 and 5).

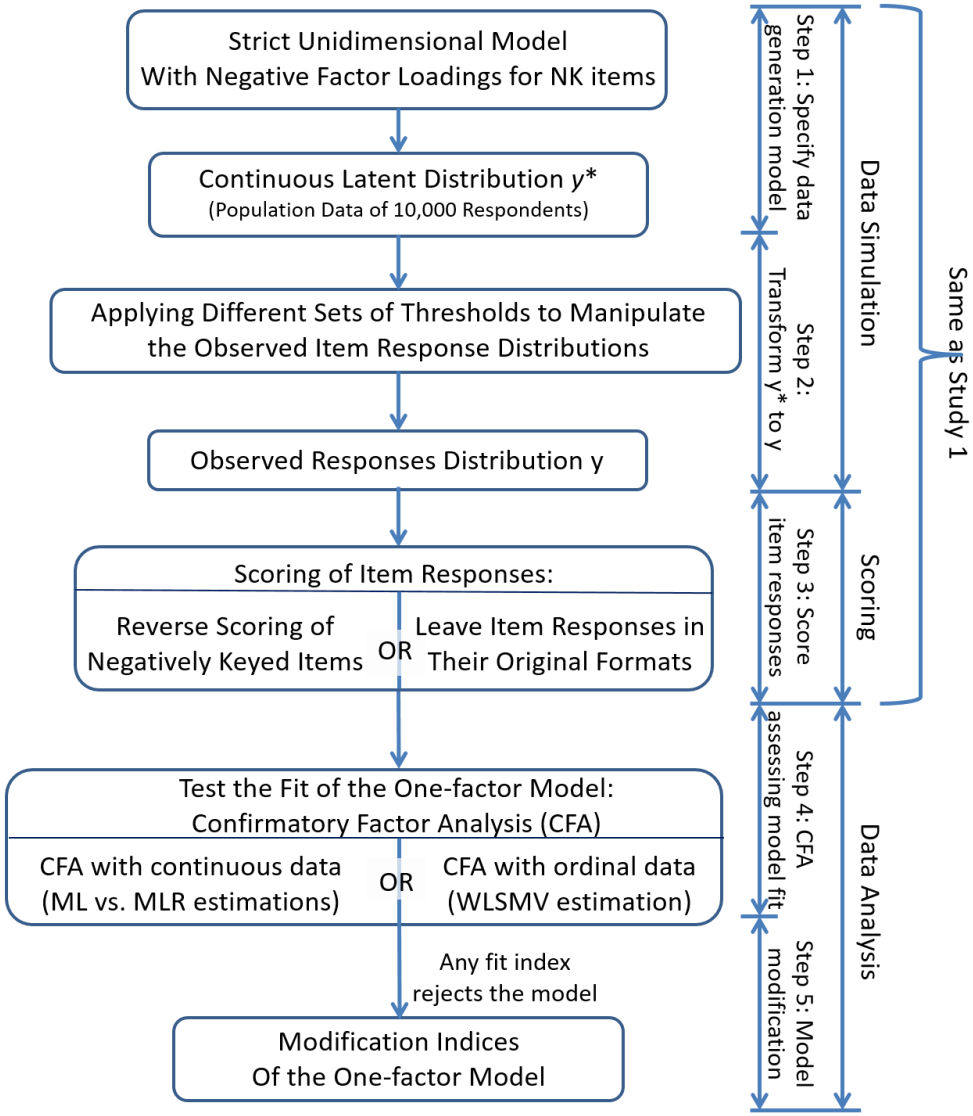


Figure 12.

A flow chart representing the research process of Study 2

Step 1: Specify the data generation model: Same as in Study 1.

Step 2: Transform continuous latent response distributions to observed responses: Same as in Study 1.

Step 3: Score the item responses: Same as in Study 1.

Step 4: Evaluate the fit of the one-factor CFA model. To examine whether the one-factor model fits well for each of the simulated datasets, all the datasets are analyzed using CFA. For each dataset, item responses are treated as either continuous or ordinal. Therefore, the same one-

factor model is tested via ML or MLR with continuous item responses and WLSMV with ordinal responses.

Step 5: Use modification indices. When any of the fit indices rejects the one-factor model, modification indices will be requested. This is because, in day-to-day practice, researchers usually do not follow a strictly confirmatory approach to test a theorized model. They are often willing to consider minor modifications based on the results of analyses.

Results and conclusions

The results show that the judgment of the model fit is not affected by the scoring methods applied to the negatively keyed items. Reverse scoring these items or leaving all item responses in their original format lead to the same statistical conclusion regarding the model fit. Therefore, the following results are presented without distinguishing the scoring method used for negatively keyed items.

One-factor CFA model fit indices with ML estimation

When the item responses are treated as continuous, the one-factor model is first tested via CFA using ML estimation. When the observed item response distributions are symmetric, a one-factor model is always supported by the Chi-square test ($p > .05$), CFI (> 0.95), TLI (> 0.95), and RMSEA (< 0.05). When the observed item response distribution is asymmetric, however, the decision regarding model fit varies depending on the number of negatively keyed items, item communality levels, and the decision rules. Table 17 presents the Chi-square test results and the model fit indices under each of the conditions with asymmetric item response distributions. The first column on the left lists the number of negatively keyed items, while the second indicates the item loadings specified in the simulation model. These different item loadings represent different levels of item communality. The two columns in the middle list the judgment of the model fit, which is made either based on the Chi-square test or on a combined rule of CFI, TLI, and RMSEA values. The last column indicates the consistency of the judgment based on different rules.

As Table 17 makes clear, out of the 12 simulation conditions with different numbers of negatively keyed items and levels of communality, the Chi-square test ($p < .05$) rejected the one-factor model under seven of them. The results suggest that when the observed item responses are

distributed asymmetrically and the level of communality is high (h^2 = 0.56 or factor loadings = 0.75), the one-factor model is likely to be rejected by the Chi-square test regardless of whether any negatively keyed items are included. If negatively keyed items are present (n ≥ 2), the Chi-square test rejects the one-factor model when the observed item responses are distributed asymmetrically and the level of communality is medium to high (h^2 = 0.25 to 0.56).

Descriptive fit indices, such as CFI, TLI, and RMSEA, are widely used. The cut-off values used for judging the model fit are CFI ≥ 0.95, TLI ≥ 0.95, and RMSEA ≤ 0.05. The results regarding the conclusion of model fit are consistent across these three fit indices and any combination of them. Therefore, the judgment of model fit based on these fit indices is presented in one column. Out of the 12 simulation conditions (i.e., 3 levels of communality × 4 levels of the number of negatively keyed items) with asymmetric item response distributions, the fit indices wrongly rejected the one-factor model under two (see Table 17). These two conditions feature asymmetric observed item responses, a high communality level (h^2 = 0.56), and a relatively large number of negatively keyed items (4 or 6 items).

Table 17.

One-factor CFA model with ML estimation

Number of NK Items	Item Loadings	Observed Item Response Distribution Asymmetric (Skewness = -2)		
		Model fit assessed by Chi-square test	Model fit assessed by CFI / TLI / RMSEA	Decision consistency: Chi-square test vs. other fit indices
0 (Baseline)	0.25	Y	Y	CON
	0.50	Y	Y	CON
	0.75	N	Y	Inconsistent
2	0.25	Y	Y	CON
	0.50	N	Y	Inconsistent
	0.75	N	Y	Inconsistent
4	0.25	Y	Y	CON
	0.50	N	Y	Inconsistent
	0.75	N	N	CON
6	0.25	Y	Y	CON
	0.50	N	Y	Inconsistent
	0.75	N	N	CON

Note: NK denotes negatively keyed items. N means that the one-factor model is rejected by the Chi-square test ($p < 0.05$), or CFI (< 0.90), TLI (< 0.90), and RMSEA (> 0.08); Y denotes that the model is supported by the Chi-square test or the three descriptive fit indices; CON means that different decision rules reach consistent decisions regarding overall model fit.

The results presented in Table 17 also suggest that using the Chi-square test or the descriptive fit indices leads to the same conclusions regarding model fit in many cases. However, the conclusions are inconsistent under five conditions, all of which have asymmetric item response distributions and are marked by "Inconsistent" in the last column of Table 17. When this inconsistency between the Chi-square test and the descriptive fit indices occurs, the pattern is that the former rejects the one-factor model while the latter support it. This suggests that the Chi-square test wrongly rejects the one-factor model under more conditions than the descriptive fit indices do when the observed item response distribution is asymmetric.

Modification indices in cases where model fit is poor (ML estimation)

Parameters with large modification indices are recorded when the one-factor model is rejected. This can be viewed as a supplement to the assessment of overall model fit, as it

indicates the area of potential misfit. This is meant to mimic the common practice of using modification indices to modify the originally hypothesized model.

As Table 17 demonstrates, the Chi-square test rejects the one-factor model under seven conditions. Modification indices are reviewed under these conditions. The following table (Table 18) presents the parameters recommended to be added to the one-factor model to achieve a better model fit. The suggested parameters are listed in the order of their Chi-square change values, from high to low. When there are more than five parameters with significant modification indices, only the five with the largest Chi-square change values are listed. Note that none of the modification indices is statistically significant ($p > 0.05$) under the condition where item loading equals 0.75, none of the items is negatively keyed, and the observed item response distribution is asymmetric. Therefore, Table 18 presents only the results from the other six conditions.

Table 18.

Parameters suggested to be freed by modification indices (ML estimation)

	Observed Item Response Distribution: Asymmetric (Skewness = -2)					
Number of NK Items	NK = 2		NK = 4		NK = 6	
Factor Loading	L0.50	L0.75	L0.50	L0.75	L0.50	L0.75
Suggested Parameters	I12 with I11	I12 with I11	I12 with I11	I10 with I9	I12 with I7	I5 with I3
			I10 with I9	I12 with I9	I8 with I7	I5 with I2
			I11 with I9	I11 with I9	I5 with I1	I12 with I10
			I12 with I9	I11 with I10	**I7 with I5**	I6 with I3
			I12 with I10	I12 with I11	**I7 with I3**	I3 with I2

Note: Only parameters associated with a significant Chi-square improvement are listed; when multiple parameters meet this criterion, only the first five are listed and they are ordered from top to bottom by their Chi-square change value. Parameters represent correlated error terms between items keyed in different directions are highlighted in bold.

Each column in Table 18 represents a simulated condition where a one-factor model is rejected by the Chi-square test. For example, the first two columns list the two conditions with asymmetric item response distributions and two negatively keyed items. These conditions differ in their communality levels or the factor loadings in the simulation model. Similarly, the remaining columns contain conditions grouped by the number of negatively keyed items. As Table 18 confirms, the suggested parameters are mostly between negatively keyed items. Only in

balanced-tests, where the number of positively and negatively keyed items is equal, do the suggested changes include parameters between items keyed in different directions. As observed in this simulation study, if researchers draw on modification indices to revise their one-factor model, it is likely that they will improve it by allowing correlated error terms between negatively keyed items when these items do not exceed the positively keyed items in a test (i.e., see Model 3b in Figure 5). It is worth noting that, although the one-factor model is the correct choice, and the models based on modification indices are incorrect, the model fit improved, as indicated by various fit indices.

One-factor CFA model fit indices with robust estimation method MLR

As shown above, the one-factor model can be wrongly rejected when the observed item response distributions are skewed and the ML estimator is used in CFA. The literature suggests that continuous MLR often performs as well as categorical data estimation methods on response data that have five or more rating points (Rhemtulla et al., 2012). Thus, the robust estimator is often preferred to ML with continuous data since it is believed to reach more accurate results. To explore whether using a robust estimator will lead to different decisions on the model fit than would drawing on ML estimation, the same one-factor model is tested using robust maximum likelihood (MLR) under all the simulated conditions.

When the observed item response distributions are symmetric, the one-factor model is always supported by the Chi-square test, with *p*s > 0.05, CFIs > 0.95, TLIs > 0.95, and RMSEAs < 0.05, regardless of the number of negatively keyed items or the item communality levels. The model is rejected only under some conditions with asymmetric response distributions and negatively keyed items. The judgment of model fit under conditions with asymmetric item response distributions is summarized in Table 19. The structure of Table 19 is the same as that of Table 17. The two columns in the middle list the judgment of the model fit, either based on the Chi-square test or on a combined rule of CFI, TLI, and RMSEA values. Just as when ML estimation is used with CFA, the judgment of model fit, as indicated by CFI, TLI, and RMSEA, is consistent across all the conditions, and thus, these results are presented together. The last column shows the consistency of the judgment based on different rules.

As Table 19 shows, the Chi-square test rejects the one-factor model under five conditions. All feature asymmetric item response distributions, negatively keyed items, and relatively high

communality levels. When the fit is judged by the descriptive fit indices, the one-factor model is rejected with RMSEA > 0.08, CFI < 0.90, and TLI < 0.90 under two conditions, both of which have skewed item response distributions, high communality levels (h^2 = 0.56, or factor loadings = 0.75), and four or more negatively keyed items out of 12 (approximately 33% or more of the total). Note that when the one-factor model was rejected by the descriptive fit indices (CFI, TLI, and RMSEA), the model was also rejected by the Chi-square test.

Table 19.

One-factor CFA model with MLR estimation

Number of NK Items	Item Loadings	Observed Item Response Distribution Asymmetric (Skewness = -2)		
		Model fit assessed by Chi-square test	Model fit assessed by CFI / TLI / RMSEA	Decision consistency: Chi-square test vs. other fit indices
0 (Baseline)	0.25	Y	Y	CON
	0.50	Y	Y	CON
	0.75	Y	Y	CON
2	0.25	Y	Y	CON
	0.50	Y	Y	CON
	0.75	N	Y	Inconsistent
4	0.25	Y	Y	CON
	0.50	N	Y	Inconsistent
	0.75	N	N	CON
6	0.25	Y	Y	CON
	0.50	N	Y	Inconsistent
	0.75	N	N	CON

Note: NK denotes negatively keyed items. N means that the one-factor model is rejected by the Chi-square test (p < 0.05), or CFI (< 0.90), TLI (< 0.90), and RMSEA (> 0.08); Y denotes that it is supported by the Chi-square test, or the other three fit indices; CON means that different decision rules reach consistent decisions regarding overall model fit.

To compare the performance of the ML and MLR estimators, the consistency of the judgment regarding model fit is evaluated. When utilizing the three fit indices, the judgment is not affected by the estimator. However, when the model fit is judged by the Chi-square test, the conclusion is different in two cases, as presented in Table 20. The results suggest that the decisions on the model fit are consistent in most of the conditions for ML and MLR estimations. Compared with ML, the robust estimator (i.e., MLR) performs better when the observed item response distribution is asymmetric, and there are few or no negatively keyed items. Under such

conditions, CFA with MLR estimation leads to fit indices that support the correct judgment of model fit in more conditions than with ML estimation. The one-factor model is supported when MLR is used in two of the conditions where fit statistics obtained from ML wrongly reject it.

Table 20.

Consistency of the decision on model fit based on the Chi-square test or fit indices (ML vs. MLR)

Number of NK Items	Item Loadings	Observed Item Response Distribution	
		Asymmetric (Skewness = -2)	Symmetric
0 (Baseline)	0.25	CON	CON
	0.50	CON	CON
	0.75	Inconsistent	CON
2	0.25	CON	CON
	0.50	Inconsistent	CON
	0.75	CON	CON
4	0.25	CON	CON
	0.50	CON	CON
	0.75	CON	CON
6	0.25	CON	CON
	0.50	CON	CON
	0.75	CON	CON

Note: NK denotes negatively keyed items; CON means that the statistical decisions regarding model fit are consistent between the two estimators.

Modification indices in cases where model fit is poor (MLR estimation)

When the one-factor model was rejected, the modification indices and the parameters suggested to be freed are recorded in Table 21. Each parameter reported is associated with a significant Chi-square change ($p < .05$) if it is freed. Each column in Table 21 represents a simulated condition where a one-factor model is rejected by the Chi-square test. Within each column, the parameters are listed in the order of their corresponding Chi-square change values, from large to small. When more than five parameters have significant modification indices, only the five with the largest Chi-square change value are listed.

Table 21.

Parameters suggested to be freed by modification indices (MLR estimation)

	Observed Item Response Distribution: Asymmetric (Skewness = -2)				
Number of NK Items	NK = 2	NK = 4		NK = 6	
Factor Loading	L0.75	L0.50	L0.75	L0.50	L0.75
Suggested Parameters	I12 with I11	I12 with I11	I10 with I9	I12 with I7	I5 with I3
		I10 with I9	I12 with I9	I8 with I7	I5 with I2
		I11 with I9	I11 with I9	I5 with I1	I12 with I10
		I12 with I9	I11 with I10	**I7 with I5**	I6 with I3
		I12 with I10	I12 with I11	**I7 with I3**	I3 with I2

Note: Only parameters associated with a significant Chi-square improvement are given; when multiple parameters meet this criterion, only the first five are listed and they are ordered from top to bottom by their corresponding Chi-square change value. Parameters represent correlated error terms between items keyed in different directions are highlighted in bold.

The one-factor model is rejected under five conditions when CFA is conducted with MLR estimation (see Table 19). Modification indices are reported under these conditions. Table 21 presents the parameters suggested to be freed to achieve a better model-data fit. As is observed in Table 18, the suggested parameters are mostly correlations between negatively keyed items. Only under conditions where the number of positively and negatively keyed items is balanced do the suggested parameters, as indicated by large modification indices, include correlations between error terms of positively keyed items and those keyed in different directions. If researchers revise their original model solely based on modification indices with the aim to improve the model fit, it is likely to result in a model with one dominant factor and a factor accounting for the item keying effect. This might, to some extent, explain why in empirical studies with balanced tests, a positive keying effect sometimes emerges in one sample and a negative keying effect emerges in another with the same measure.

One-factor CFA model fit indices with WLSMV

CFA models treating item responses as ordinal and using WLSMV show a good model fit across all the simulated conditions.

The results from Study 2 are summarized below by responding to each of the research questions that were previously raised.

Q2.1: Does the reverse scoring of negatively keyed items affect the assessment of model fit?

Neither reversed scoring the negatively keyed items nor leaving them in their original format leads to any differences in the judgment of the one-factor model fit.

Q2.2: Under what conditions will the one-factor model be rejected either by the Chi-square test or other fit indices?

When item responses are treated as continuous and ML estimation is used in CFA, the one-factor model may be wrongly rejected when the observed item response distribution is asymmetric. Chi-square tests reject the right model under conditions with skewed item response distributions and high communality, even without negatively keyed items. Descriptive fit indices are more robust with asymmetric response distributions, but they still wrongly reject the model under conditions with high communality and a relatively large number of negatively keyed items.

MLR performs similarly to ML in assessing the model fit under most conditions. When judged by fit indices, CFA models with MLR or ML always lead to the same conclusions. When the model fit is judged by the Chi-square test, MLR does slightly better than ML. The two conditions where MLR estimation leads to the correct judgment of model fit while ML estimation does not are those with few negatively keyed items.

When item responses are treated as ordinal and the one-factor model is tested via CFA using WLSMV, both the Chi-square test and other fit indices suggest a good model fit under all the simulated conditions. This implies that having negatively keyed items does not affect the judgment of model fit when the item responses are treated as ordinal in CFA with WLSMV estimation.

Q2.3: When the one-factor model is rejected, which parameters will the modification indices suggest to be freed? Will these additional parameters support an alternative model that accounts for the item keying effect?

Under the conditions where the one-factor model is rejected by fit statistics from CFA, the modification indices are likely to suggest freeing parameters that represent correlations between negatively keyed items. Relying on modification indices to modify the original factor

model tends to lead to factor solutions with one dominant factor and correlated error terms between negatively keyed items.

Section Two: A Reversed Threshold Model for Negatively Keyed Items

The two simulation studies in this section document the effect of negatively keyed items on the decision of the number of factors using exploratory and confirmatory approaches. The simulation model used in this section differs from the one in studies 1 and 2 because the factor loadings in the latent response model are all positive and the latent response thresholds are reversed for negatively keyed items. That is, for positively keyed items, higher values on the latent response distribution correspond to higher values on the response category; however, for negatively keyed ones, higher values on the former correspond to lower values on the latter. Chapter Two describes this psychometric model as a "reversed threshold model for negatively keyed items."

Study 3: The impact of negatively keyed Items on the decision of the number of factors using exploratory approaches

Similarly to Study 1, the research questions this study aims to address are as follows.

Q3.1: Does the reverse scoring of negatively keyed items affect the number of factors pointed by the K-G rule and PA?

Q3.2: Under what conditions will the K-G rule correctly point to the number of factors?

Q3.3: Under what conditions will PA point to the correct number of factors?

Q3.4: Will the two simulation models (Study 1 vs. Study 3) lead to the same conclusions regarding the number of factors under different conditions?

Method

Study design

This study mimics negatively keyed items by reversing the relationship between the latent response distribution and observed response categories. The factors manipulated in the data simulation stage are presented in Table 1. As in studies 1 and 2, three factors are systematically varied and fully crossed. These three factors are: (a) the number or proportion of negatively keyed items, (b) the magnitude of communality in the factor structure, and (c) the

distribution of observed item responses. There are four levels of the number of negatively keyed items (0, 2, 4, 6 items out of 12), three levels of communality (0.06, 0.25, and 0.56), and two levels of the distribution of observed item responses (symmetric and asymmetric distributions; see Table 1). The length of the test (i.e., the total number of items) is fixed at twelve, and the observed responses are on a scale from one to five. In addition, the "true" factor structure of the test is fixed to follow a strict one-factor model. This results in a total of 24 (i.e., $4 \times 3 \times 2$) datasets after the data simulation.

Besides the factors manipulated in the data simulation process, the effect of two scoring methods for negatively keyed items is also investigated. These scoring methods are (a) reverse scoring negatively keyed items, and (b) leaving all the responses in their original format. Each scoring method is applied to the negatively keyed items once the response data are simulated. There are six conditions that contain no negatively keyed items, and that hence do not need these scoring methods. Therefore, a total of 42 (i.e., $4\times3\times2\times2-6$) unique datasets are produced after the data simulation and item response scoring stages.

In this study, the primary outcome variable of interest is the decision of the number of factors. PCA is conducted for each dataset to obtain eigenvalues. The K-G rule and PA are applied to determine the number of factors. Following the routine procedure in applied research, when PA points to more than one factor to retain, follow-up EFA modeling is conducted and evaluated. These EFA models examine whether factors associated with item keying direction will emerge.

Procedures

Although this study and Study 1 differ in their data simulation models, the procedures in both are the same. The major steps in conducting this study are presented in Figure 13. As the flow chart indicates, this study begins with data simulation and proceeds to apply different scoring methods to the simulated item responses. After scoring, the datasets are ready to be analyzed. In the analysis stage, each dataset (i.e., each combination of factors manipulated before data analysis stage) is analyzed through PCA to obtain eigenvalues. The K-G rule and PA are then utilized to determine the number of factors. In cases where PA suggests more than one factor, EFA with different types of rotation is applied to explore the factor structures. The following paragraphs describe the steps that diverge from Study 1.

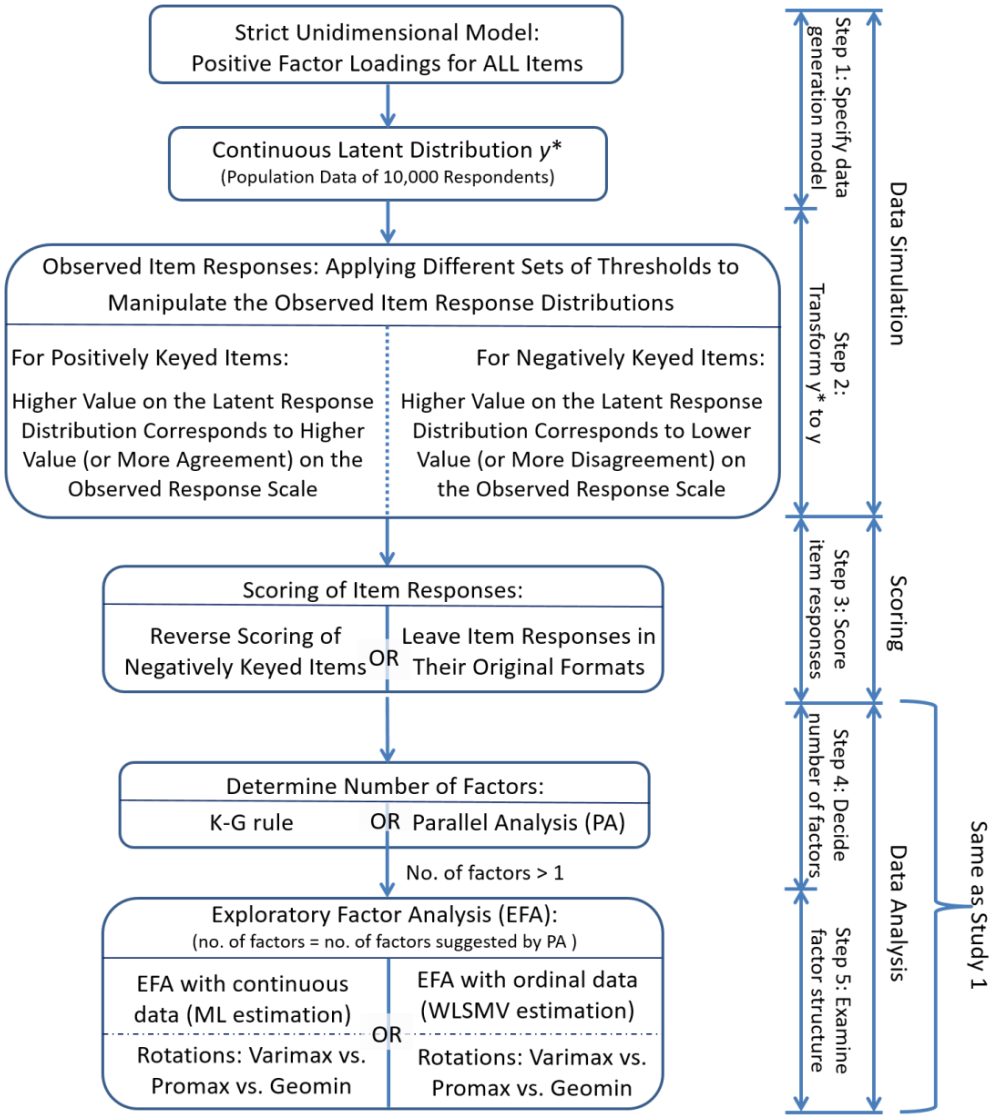


Figure 13.

A flow chart representing the research process of Study 3

Step 1: Specify the data generation model. Observed responses are simulated through latent response *y**. The data were generated using a one-factor model with 12 items (see Figure 11). The continuous *y** distributions for the latent responses follow multivariate normal distributions, and the loadings for all the items are positive. The magnitude of the item loadings is manipulated to represent different levels of communality.

Step 2: Transform continuous latent response distributions to observed responses. The continuous latent response (y^*) is broken down into K ordered response categories through a set of (K-1) thresholds. By manipulating the values of the thresholds, the distribution of the observed response can be changed. As in the two previous studies, the number of response categories is fixed at five. Two sets of four thresholds are used to transform the latent responses (y^*) into observed responses (y), forming two types of distributions: symmetric and skewed, with an absolute skewness value of 2. The correspondence between latent response (y^*) and observed response categories (y) for each distribution condition is presented in Table 22.

Table 22.

Thresholds used in the response transformation

Response categories for positively keyed items	Corresponding y^* values		Response categories for negatively keyed items
	Symmetric	Skewed (|skewness| = 2)	
1	Lowest thru -1.8000	Lowest thru -1.66429	5
2	-1.7999 thru -0.6000	-1.66428 thru -1.27956	4
3	-0.5999 thru 0.6000	-1.27955 thru -1.02406	3
4	0.6001 thru 1.8000	-1.02405 thru -0.68564	2
5	Higher than 1.8000	Higher than -0.68564	1

Step 3: Score the item responses: Same as in Study 1.

Step 4: Decide on the number of factors: Same as in Study 1.

Step 5: Examine the factor structures: Same as in Study 1.

Checking the simulation method: Descriptive statistics from one of the simulated datasets.

To serve as a check on the simulation methodology, the descriptive statistics for each dataset are reviewed before conducting any further analyses. The statistics for one of these datasets are reported to show the method's credibility. This demonstration draws on the dataset with the same constraints as the example dataset used in Study 1. These two sample datasets are the same in terms of item communality level (i.e., $h^2 = 0.56$ or item loading = 0.75), observed item response distribution (i.e., symmetric), and number of negatively keyed items (i.e., six out of twelve). Placing the same constraints on the simulation factors allows us to compare the datasets using different approaches to conceptualizing negatively keyed items. This facilitates the interpretation and discussion of the results from this study in relation to those from Study 1.

Table 23 presents the mean and skewness of the simulated data. Item responses are analyzed in their original format, that is, those to negatively keyed items are not reverse scored. The first column on the left lists item code, while the second column indicates the intended keying direction. In this simulated condition, the last six items are negatively keyed. No difference is observed between positively and negatively keyed items in terms of their means and the values of skewness. Table 24 displays the item correlations, whose directions match the intended keying directions for the items. The correlations between items keyed in the same direction (either positively or negatively keyed) are positive. The correlations between items keyed in different directions (i.e., those between positively and negatively keyed items) are negative, suggesting that a higher response category on one type of item is associated with a lower response category on the other. The absolute values of the inter-item correlations are all similar.

As mentioned in the first study, a negatively keyed item is one whose original response is negatively correlated with the total score of a test, as well as with other positively keyed items. A positively keyed item should correlate positively with the total test score and with other positively keyed-items. Based on the descriptive statistics on the observed item responses, the simulation strategy used in this study produces items with different keying directions.

Table 23.

Mean and skewness of observed item responses (loading = 0.75, symmetric, $N_{NK} = 6$)

Item Id	Keying Direction	Mean	Skewness
I1	PK	2.99	0.00
I2	PK	3.00	0.00
I3	PK	2.99	-0.01
I4	PK	3.00	0.00
I5	PK	2.99	-0.02
I6	PK	3.00	0.03
I7	NK	3.01	0.02
I8	NK	3.00	0.04
I9	NK	3.00	0.04
I10	NK	2.99	0.00
I11	NK	3.01	0.04
I12	NK	3.01	0.00

Note: PK denotes positively keyed items, and NK denotes negatively keyed ones.

Table 24.

Correlation matrix of observed item responses (loading = 0.75, symmetric, $N_{NK} = 6$)

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11
I1	--										
I2	0.516	--									
I3	0.505	0.505	--								
I4	0.502	0.502	0.500	--							
I5	0.508	0.507	0.511	0.498	--						
I6	0.500	0.496	0.496	0.497	0.500	--					
I7	-0.506	-0.509	-0.504	-0.502	-0.502	-0.497	--				
I8	-0.503	-0.514	-0.504	-0.507	-0.496	-0.494	0.509	--			
I9	-0.502	-0.497	-0.502	-0.501	-0.508	-0.488	0.505	0.497	--		
I10	-0.497	-0.510	-0.493	-0.507	-0.505	-0.489	0.491	0.502	0.494	--	
I11	-0.510	-0.510	-0.502	-0.498	-0.497	-0.499	0.494	0.500	0.506	0.502	--
I12	-0.487	-0.503	-0.497	-0.496	-0.491	-0.489	0.504	0.499	0.492	0.493	0.494

Compare the data presented in tables 23 and 24 with that in tables 3 and 4 in Study 1. The mean scores, skewness of the item responses, and inter-item correlations all have similar values. This suggests that when the simulated observed response distribution is symmetric, the two

conceptualizations of negatively keyed items used in Study 1 and Study 3 produce item responses with similar statistical features.

Results and conclusions

Decision on the number of factors

As expected, both scoring methods produce identical eigenvalues after extraction. As a result, the number of factors that the K-G rule or PA identifies is the same for the datasets that differ only in their scoring methods for negatively keyed items. Therefore, the results are presented below without explicitly referring to the scoring methods used.

As in Study 1, PCA is first conducted on a Pearson correlation matrix for each dataset. The K-G rule and PA point to the right number of factors when the observed item response distribution is symmetric. When this observed item response distribution is skewed, however, the number of factors indicated by the K-G rule and PA under each condition is presented in Table 25. The table also shows the consistency of the results of this study with those of Study 1. The first column on the left lists the number of negatively keyed items. Next is a column containing the factor loadings used in the simulation model. The following two columns present the number of factors pointed by the K-G rule and PA. As shown in the table, under most conditions where the observed item response distribution is asymmetric, the number of factors is correctly identified as one. Only under conditions with low communality ($h^2 = 0.06$, or factor loadings equal 0.25) and skewed item response distributions does the K-G rule inflate this number. The last two columns evaluate the consistency of the number of factors indicated by the two methods under different simulation frameworks. As was observed in Study 1, the K-G rule inflates the number of factors under conditions where the observed item responses are asymmetric and the communality level is low (i.e., item loadings equal 0.25). In contrast to Study 1, the K-G rule points to the right number of factors when the item communality level is median or high, even with asymmetric observed item response distributions and different numbers of negatively keyed items.

Table 25.

Number of factors based on a Pearson correlation matrix

Number of NK Items	Item Loadings	Observed Item Response Distribution: Asymmetric (Skewness = -2)			
		Study 3 Results		Results Consistency: Study 3 vs. Study 1	
		K-G rule based on Pearson correlations	PA based on Pearson correlations	K-G rule based on Pearson correlations	PA based on Pearson correlations
0 (Baseline)	0.25	3	1	CON	CON
	0.50	1	1	Inconsistent	CON
	0.75	1	1	CON	CON
2	0.25	3	1	CON	CON
	0.50	1	1	Inconsistent	CON
	0.75	1	1	Inconsistent	Inconsistent
4	0.25	3	1	CON	CON
	0.50	1	1	Inconsistent	CON
	0.75	1	1	Inconsistent	Inconsistent
6	0.25	3	1	CON	CON
	0.50	1	1	Inconsistent	Inconsistent
	0.75	1	1	Inconsistent	Inconsistent

Note: NK denotes negatively keyed items; Inconsistent represents a condition where the number of factors differs between Study 3 and Study 1; CON means that the numbers of factors identified in Study 3 and Study 1 are the same.

Although PA points to the correct number of factors under all the simulation conditions, the K-G rule still shows inflation when PCA is conducted on Pearson correlations while the item communality is low and the item response distribution is asymmetric. As the results from Study 1 suggest, the K-G rule and PA using eigenvalues from PCA on polychoric correlations seem to be more robust to the presence of skewed item response distributions. Table 26 presents the number of factors suggested by the K-G rule and PA with PCA on polychoric correlations. This is to examine whether switching from Pearson to polychoric correlations affects the performance of the K-G rule and PA under the current simulation framework.

When the observed item response distribution is symmetric, the K-G rule and PA always point to the right number of factors. Hence, Table 26 displays the results for only asymmetric distributions. The data show that using Pearson or polychoric correlations in the PCA stage does

not affect the number of factors suggested by the K-G rule and PA. The results align with Study 1, in which eigenvalues are extracted from a polychoric correlation matrix.

Table 26.

Number of factors: Results based on a polychoric correlation matrix

		Observed Item Response Distribution: Asymmetric (|Skewness| = 2)	
Number of NK Items	Item Loadings	K-G rule based on *polychoric* correlations	PA results based on *polychoric* correlations
0 (Baseline)	0.25	3	1
	0.50	1	1
	0.75	1	1
2	0.25	3	1
	0.50	1	1
	0.75	1	1
4	0.25	3	1
	0.50	1	1
	0.75	1	1
6	0.25	3	1
	0.50	1	1
	0.75	1	1

Note: NK denotes negatively keyed items.

As in Study 1, PA outperforms the K-G rule in the overall accuracy of the number of factors it identifies across all the simulated conditions. When negatively keyed items are simulated through negative factor loadings in the latent response model (Study 1), the number of factors suggested by the K-G rule or PA with PCA on Pearson correlations tends to be inflated when the observed item response distributions are asymmetric. However, when negatively keyed items are simulated through a reversed transformation from the latent to the observed responses, both approaches suggest the right number of factors except for conditions with low item communality combined with the use of K-G rule.

Because PA points to one factor under all the simulated conditions in this study, a subsequent EFA is not conducted. In general, when negatively keyed items are simulated through reversing the relationship between the latent and observed responses rather than through

utilizing negative factor loading in the latent response model, the assessment of dimensionality is not as dependent on the data analysis methods.

The study results are summarized by responding to the research questions presented above.

Q3.1: Does the reverse scoring of negatively keyed items affect the number of factors pointed by the K-G rule and PA?

When the K-G rule and PA are used to explore test dimensionality, the reverse scoring of negatively keyed items does not influence the subsequent data analysis or its results.

Q3.2: Under what conditions will the K-G rule correctly point to the number of factors?

When the observed item response distribution is symmetric, the K-G rule always identifies the right number of factors. When it is not, the K-G rule inflates the number of factors when the item communality level is low. This inflation is not associated with the presence of negatively keyed items.

Q3.3: Under what conditions will PA point to the correct number of factors?

PA always points to the right number of factors in this simulation study.

Q3.4: Will the two simulation models (Study 1 vs. Study 3) lead to the same conclusions regarding the number of factors under different conditions?

When the simulated item responses follow a symmetric distribution, the results regarding the number of factors to retain are consistent between Study 1 and Study 3. This suggests that for symmetrically distributed observed item responses, different simulation procedures for negatively keyed items do not make a difference in the exploratory assessment of test dimensionality. When the observed item responses are asymmetrically distributed, however, the situation becomes complicated. Depending on the generation process for negatively keyed items and the method used to determine the number of factors, the conclusion regarding test dimensionality can be different.

Study 4: The impact of negatively keyed items on the model fit in CFA

As in Study 2, the purpose of this study is to investigate the effect of negatively keyed items on the factor structure of a test when CFA is used to assess the model fit. It considers four fit indices: (a) Chi-square statistics, (b) CFI, (c) TLI, and (d) RMSEA with associated confidence

intervals (CIs). Poor fit indices usually lead to the rejection of a model, while acceptable ones lead to its acceptance.

In correspondence with Study 2, this study seeks to answer the following questions:

Q4.1: Does the reverse scoring of negatively keyed items affect the assessment of model fit?

Q4.2: Under what conditions will the Chi-square test or other fit indices reject the one-factor model?

Q4.3: Do different simulation models (Study 2 vs. Study 4) lead to the same judgment regarding model fit?

Method

Study design

The same datasets from Study 3 are used here. In total, 42 unique datasets are simulated, each consisting of twelve items and 10,000 respondents. CFA is applied and the item responses are treated either as continuous or ordinal. ML and MLR estimations are used for continuous item responses, and WLSMV estimation for ordinal ones. The global model fit indices are of primary interest in this study.

Procedures

The flow of this study, as presented in Figure 14, is similar to that of Study 2. The data simulation strategy (i.e., steps 1 to 3) is the same as the one described in Study 3. The data analysis plan (i.e., steps 4 and 5) replicates that in Study 2.

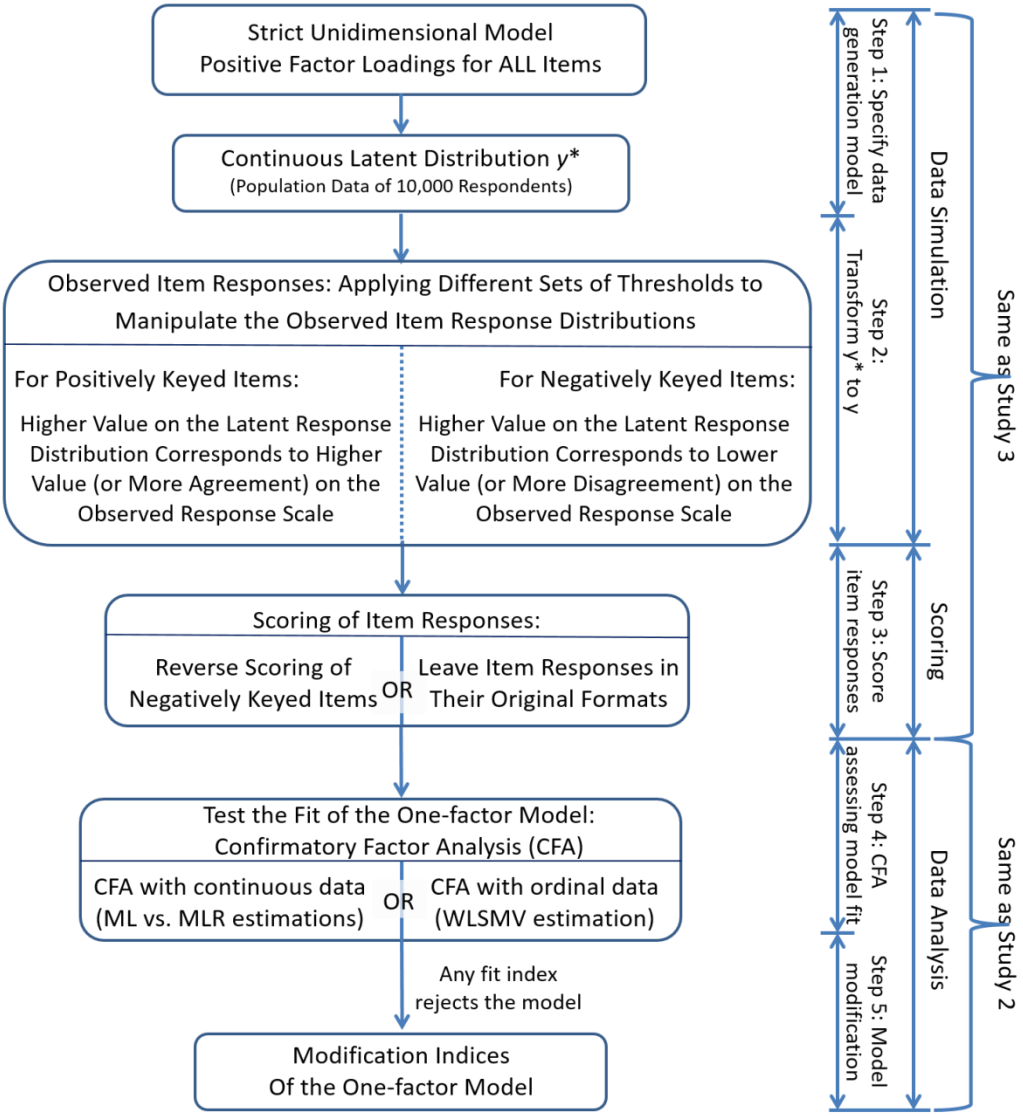


Figure 14.
A flow chart representing the research process of Study 4

Step 1: Specify the data generation model: Same as in Study 3.

Step 2: Transform continuous latent response distributions to observed responses: Same as in Study 3.

Step 3: Score the item responses: Same as in Study 3.

Step 4: Evaluate the fit of the one-factor CFA model: Same as in Study 2.

Step 5: Use modification indices: Same as in Study 2. This step is only conducted for conditions where the one-factor model is rejected.

Results and conclusions

The judgment of model fit does not vary between datasets that differ only in their scoring methods for negatively keyed items. The fit statistics obtained from the reverse scoring of item responses for negatively keyed items are the same as those obtained from analysing them in their original format. In the remainder of this study, the results for datasets using different scoring methods will not be presented separately.

One-factor CFA model fit indices with ML estimation

When item responses are treated as continuous, the one-factor model is tested via CFA using ML estimation. Chi-square tests and the other three fit indices are in agreement, and all suggest a good model fit when observed items responses are symmetric, regardless of the communality levels and the number of negatively keyed items. Hence, Table 27 includes only conditions with asymmetric item response distributions. Since the judgment of model fit is consistent across the three descriptive fit indices, their results are presented together. The results from this study and Study 2 are compared, and their consistency is indicated in the last two columns on the right.

The Chi-square test ($p < 0.05$) rejects the one-factor model under four conditions (see Table 27). It suggests that when the observed item responses are distributed asymmetrically and the level of communality is high ($h^2 = 0.56$ or factor loadings = 0.75), the one-factor model is likely to be rejected by the Chi-square test regardless of whether *any* negatively keyed items are included.

When using CFI, TLI, and RMSEA, the one-factor model is supported under all the simulated conditions. The cut-offs used for these three indices are 0.95, 0.95, and 0.05, respectively. This result is slightly different from what is found in Study 2, where the one-factor model is rejected by the fit indices under conditions with a high communality level ($h^2 = 0.56$), a relatively *large* number of negatively keyed items (4 to 6 negatively keyed items), and asymmetrically distributed observed item responses.

The results also show that conclusions regarding the model fit are inconsistent between the Chi-square tests and the other three fit indices under four conditions. These conditions all have asymmetric observed item response distributions and high communality levels. Under such

conditions, the Chi-square tests reject the one-factor model while CFI, TLI, and RMSEA support it (Table 27).

Table 27.

One-factor CFA model with ML estimation

		Observed Item Response Distribution: Asymmetric (Skewness = -2)			
		Study 4 Results		Results Consistency: Study 4 vs. Study 2	
Number of NK Items	Item Loadings	Model fit assessed by Chi-square test	Model fit assessed by CFI / TLI / RMSEA	Model fit assessed by Chi-square test	Model fit assessed by CFI / TLI / RMSEA
0 (Baseline)	0.25	Y	Y	CON	CON
	0.50	Y	Y	CON	CON
	0.75	N	Y	CON	CON
2	0.25	Y	Y	CON	CON
	0.50	Y	Y	Inconsistent	CON
	0.75	N	Y	CON	CON
4	0.25	Y	Y	CON	CON
	0.50	Y	Y	Inconsistent	CON
	0.75	N	Y	CON	Inconsistent
6	0.25	Y	Y	CON	CON
	0.50	Y	Y	Inconsistent	CON
	0.75	N	Y	CON	Inconsistent

Note: NK denotes negatively keyed items. N means that the one-factor model is rejected by the Chi-square test ($p < 0.05$), or CFI (< 0.90), TLI (< 0.90), and RMSEA (> 0.80); Y denotes that it is supported by the Chi-square test ($p \geq 0.05$) or the other three fit indices. In the last two columns; Inconsistent represents a condition where the results of Study 4 are different from those of Study 2; CON means that the results of Study 4 are consistent with those of Study 2.

In practice, when only the Chi-square test rejects a hypothetical model, it is still likely to be accepted, especially if the sample size is large. Given that the one-factor model is supported by CFI, TLI, and RMSEA, its modification is not considered in this study.

One-factor CFA model fit indices with robust estimation

To explore whether relying on a robust estimator will lead to different decisions on the model fit compared with employing ML estimation, the one-factor model is also tested using

MLR. The model is always supported by the Chi-square test with *p*s > 0.05 and the other three fit indices under all the simulated conditions.

The judgment of model fit when CFA is conducted with ML or MLR is consistent in most conditions. Only when Chi-square test is used to make the model fit judgment, observed item response distribution is asymmetric, and item communality level is high ($h^2 = 0.56$), the model fit judgment is different depending on the estimation method used in CFA (see Table 28).

Table 28. Consistency of the decision on model fit based on Chi-square test (ML vs. MLR)

Number of NK Items	Item Loadings	Observed Item Response Distribution	
		Asymmetric (|Skewness| = -2)	Symmetric
0 (Baseline)	0.25	CON	CON
	0.50	CON	CON
	0.75	Inconsistent	CON
2	0.25	CON	CON
	0.50	CON	CON
	0.75	Inconsistent	CON
4	0.25	CON	CON
	0.50	CON	CON
	0.75	Inconsistent	CON
6	0.25	CON	CON
	0.50	CON	CON
	0.75	Inconsistent	CON

Note: NK denotes negatively keyed items; CON means that the statistical decisions regarding model fit are consistent between the two estimators; Inconsistent denotes the conditions where the decisions are inconsistent.

One-factor CFA model fit indices with WLSMV

CFA models treating item responses as ordinal and using WLSMV estimator showed good model fit across all the simulated conditions.

To summarize the findings of this study, the research questions listed at the beginning of this study are responded to as follows.

Q4.1: Does reverse scoring of negatively keyed items affect the assessment of model fit?

No. Reverse scoring the negatively keyed items or not does not lead to any differences in the final judgment of the one-factor model fit.

Q4.2: Under which conditions will the one-factor model be rejected either by Chi-square test or other fit indices?

Under this simulation framework, the one-factor is supported by descriptive fit indices (i.e., CFI, TLI, and RMSEA) in all the conditions regardless of the estimator used in CFA. Only when item responses are treated as continuous, the one-factor model is tested via ML estimation, the observed item responses are asymmetric, and the item communality level is high, would the Chi-square test wrongly reject the one-factor model. However, this happens independently of the presence of negatively keyed items.

Q4.3: Do different simulation models (Study 2 vs. Study 4) lead to the same judgment regarding the model fit?

Study 2 and Study 4 are different in the simulation strategy used to generate negatively keyed items. They represent different ways to conceptualize a negatively keyed item. When CFA models are conducted with WLSMV estimation, the model fit judgment is consistent between these two studies, and the one-factor model shows good fit under all the simulation conditions. Also, when the observed item response distribution is symmetric, the judgment of model fit is consistent regardless of the simulation strategy, CFA estimation methods, and other factors manipulated in the simulation. However, when CFA with ML or MLR estimation is applied under conditions where the observed item response distribution is asymmetric, the overall judgment of model fit can differ depending on the simulation strategy to negatively keyed items and the method assessing model fit (Chi-square test vs. descriptive fit indices).

CHAPTER FOUR:
DISCUSSION AND RECOMMENDATIONS

This dissertation investigated how including negatively keyed items may affect statistical conclusions about the dimensionality and factor structure of short psychological measures containing Likert-type items. These measures are typically postulated to be unidimensional and hence report an overall test score. The dissertation conducted four simulation studies both to inform psychometric theories and provide recommendations to researchers in their day-to-day practice. This chapter discusses the findings from the simulations in the context of revisiting the research questions listed in Chapter Two. It then offers guidelines for researchers based on these findings. Next, the novel contributions of this dissertation are described. The chapter closes with a discussion of the future research directions that arise from the limitations of this research.

Revisiting the Research Questions

Self-report Likert-type tests are commonly used in educational and psychological research, and these measures frequently contain both positively and negatively keyed items. A major assumption underlying this practice is that the negatively keyed items will function in the same manner as their positively keyed counterparts (Marsh, 1996). In other words, regardless of the keying direction, items intended to measure the same construct should be psychometrically comparable. An individual's score on a typical Likert-type test is a sum, an average, or another similar composite derived from his or her responses to all the items—i.e., the overall test score. Hence, the scoring of item responses rests on the implicit supposition that the measure is unidimensional (i.e., contains one latent factor).

Unfortunately, empirical evidence calls this assumption into question for several different measurement instruments and populations (e.g., Barnette, 2000; Lai, 1994; Marsh, 1986; Motl et al., 2000; Pilotte & Gable, 1990; Schriesheim & Hill, 1981). This has serious implications, since the scoring of a test and, more importantly, the validity of inferences based on test scores depend on its factor structure. The emergence of additional factor(s), which are essentially defined by item keying direction, raises concerns about the credibility of interpretations based on the overall test score. Moreover, the incorrect identification of a test's dimensionality or factor structure

may obscure the understanding of the construct and lead to misinterpretation. Finally, any subsequent statistical analysis that relies on the overall test score is also adversely affected.

To understand the mechanisms that drive the emergence of additional factors in tests with mixed-keyed items, a large number of empirical data-based research and three simulation studies have been reported in the research literature. It has been argued that including negatively keyed items in a test tends to, but does not always, introduce a systematic method effect (e.g., Finney, 2001; Marsh, 1986; Motl et al., 2000; Pilotte & Gable, 1990; Schmitt & Stults, 1985; Schriesheim & Hill, 1981; Spector et al., 1997; Woods, 2006). These mixed findings may be due to the unclear terminology and conceptualization used for keying and wording effects, as well as to their close interconnection. Most of the empirical studies have paid attention only to the item wording effect, in particular, to negatively worded items. This may be attributed to the avoidance of using items that are negatively worded but positively keyed (e.g., an item like "I am not sad" in a test measuring happiness). It should be noted, however, that the negative wording effect is often nested within the keying effect. That is, negatively worded items are often negatively keyed, but negatively keyed items are not necessarily negatively worded.

If the item keying effect cannot be ruled out, the method effect that is often ascribed to item wording can also be attributed to item keying or a combination of these factors. Without investigating and documenting the keying effect, we can neither get a full picture of the issues related to the factor structure of tests with mixed-keyed items nor clearly understand the effect of item wording. To extend the previous research on negatively keyed items and their impact on the assessment of test dimensionality, I systematically investigated this issue via four inter-related computer simulation studies in this dissertation.

Item response data are generated through threshold models to represent responses collected on Likert-type tests. To guide the simulation procedures, two ways to conceptualize the responses to negatively keyed items are used. One is to conceive of them as items whose latent response functions (y^*) are negatively correlated with the construct of interest. In this case, these item responses are attributed to a latent response distribution that is negatively correlated with the construct. This psychometric model is called the "negative factor loading model for negatively keyed items" and is investigated in Section I of Chapter Three (studies 1 and 2). To better understand the study results, we can relate this psychometric framework to items that are negatively keyed but not necessarily negatively worded. These items measure the polar opposite

of the target construct, and mainly fall into the category of positively worded but negatively keyed items (see Figure 2 in Chapter Two).

The second way to conceptualize the responses to negatively keyed items is to view them as deriving from a reversed function from the latent response distribution to observed responses. In this case, the relationship between the latent responses and the construct is the same for items keyed in both directions. However, the relationship between thresholds and their corresponding response categories is reversed for negatively keyed items. This psychometric model is called the "reversed thresholds model for negatively keyed items" and is investigated in Section II of Chapter Three (studies 3 and 4). To connect this psychometric framework to daily practice, we can see it as presenting responses to negatively worded and negatively keyed items. Typical examples are items containing negation markers, such as "not," "no," and "never."

The connections made between the two psychometric models and the two types of negatively keyed items are meant to facilitate the understanding of study results and promote further conversations. According to the research on item response process, individuals use various verbal, visual and contextual cues to answer an item (e.g., Schwarz, Strack, Hippler, & Bishop, 1991; Tourangeau, Couper, & Conrad, 2007, 2013). Item features, such as wording directions, and response scale features, such as descriptors attached to each response category (e.g., Cabooter, Weijters, Geuens, & Vermeir, 2016) are some examples of such cues. It is reasonable to postulate that the two types of negatively keyed items that are different in their wording direction may prompt different response processes and are better represented by different psychometric models. However, it is worth noting that, when they were first proposed in the research literature, the item factor models were not (response) process models and were not necessarily designed to describe different response processes. Rather, they were introduced as a statistical estimation method to reproduce the covariance structure. As mentioned in Chapter Two, the directionality of this argument is important in understanding the relationship between these psychometric models and the two types of negatively keyed items. As was stated earlier, psychometric modeling is an interplay of theory, model, and data, and therefore, while the psychometric models may imply a type of item responding, the item responses (on their own) do not necessarily imply a psychometric model exclusively.

A comparison of results based on these two psychometric frameworks (i.e., Section I vs. Section II in Chapter Three) shows that the conceptualization and simulation strategy employed

for negatively keyed items can affect the statistical assessment of test dimensionality and factor structure. Generally speaking, when negatively keyed items are simulated through reversing the relationship between latent response distributions and observed responses (Section II), their presence does not affect the statistical judgment of test dimensionality. When item responses are simulated using the first strategy, however, the negative keying direction comes from negative loadings in the latent response model (Section I). To assess test dimensionality in such cases, the item responses are better treated as ordinal rather than continuous. When the observed item response distribution is skewed and item responses are treated continuously, the presence of negatively keyed items can inflate the number of factors or suggest an incorrect model. Using PA with polychoric correlations in PCA or using estimators that treat item responses as ordinal in factor analysis can prevent the misidentification of the factor structure.

Another observation is that the reverse scoring of negatively keyed items does not change the statistical judgment of test dimensionality. It is conventional wisdom to reverse score the negatively keyed items before analyzing the data, but this dissertation suggests that it is not always necessary. When the analyses are conducted using the EFA or CFA methods reported in the simulation studies, whether or not the negatively keyed items are reverse scored does not affect the judgment. When factor analysis techniques are applied to tests with negatively keyed items, the different keying directions of items are reflected by the signs of the factor loadings. Moreover, under conditions where negatively keyed items do affect the statistical decisions about the test dimensionality, reverse scoring these items does not solve the problem.

Two of the simulation studies (studies 1 and 3) focus on assessing dimensionality through exploratory methods. They utilize two PCA-based approaches, the K-G rule and PA, to determine the number of factors to retain. Both methods point to the right number of factors when the observed item response distributions are symmetric. However, when the observed item response distributions are asymmetric, the K-G rule tends to show an inflated number of factors even without negatively keyed items. The finding that PA outperforms the K-G rule in many conditions is consistent with previous studies (Hakstian et al., 1982; Zwick & Velicer, 1982, 1986). The results from studies 1 and 3 suggest that researchers should prefer PA using eigenvalues based on polychoric correlations when deciding on the number of factors. Of special note is that a polychoric correlation matrix should be used in calculating eigenvalues, especially

when the response distributions are asymmetric, the item loadings are relatively high, and the number of negatively keyed items is large (close to 50% of the total).

In cases where an inflated number of factors is proposed, PA or the K-G rule usually suggests a two-factor structure. If EFA with ML estimation is conducted as a follow-up to explore the factor structure, a two-factor model will be supported by the good model fit. The two factors in the model are likely to be defined by the item keying direction. The three rotation methods investigated in Study 1, Varimax, Promax, and Geomin, all lead to similar factor structures. The emergence of a second factor defined by negatively keyed items occurs even with a small number of these items (i.e., as few as two). Although the "true" structure of the data under investigation is always a one-factor model, EFA with ML estimation wrongly points to two-factor solutions, which also seem to be supported by model fit statistics, acceptable item loadings, and interpretable factor structures. This implies that researchers should be cautious about making substantive interpretations of EFA results with tests having negatively keyed items. Because it may lead to an over-extraction of the number of factors. This over-extraction may result in researchers trying to substantively interpret statistical artifact.

When a test's dimensionality or factor structure is examined through a confirmatory approach (studies 2 and 4), the results suggest that it is beneficial to conduct CFA using an estimator that treats item responses as ordinal (e.g., WLSMV) rather than one that treats them as continuous (e.g., ML or MLR). When the observed item responses are asymmetric and CFA is conducted with continuous item responses, the fit statistics may wrongly reject the one-factor model. The results also show that the practice of revising models based on modification indices to improve the fit can be misleading. As Study 2 indicates, modification indices tend to suggest adding correlations between the error terms of negatively keyed items. In this case, the better-fitting model is actually the wrong one. Usually, researchers do not change their hypothetical models solely based on one index, but it is common for them to use statistics to inform their decisions about model modification and selection. To make matters worse, the method effects of "negative keying" or "negative wording" have been widely reported in the literature, making it easier for researchers to accept their existence in the tests they use. Although fit indices are generally considered to be useful in assessing the overall model fit, it is worth noting that they are inadequate to guard against invalid models. Indeed, models that such indices deem to be well fitting can still have some poorly fitting parts (Reisinger & Mavondo, 2006; Tomarken & Waller,

2003). In more serious cases, fit indices can wrongly reject an acceptable model (Marsh, Han, & Wen, 2004).

These findings suggest that the observed item response distribution is the driving force behind the statistical decision of test unidimensionality. When the observed item responses are symmetric and normal-like, the unidimensional data structure can be correctly identified under most of the conditions. That is, with a symmetric item response distribution, none of the other factors, including the number of negatively keyed items, the psychometric models used, the reverse scoring of negatively keyed items, and the methods assessing dimensionality, affect the statistical judgment of a unidimensional test. When the observed item responses are asymmetric, the statistical methods employed play an important role in the correct identification of test unidimensionality. Methods that treat item responses as ordinal outperform those that treat them as continuous. It is worth noting that methods treating item responses as ordinal are consistent with the data assumptions made in the simulation models; hence, it is not surprising that categorical methodology is more likely to identify the right data structure in these cases.

Let us now turn to the research questions stated near the end of Chapter Two.

Q1: Do the different psychometric models (i.e., simulation models) of negatively keyed items affect the statistical judgment of test dimensionality under different conditions?

It depends on the shape of the observed item response distribution. When these distributions are symmetric, data generated from the different psychometric models of negatively keyed items result in the same statistical conclusions of test dimensionality. When the observed item response distributions are asymmetric, however, the situation becomes complicated. Depending on the way negatively keyed items are generated, the simulation conditions (i.e., the number of negatively keyed items, item communality levels, and the distribution of observed item responses), and the analytical methods used to assess the dimensionality, the conclusions regarding test dimensionality can differ. For example, when CFA with ML estimation is applied when the number of negatively keyed items is relatively large (i.e., four or six in a twelve-item test), the item communality level is high (factor loading = 0.75), and the observed item response distribution is asymmetric, fit statistics reject the one-factor model if the negatively keyed items are simulated through negative factor loadings in the simulation model. If the negatively keyed items are simulated through the reversed threshold model, these fit indices will support the one-

factor solution when the same methods are applied to the datasets under the same simulation conditions.

Q2: Does the reverse scoring of negatively keyed items affect the statistical judgment of test dimensionality?

No, the reverse scoring of negatively keyed items has no impact on the statistical judgment of test dimensionality.

Q3: Under what conditions will the K-G rule or PA point to the correct number of factors?

When the observed item response distributions are symmetric, both these methods always suggest the proper number of factors. When these distributions are asymmetric, however, the accuracy of the K-G rule depends on the particular psychometric models used to simulate (generate) the data, the simulation conditions, and the type of correlation matrix used in eigenvalue calculation. When negatively keyed items are simulated through negative factor loadings in the latent response model, the K-G rule applied to eigenvalues obtained from PCA based on Pearson correlations tends to inflate the number of factors. The K-G rule applied to eigenvalues obtained from PCA based on polychoric correlations is more robust in the presence of skewed response distributions. However, it still inflates the number of factors when the communality level is low, even without negatively keyed items.

In general, PA performs better than the K-G rule when the observed item response distribution is asymmetric, pointing to the right number of factors in more conditions (i.e., simulation and analysis conditions). Like the K-G rule, the accuracy of PA depends on the psychometric models that generated the data, the simulation conditions, and the type of correlation matrix used to calculate eigenvalues. Indeed, PA points to the right number of factors under all the conditions when negatively keyed items are simulated through a reversed relationship between the latent response and observed response categories. When these items are simulated through negative factor loading in the latent response model, PA still indicates the right number of factors under most of the simulated conditions. Only under conditions where the observed item response distribution is asymmetric, the communality level is high, the number of negatively keyed items is relatively large, and the Pearson correlation matrix is used at the PCA stage does PA overestimate the number of factors.

Q4: When EFA is conducted with an inflated number of factors to retain, what will the factor structure look like? Will factors emerge according to the keying directions of items?

When PA points to an inflated number of factors, it always proposes two. Conducting EFA with two factors using ML estimation results in a two-factor model, with each factor essentially being defined by item keying direction.

Q5: Under what conditions will the one-factor model be rejected either by the Chi-square test or other fit indices?

Only when negatively keyed items are simulated through negative factor loadings and the observed item responses are asymmetric does the estimation method matter. When item responses are treated as continuous and ML estimation is used in CFA modeling, the one-factor model may be wrongly rejected when the observed item response distribution is asymmetric. The Chi-square test rejects the true model under conditions with skewed item response distributions and high communality, even without negatively keyed items. Descriptive fit indices are more robust in the presence of asymmetric response distributions, but they still wrongly reject the model under conditions with high communality and a relatively large number of negatively keyed items. The performance of MLR is similar to that of ML regarding the judgment of model fit under most conditions. When the model fit is judged by fit indices, CFA models with MLR or ML always lead to the same scientific conclusions (rather than statistical results). When the model fit is judged by the Chi-square test, MLR estimation performs slightly better than ML estimation. The two conditions where MLR leads to the correct judgment of model fit while ML does not both have few negatively keyed items. When item responses are treated as ordinal and the one-factor model is tested via CFA using WLSMV estimation, both the Chi-square test and other fit indices suggest a good fit under all the simulated conditions.

When the observed item response distribution is symmetric, the judgment of model fit is consistent regardless of the simulation strategy, CFA estimation methods, or other factors manipulated.

When the observed item response distribution is asymmetric and CFA models are employed with WLSMV estimation, the judgment of model fit is still consistent between the two psychometric models of negatively keyed items. The one-factor model shows a good fit to the data under all the simulation conditions. However, when CFA with ML or MLR estimation is applied to the asymmetric observed item response distributions, the overall judgment of model fit can differ depending on the simulation strategy for negatively keyed items and the method for assessing model fit (Chi-square test vs. descriptive fit indices).

Q6: When the one-factor model is not supported by fit statistics in CFA, what are the consequences of modifying the model using modification indices?

When the Chi-square test or other fit indices, including CFI, TLI, and RMSEA, reject the one-factor model, modification indices can show the parameters to free to achieve a better model fit. It appears that most suggested parameters associated with large Chi-square changes (as indicated by large modification indices) are correlations between negatively keyed items. Revisions solely based on modification indices tend to produce a one-factor model with correlated error terms between negatively keyed items, although the true measurement model is a one-factor model without correlated uniqueness terms.

Taken together, the findings highlight the benefit of using categorical data analytic techniques when assessing the dimensionality of Likert-type tests, especially when the items are keyed in different directions and the observed response data are skewed. When the observed item response distribution is symmetric, continuous methodology performs as well as the categorical estimator. Also, the four simulation studies show that more than one psychometric model can be used to generate responses to negatively keyed items, although they may lead to different statistical conclusions. These results call attention to the assumptions of different models (both for data analysis and for data generation) used in the validation process, because their selection has implications for the eventual validation results. Understanding a model's assumptions is necessary for a comprehensive discussion of its validity (Zumbo, 2007, 2017).

Guidelines and Implications for Researchers

An objective of this study was to provide suggestions and guidelines for researchers when interpreting the statistical results of measures that consist of items keyed in different directions. The following recommendations are based mainly on the simulation results outlined in Chapter Three. They are preliminary in that further research is needed to investigate how they may work under other simulated conditions, as well as in empirical studies. The guidelines provide recommendations based on certain conditions (i.e., the values in the simulation studies), which researchers must compare to their own data before use.

In summary, negatively keyed items are not inherently flawed and including them in a measure does not automatically produce complications in the factor structure of the test. However, skewed item response distributions and inappropriate data analysis methods may lead

to a poor statistical judgment of the test's dimensionality and a subsequent misinterpretation of the factor structure. The observed item distributions can be skewed for different reasons. For example, extreme items tend to lead to skewed response distributions (e.g., the ceiling and the flooring effect), as can response scales with few rating points. Also, the construct can be a truly skewed phenomenon in a population because of the population's characteristics and/or the construct itself. When the response distributions are skewed, researchers must be cautious when choosing a data analysis method to assess test dimensionality.

For applied researchers who collect data using mixed-keyed tests with five-point Likert-type response scales, the following are nine suggestions that are drawn from the study results. These guidelines are not black and white, but are intended to serve as advice in the decision-making process of dimensionality assessment.

(a) During the test administration, try to ensure that respondents understand the items correctly and interpret the response scale consistently. Either careless or acquiescence responding can lead to the emergence of additional factors in the dimensionality assessment.

(b) The scoring of a test contains implied assumptions about its dimensionality. When computing it as a total, an average, or a factor score of all the items, researchers assume that the test is unidimensional. Evidence from dimensionality assessment must be provided to support the scoring of the test and the use of the test score.

(c) When item responses follow the assumption of threshold models—that is, for each item, the relationship between the observed responses and the construct is monotonic—the conventional reverse scoring of negatively keyed items is reasonable.

(d) The reverse scoring of negatively keyed items is unnecessary in the assessment of test dimensionality or factor structure through factor analysis.

(e) When using EFA to explore the dimensionality of a test with PCA-based pointers to the number of factors, it is advised to employ polychoric correlations to obtain the eigenvalues. Also, PA performs better than the K-G rule when the observed response distributions are asymmetric.

(f) When determining the factor structure using EFA or CFA, it is better to choose estimators that treat observed item responses as categorical (e.g., WLSMV with a

polychoric correlation matrix) than ones that treat them as continuous (e.g., ML or MLR with a Pearson matrix).

(g) When using CFA to confirm the factor structure, fit indices can wrongly reject a correct model when the observed item responses are skewed and continuous data analytic methods are used.

(h) Revising the initial model based on modification indices in CFA should be done carefully. In such cases, a more appropriate approach is to run an EFA or an exploratory structural equation model (ESEM).

(i) When assessing dimensionality for validation purposes, comparing competing models by overall model fit does not strengthen the validity argument. A well-fitting model can still be wrong. Instead of comparing the fit of numerous plausible measurement models for a test, it might be more helpful to consider the question of the factor structure in a larger conceptual framework with a matrix of variables and their hypothesized relationships (e.g., a multitrait-multimethod matrix, or MTMM).

For researchers who are interested in studying the item wording or keying effect, the main conclusions drawn from this dissertation are:

(a) It is important to properly differentiate among and describe item features, such as wording, keying, and social-psychological meaning. Using terms ambiguously or interchangeably makes the results difficult to interpret, compare, and synthesize.

(b) Researchers may conceptualize the responses to negatively keyed items differently and test takers may employ various psychological procedures to answer them. The difference in the conceptualization and response process of negatively keyed items can be reflected by different mathematical models. The conclusions regarding the "method effect" associated with item keying can therefore diverge depending on the assumptions made in the modeling process.

(c) Model fit indices alone are insufficient to determine if the presence of the "method effect" associated with item keying or wording is meaningful or an artifact.

Novel Contributions

Likert-type tests are widely used for data collection in the social, behavioural, and health sciences. Hence, most researchers in these fields need to work with data collected through such tests in their daily research practice. It is therefore important that they are made aware of the possible impact of negatively keyed items on the analysis and interpretation of test factor structure, especially considering the widespread use in Likert-type tests. However, there has been virtually no clear guidance or systematic discussion on the strategies to handle responses from mixed-keyed tests, and how the various methods perform under different conditions. As a first step in filling this gap, this dissertation has made three novel contributions to understanding how negatively keyed items may affect the statistical conclusions regarding test dimensionality.

To my knowledge, this dissertation is the first to systematically document the impact of negatively keyed items from different psychometric models on the statistical judgment of test dimensionality. Test dimensionality supports test scoring strategies, as well as test score interpretation. It is one of the most regularly reported pieces of validity evidence to support test score use (Zumbo & Chan, 2014). Given the popularity of mixed-keyed Likert-type tests, and the crucial role of test dimensionality, the additional factors formed by item keying direction have attracted much attention and generated heated discussions regarding their interpretation. With an eye towards communicating with social and behavioural researchers, this dissertation shows practitioners, psychometricians, and methodologists how having negatively keyed items can influence the statistical decision about the factor structure in a factor analysis framework.

Unlike most previous empirical studies that are observational in nature, or the three very limited simulation studies noted in Chapter Two, this dissertation used four computer simulation experiments to investigate the impact of negatively keyed items on the assessment of test dimensionality. A relatively large number of factors were considered in the research design. These design factors, including data generation models for negatively keyed items, test characteristics (e.g., different observed response distributions, numbers of negatively keyed items, and item communality levels), scoring methods for negatively keyed items, and statistical methods to assess dimensionality, cover a wide range of factors that may affect the statistical judgement of test dimensionality. Prior to this research, many of these factors had never been investigated in the context of mixed-keyed tests.

The second contribution is that this dissertation made two important distinctions that may help open future research directions in understanding the factor structure of mixed-keyed tests. Firstly, it distinguished negatively keyed items from negatively worded items. This separation is important, as respondents may engage with items that researchers view as negative in various ways, leading to different response patterns. In turn, these patterns may have different effects on the results from different statistical methods. Indeed, although many empirical studies investigated the impact of negatively worded items on test dimensionality, few paid attention to the item keying effect.

This dissertation further classified negatively keyed items into two types, which it then examined separately. To my knowledge, this is the first simulation study that has described and deliberately distinguished between two psychometric models for negatively keyed items. By making this distinction, this dissertation attempted to caution future researchers about the impact of the simulation strategies they choose for these items on the interpretation of their results. As described in Chapter Two, one operationalization of negatively keyed items is that their item responses (without reverse scoring) correlate negatively with the responses to positively keyed ones, and positively with the total test score. The expected data pattern described in this operational definition of negatively keyed items can be achieved by both the negative factor loading model and the reversed threshold model. If we focus only on the outcome (i.e., the generated item responses), we will miss the potential difference implied by these two psychometric models.

The final contribution of this dissertation is that it offered advice to applied researchers on how to assess dimensionality using data collected through mixed-keyed tests. It is also useful for those who are interested in studying negatively keyed or negatively worded items. One advantage of conducting computer simulation studies is that they enable us to systematically manipulate specific conditions of interest (e.g., the number of negatively keyed items). Although they apply to idealized situations, the results suggest some general guidelines. In practice, the true factor structure of a test is almost always unknown to the researchers who are working with empirical data. Without knowing the "true" structure of the data in the population, the accuracy of the statistical methods and decision rules cannot be evaluated. Thus, researchers need to rely on results from simulation studies to more clearly understand how the item and test features, the

resulting item response distributions, as well as their chosen statistical methods may affect the results they have observed in empirical studies.

This dissertation also included some follow-up steps (e.g., EFA with two factors, and CFA with modification indices) in each of the studies to mimic what researchers usually do when the suggested number of factors is more than one, or the prior measurement model is rejected by fit statistics. By doing so, it sought to provide some insight into the potential reasons for some of the reported findings in the literature. The results from the follow-up steps showed how the presence of negatively keyed items, together with inappropriate analysis methods, may distort the judgment and interpretation of a test's factor structure.

In summary, by (a) conducting simulation experiments to disentangle the keying effect from others, (b) making distinctions between two psychometric models for negatively keyed items, and (c) providing guidelines for other researchers based on the findings, this dissertation fills an important gap in the research literature and contributes to the critical issue of better understanding negatively keyed items and their impact on the assessment of test dimensionality.

Future Directions

Responding to items is a complicated process that can be affected by many factors, including item features, respondent characteristics, administration mode, and other social and cultural variables. Because they are highly inter-related, separating these factors and studying their effects in isolation using empirical data is difficult. As an attempt to detach the item keying effect from the wording effect and to account for the responding process, the four studies reported here were performed using computer simulation. Admittedly, despite its advantages, this method may oversimplify the phenomenon; however, the priority was placed on this experimental setting for its capacity to isolate causal effects. As mentioned in Chapter One, these four inter-related studies conducted in this dissertation aim to create a baseline for future work in this area. Their simulated conditions were somewhat idealized to help with the interpretation. A limited number of influential factors that are believed to have some impact on the statistical judgment of test dimensionality and factor structure were investigated. In essence, the studies reported in this dissertation are controlled experiments in which a few factors were manipulated and studied. Some other variables that may influence the statistical conclusions on factor structure were not considered, such as the sample size, the number of points on the rating

response scale, sub-populations in the respondent population, and the degree of misfit between the model and the data at the population level. In light of the limitations of this dissertation, future research directions are discussed below.

Measurement model

In the current study, the simulation factors, including item communality levels and observed item response distributions, are simulated in idealized conditions. That is, they are kept consistent across all the items. This is only possible when all the items in a test are equivalent, which seldom happens in practice. In reality, items rarely have an equal magnitude in their loadings on the factor(s) and the same thresholds on their rating response scale. The population-level "true" model is assumed to follow a strict unidimensional structure. This refers to a type of structure that has one dominant factor without secondary minor factors. While this situation meets the assumptions of classical testing theory (CTT), several other scenarios may happen in practice. For example, items are likely to have different factor loadings, thresholds may be different for different items, and the observed response distribution may differ across items. Moreover, some of the items in the measure may load on one minor factor (e.g., method effect).

This dissertation focused on unidimensional tests. This is largely because, of the studies that have investigated issues associated with negatively keyed items, most have examined measures that are purportedly unidimensional. Also, a typical Likert-type or summative test assumes a unidimensional structure as indicated by the scoring practice which uses a total or an average of all the items. However, analyzing unidimensional tests makes it impossible to observe the deflation of the number of factors that may result from different statistical methods. When multidimensional tests are investigated, the measurement model can get complicated, with cross-loadings potentially occurring (e.g., one item loading on more than one factor). Based on the findings of this dissertation, future research may begin investigating more of these complexities in the measurement model.

Psychometric models of negatively keyed items

Two psychometric models were used to help conceptualize and simulate the responses to negatively keyed items. It is important to note that the two psychometric response models (negative factor loading and reversed thresholds models) are mathematical-statistical models of

item responding. As Zumbo (2017) states, such models have a long and fruitful history in psychometrics and statistics. It should be noted, however, that these models are representations of idealized item responding and that the actual process may be more complex and varied depending on item features, respondent characteristics, and the context of the test. To the best of my knowledge, there is no clear evidence from the investigation of the actual cognitive process of item responding to clearly confirm or rule out either of these possible psychometric models. Also, no direct evidence has been found to support the connections made between the psychometric framework and the item types (i.e., wording and/or keying). One may argue for different explanations of these two psychometric frameworks, and the ones presented here are just examples.

In this dissertation, threshold models are used to generate observed item responses. It assumes that all the observed responses can be explained by a latent response model. That is, the observed response to an item can be fully attributed to a continuous latent response distribution underlying the item. Thresholds are applied to transform continuous latent responses into observed ordinal responses. Other theoretical frameworks, such as the ideal point model (Cliff et al., 1988; Thurstone, 1928), have been proposed in the literature to explain the responding process. The study results based on the threshold model may thus not be generalizable to response data simulated from other theoretical models.

Respondent population

The simulations conducted in this dissertation assumed that all the respondents understood and responded to the items in the same way. In other words, there are no sub-groups of respondents that respond to the items in different ways. In reality, participants in a study usually come from diverse backgrounds. It is possible that the respondent population is heterogeneous, and that subpopulations may differ in their responding process to negatively keyed items. For example, two psychometric models of negatively keyed items were investigated separately in this dissertation. However, these psychometric models may not represent the difference in negatively keyed items, but the difference in how respondents may react to an item. In this case, some of the respondents may employ an item responding process that is more similar to one of the psychometric models, while others may employ one that better suits a different psychometric framework.

Other factors can give rise to subpopulations. Previous studies show that if a small group of test takers is careless and misresponds to the negatively worded and keyed items, an extra factor may appear in the factor analysis (Schmitt & Stults, 1985; Woods, 2006). Future studies may explore how different types of subpopulations may interact with the presence of negatively keyed items.

Sample size

Only population-level data were simulated and analyzed in this dissertation, since it sought to document the statistical conclusions on the factor structure in the presence of negatively keyed items. The current study was conducted on a population data of 10,000 simulated respondents—that is, a population analogue. Although it is possible to obtain such a large sample size with extensive testing, such as the Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS), it would be in many day-to-day researchers' interest to explore the effect of smaller sample sizes on the statistical judgment of test dimensionality and factor structure. The results based on population-level data may not be replicated, especially in cases where the sample size is small.

The impact of negatively keyed items on other statistical and psychometric procedures

Besides the most prominent issues, others, such as how misidentifying the factor structure of tests with negatively keyed items might bias the subsequent analysis, are worth investigating. As discussed at the beginning of this dissertation, the assessment of test dimensionality and factor structure usually serves as a fundamental step in test validation by supporting its scoring. Test-level scores are often used for research purposes and to make various decisions. When including negatively keyed items poses challenges to the statistical judgment of test dimensionality, it calls into question the accuracy and appropriateness of the test score. In such cases, it may also affect other methods that rely on this data, such as t-tests and regression analyses. We may ask two questions that go in opposite directions: (a) When an incorrect but better-fitted measurement model with the "method effect" is selected to represent the factor structure of a test with negatively keyed items, what is the impact of this mis-identification on the subsequent analysis? and (b) What are the consequences when method effects exist among some items but are ignored (e.g., Gu et al., in press)?

In conclusion, more research is needed to understand negatively keyed items and their impact on the results of psychometric and statistical analysis. With the limitations of this dissertation in mind, future research can be designed to focus on two broad directions. On the one hand, further simulation studies are needed to investigate more factors, so we can fully comprehend the effect of negatively keyed items on the assessment of test dimensionality. On the other, empirical studies should be undertaken to confirm or disconfirm the connections between the psychometric models of item response and the processes in which respondents actually engage. Although many mathematical-statistical models have been used to generate and analyze item response data, far less is understood about how these models can contribute to elucidating the item response process. Indeed, these models are developed to capture or reproduce the data pattern of the observed item responses, rather than to replicate how participants produced these data. However, we should recognize that focusing on the measurement outcome (i.e., response data) and ignoring the response process narrows our understanding of the phenomenon we wish to study.

References

Ahlawat, K. S. (1985). On the negative valence items in self-report measures. *The Journal of general psychology, 112*(1), 89-99.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Armstrong, J. S., & Soelberg, P. (1968). On interpretation of factor analysis. *Psychological Bulletin, 70*, 361-364.

Aron, A., Coups, E. J., & Aron, E. N. (2013). *Statistics for psychology* (6th Ed.). Upper Saddle River, NJ: Pearson.

Asparouhov, T. & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*, 397-438.

Bachman, J. G., & O'Malley, P. M. (1986). Self-concepts, self-esteem, and educational experiences: The frog pond revisited (again). *Journal of Personality and Social Psychology, 50*, 35-46.

Bagozzi, R. P. (1993). An examination of the psychometric properties of measures of negative affect in the PANAS-X scales. *Journal of Personality and Social Psychology, 65*, 836-851.

Barnette, J. J. (1997). Effects of items and response set reversals on survey statistics. *Paper presented at the Annual Meeting of the American Educational Research Association*, Chicago, IL. March, 24-28.

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement, 60*, 361-370.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13*(2), 186-203.

Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary school children. *Journal of Educational Measurement, 22*(3), 231-240.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246.

Bentler, P. M. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences, 42,* 825-829.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588-606.

Bentler, P. M., Jackson, D. N., & Messick, S. (1971). Identification of content and style: a two-dimensional interpretation of acquiescence. *Psychological Bulletin, 76*(3), 186-204.

Bieling, P. J., Antony, M. M., & Swinson, R. P. (1998). The State-Trait Anxiety Inventory, Trait version: Structure and content re-examined. *Behaviour Research and Therapy, 36,* 777-788. doi:10.1016/S0005-7967(98)00023-0

Borgers, N., Hox, J., & Sikkel, D. (2004). Response effects in surveys on children and adolescents: The effect of number of response options, negative wording, and neutral mid-point. *Quality & Quantity, 38,* 17-33.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36*(1), 111-150.

Cabooter, E., Weijters, B., Geuens, M., & Vermeir, I. (2016). Scale format effects on response option interpretation and use. *Journal of Business Research, 69*(7), 2574-2584.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessmen*t. Beverly Hills, CA: Sage.

Carroll, J. B. (1953). An analytical solution for approximating simple structure in factor analysis. *Psychometrika, 18*(1), 23-38.

Chang, L. (1995a). Connotatively inconsistent test items. *Applied Measurement in Education, 8,* 199-209.

Chang, L. (1995b). Connotatively consistent and reversed connotatively consistent items are not fully equivalent: Generalizability study. *Educational and Psychological Measurement, 55*, 991-997.

Cliff, N. (1988). The eigenvalues-greater-than-one rule and the reliability of components. *Psychological Bulletin, 103*, 276-279. doi:10.1037/0033-2909.103.2.276

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

Cloud, J., & Vaughn, G. M. (1970). Using balanced scales to control acquiescence. *Sociometry, 33*, 193-206.

Coleman, C. M. (2013). *Effects of negative keying and wording in attitude measures: A mixed-methods study* (Unpublished doctoral dissertation). James Madison University, Virginia.

Colosi, R. (2005). Negatively worded questions cause respondent confusion. *Proceedings of the Survey Research Methods Section, American Statistical Association 2005*, 2896-2903.

Comrey, A. L. (1978). Common methodological problems in factor analytic studies. *Journal of Consulting and Clinical Psychology, 46*, 648-659.

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

Cordery, J. L., & Sevastos, P. P. (1993). Responses to the original and revised job diagnostic survey: Is education a factor in responses to negatively worded items? *Journal of Applied Psychology, 78*, 141-143.

Corwyn, R. (2000). The factor structure of global self-esteem among adolescents and adults. *Journal of Research in Personality, 34*, 357-379.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Cota, A. A., Longman, R. S., Holden, R. R., Fekken, G. C., & Xinaris, S. (1993). Interpolating 95th percentile eigenvalues from random data: An empirical example. *Educational and Psychological Measurement, 53*(3), 585-596.

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: agreeing response set as a personality variable. *The Journal of Abnormal and Social Psychology, 60*(2), 151-174.

Creed, P. A., Patton, W., & Bartrum, D. (2002). Multidimensional properties of the LOT-R: Effects of optimism and pessimism on career and well-being related variables in adolescents. *Journal of Career Assessment, 10*(1), 42-61.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Harcourt Brace Jovanovich.

Curry, J. P., Wakefield, D. S., Price, J. L., & Mueller, C. W. (1986). On the causal ordering of job satisfaction and organizational commitment. *Academy of Management Journal, 29*(4), 847-858.

De Ayala, R. J., & Hertzog, M. A. (1991). The assessment of dimensionality for use in item response theory. *Multivariate Behavioral Research, 26*(4), 765-792.

Dingman, H. F., Miller, C. R., & Eyman, R. K. (1964). A comparison between two analytic rotational solutions where the number of factors is indeterminate. *Behavioral Science, 9*(1), 76-80.

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling, 13*(3), 440-464.

Ebesutani, C., Drescher, C. F., Reise, S. P., Heiden, L., Hight, T. L., Damon, J. D., & Young, J. (2012). The importance of modeling method effects: Resolving the (uni)dimensionality of the Loneliness Questionnaire. *Journal of personality assessment, 94*(2), 186-195.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272-299. doi:10.1037/1082-989X.4.3.272

Fava, J. L., & Velicer, W. F. (1992). The effects of overextraction on factor and component analysis. *Multivariate Behavioral Research, 27*(3), 387-415.

Fava, J. L., & Velicer, W. F. (1996). The effects of underextraction in factor and component analyses. *Educational and Psychological Measurement, 56*(6), 907-929.

Ferguson, E., & Cox, T. (1993). Exploratory factor analysis: A user's guide. *International Journal of Selection and Assessment, 1*, 84-94.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466-491. doi: 10.1037/1082-989X.9.4.466

Flora, D. B., & Flake, J. K. (2017). The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement, 49*(2), 78-88. doi: 10.1037/cbs0000069

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*(3), 286-299.

Finney, S. (2001). *A comparison of the psychometric properties of negatively and positively worded questionnaire items* (Unpublished doctoral dissertation). The University of Nebraska, Lincoln.

Furr, R. M., & Bacharach, V. R. (2013). *Psychometrics: An introduction*. Thousand Oaks, CA: Sage.

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement, 55*(3), 377-393.

Gorsuch, R. L. (1983). *Factor analysis* (2nd Ed.). Hillsdale, NJ: L. Erlbaum Associates.

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment, 68*(3), 532-560.

Greenberger, E., Chen, C., Dmitrieva, J., & Farruggia, S. P. (2003). Item-wording and the dimensionality of the Rosenberg self-esteem scale: Do they matter? *Personality and Individual Differences, 35*(6), 1241-1254.

Gu, H., Wen, Z., & Fan, X. (in press). Examining and controlling for wording effect in a self-report measure: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, doi: 10.1080/10705511.2017.1286228

Guion, R. M. (1977). Content validity—the source of my discontent. *Applied Psychological Measurement, 1*(1), 1-10.

Guttman, L. (1954). A new approach to factor analysis: The radex. In P. F. Lazarsfeld (Ed.), *Mathematical Thinking in the Social Sciences* (pp. 258–349). New York, NY: Free Press of Glencoe.

Hakstian, A. R., Rogers, W. T., & Cattell, R. B. (1982). The behavior of number-of-factors rules with simulated data. *Multivariate Behavioral Research, 17*(2), 193-219.

Harman, H. H. (1976). *Modern factor analysis*. Chicago, IL: University of Chicago Press.

Harvey, R. J., Billings, R. S., & Nilan, K. J. (1985). Confirmatory factor analysis of the job diagnostic survey: Good news and bad news. *Journal of Applied Psychology, 70*, 461-468.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139-164.

Hayashi, K., Bentler, P. M., & Yuan, K. H. (2007). On the likelihood ratio test for the number of factors in exploratory factor analysis. *Structural Equation Modeling, 14*(3), 505-526.

Hazlett-Stevens, H., Ullman, J. B., & Craske, M. G. (2004). Factor structure of the Penn State Worry Questionnaire: Examination of a method factor. *Assessment, 11*, 361-370.

Hendrickson, A. E., & White, P. D. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology, 17*, 65-70.

Hensley, W. E., & Roberts, M. K. (1976). Dimensions of Rosenburg's self-esteem scale. *Psychological Reports, 38*, 583-584. doi:10.2466/pr0.1976.28.2.582

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*(3), 393-416.

Herzberg, P. Y., Glaesmer, H., & Hoyer, J. (2006). Separating optimism and pessimism: a robust psychometric analysis of the revised Life Orientation Test (LOT-R). *Psychological assessment, 18*(4), 433-438.

Holden, R. R., Fekken, G. C., & Jackson, D. N. (1985). Structured personality test item characteristics and validity. *Journal of Research in Personality, 19*(4), 386-394.

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

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185. doi:10.1007/BF02289447

Horn, L. R. (1989). *A natural history of negation*, Chicago and London: The University of Chicago Press.

Hoyle, R. H. (1995). *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: Sage Publications Inc.

Hoyle, R. H., & Duvall, J. L. (2004). Determining the number of factors in exploratory and confirmatory factor analysis. In D. Kaplan (Ed.), *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 301-313). Thousand Oaks, CA: Sage Publications.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424-453.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55.

Hubley, A. M., Wu, A. D., Liu, Y., & Zumbo, B. D. (2017). Putting flesh on the psychometric bone: Making sense of IRT parameters in non-cognitive measures by investigating the social cognitive aspects of the items. In B. D. Zumbo and A. M. Hubley (Eds.), *Understanding and Investigating Response Processes: Advances in Validation Research* (pp. 69-91). New York, NY: Springer.

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *The Journal of General Psychology, 123*(3), 207-215.

Hui, C. H., & Triandis, H. C. (1985). The instability of response sets. *Public Opinion Quarterly, 49*, 253-260.

Humphrey, L. G., & Ilgen, D. R., (1969). Note on a criterion for the number of common factors. *Educational and Psychological Measurement, 29*, 571-578.

Humphreys, L. G., & Montanelli, R. G. (1974). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research, 10*, 193-205.

IBM Corp. (2012). IBM SPSS Statistics for Windows (Version 21.0) [computer software]. Armonk, NY: IBM Corp.

Ibrahim, A. M. (2001). Differential responding to positive and negative items: The case of a negative item in a questionnaire for course and faculty evaluation. *Psychological Reports, 88*(2), 497-500.

Idaszak, J. R., & Drasgow, F. (1987). A revision of the job diagnostic survey: Elimination of a measurement artifact. *Journal of Applied Psychology, 72*(1), 69-74.

Innamorati, M., Lester, D., Balsamo, M., Erbuto, D., Ricci, F., Amore, M., Girardi, P., & Pompili, M. (2014). Factor validity of the Beck Hopelessness Scale in Italian medical patients. *Journal of Psychopathology and Behavioral Assessment, 36*(2), 300-307.

Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14*(1), 6-23.

Jackson, P. R., Wall, T. D., Martin, R., & Davids, K. (1993). New measures of job control, cognitive demand, and production responsibility. *Journal of Applied Psychology, 78,* 753-762.

Jennrich, R. I., & Sampson, P. F. (1966). Rotation for simple loadings. *Psychometrika, 31*, 313-323.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language*. Scientific Software International.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*(3), 187-200.

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

Kieruj, N. D., & Moors, G. (2013). Response style behavior: Question format dependent or personal style? *Quality & Quantity, 47*(1), 193-211.

Kline, P. (1994). *An easy guide to factor analysis*. New York, NY: Routledge.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd Ed). New York: Guilford Press.

Lai, J. C. (1994). Differential predictive power of the positively versus the negatively worded items of the Life Orientation Test. *Psychological Reports, 75*, 1507-1515.

Lai, J. C., & Yue, X. (2000). Measuring optimism in Hong Kong and mainland Chinese with the revised Life Orientation Test. *Personality and Individual Differences, 28*(4), 781-796.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 5-53.

Liu, Y., Wu, A. D., & Zumbo, B. D. (2010). The impact of outliers on Cronbach's coefficient alpha estimate of reliability: Ordinal/rating scale item responses. *Educational and Psychological Measurement, 70*(1), 5-21.

Liu, Y., Zumbo, B. D., & Wu, A. D. (2012). A demonstration of the impact of outliers on the decisions about the number of factors in exploratory factor analysis. *Educational and Psychological Measurement, 72*(2), 181-199.

MacCallum, R. C. (1983). A comparison of factor analysis programs in SPSS, BMDP, and SAS. *Psychometrika, 48,* 223-231.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490-504.

Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology, 22*, 37-49.

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70*, 810-819.

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(3), 320-341.

Martin, J. (1964). Acquiescence: Measurement and theory. *British Journal of Social and Clinical Psychology, 3*, 316-326.

McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*(1), 64-82.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*(10), 955-966.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45*(1-3), 35-44.

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development and validation of the Penn State Worry Questionnaire. *Behaviour Research and Therapy, 28*(6), 487-495.

Miles, J., & Shevlin, M. (2007). A time and a place for incremental fit indices. *Personality and Individual Differences, 42*(5), 869-874.

Motl, R. W., & Conroy, D. E. (2000). Validity and factorial invariance of the Social Physique Anxiety Scale. *Medicine and Science in Sports Exercise, 32*, 1007–1017.

Motl, R. W., Conroy, D. E., & Horan, P. M. (2000). The Social Physique Anxiety Scale: An example of the potential consequence of negatively worded items in factorial validity studies. *Journal of Applied Measurement, 1*, 327-345.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105*(3), 430-445.

Muthén, B. O. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics, 22*(1), 43-65.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*(1), 115-132.

Muthén, B., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus Web Notes, No. 4* (Version 5), 1-22.

Muthén L. K. & Muthén B. O. (1998–2012). MPlus (Version 7) [computer software]. Los Angeles, CA: Muthén and Muthén.

Nunnally, J. C. (1978). *Psychometric theory (2nd Ed.)*. New York: McGraw-Hill.

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

Osborne, J. W. (2015). What is rotating in exploratory factor analysis. *Practical Assessment, Research & Evaluation, 20*(2). Available online: http://pareonline.net/getvn.asp?v=20&n=2

Pilotte, W. J., & Gable, R. K. (1990). The impact of positive and negative item stems on the validity of a computer anxiety scale. *Educational and Psychological Measurement, 50*(3), 603-610.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879-903.

Pohl, S., & Steyer, R. (2010). Modeling common traits and method effects in multitrait-multimethod analysis. *Multivariate Behavioral Research, 45*, 45-72.

Raiche, G., & Magis, D. (2010). nFactors: Parallel Analysis and Non-Graphical Solutions to the Cattell Scree Test (Version 2.3.3) [R package].

Reisinger, Y., & Mavondo, F. (2006). Structural equation modeling: Critical issues and new developments. *Journal of Travel and Tourism Marketing, 21*(4), 41-71.

Revelle, W. (2014). psych: Procedures for Personality and Psychological Research (Version 1.4.1) [R package].

Rhemtulla, M., Brosseau-Liard, P., & Savalei, V. (2012). How many categories is enough to treat data as continuous? A comparison of robust continuous and categorical SEM estimation methods under a range of non-ideal situations. *Psychological Methods, 17*(3), 354-373. Retrieved from: http://psych.colorado.edu/~willcutt/pdfs/Rhemtulla_2012.pdf [Accessed 30 Jul. 2016].

Robinson, J. R., Shaver, P. R., & Wrightsman, L. S. (Eds.) (1991). *Measures of personality and social psychological attitudes*. San Diego, CA: Academic Press.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Rummel, R. J. (1970). *Applied factor analysis*. Evanston, IL: Northwestern University Press.

Ryff, C. D., & Keyes, C. L. M. (1995). The structure of psychological well-being revisited. *Journal of Personality and Social Psychology, 69*(4), 719-727.

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16*(4), 561-582.

Sauro, J., & Lewis, J. R. (2011). When designing usability questionnaires, does it hurt to be positive? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2215-2224). ACM.

Savalei, V., & Falk, C. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate Behavioral Research, 49*, 407-424. doi:10.1080/00273171.2014.931800

Scheier, M. F., & Carver, C. S. (1985). Optimism, coping and health: Assessment and implications of generalized outcome expectancies. *Health Psychology, 4*, 219-247.

Scheier, M. F., & Carver, C. S. (1987). Dispositional optimism and physical well-being: The influence of generalized outcome expectancies on health. *Journal of Personality, 55*, 169-210.

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the Life Orientation Test. *Journal of Personality and Social Psychology, 67*(6), 1063-1078.

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement, 9*(4), 367-373.

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment, 29*(4), 304-321.

Schriesheim, C. A., Eisenbach, R. J., & Hill, K. D. (1991). The effect of negation and polar opposite item reversals on questionnaire reliability and validity: An experimental investigation. *Educational and Psychological Measurement, 51*(1), 67-78.

Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. *Educational and Psychological Measurement, 41*(4), 1101-1114.

Schwarz, N., Strack, F., Hippler, H. J., & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology, 5*(3), 193-212.

Sliter, K. A., & Zickar, M. J. (2014). An IRT examination of the psychometric functioning of negatively worded personality items. *Educational and Psychological Measurement, 74*(2), 214-226.

Slocum, S. L. (2005). *Assessing unidimensionality of psychological scales: Using individual and integrative criteria from factor analysis* (Unpublished doctoral dissertation). University of British Columbia, Vancouver, Canada.

Spector, P. E. (1992). *Summated rating scale construction: An introduction* (Sage university paper series on Quantitative Applications in the Social Sciences, No. 07-082). Newbury Park, CA: Sage.

Spector, P. E., Van Katwyk, P. T., Brannick, M. T., & Chen, P. Y. (1997). When two factors don't reflect two constructs: How item characteristics can produce artifactual factors. *Journal of Management, 23*(5), 659-677.

Steed, L. (2001). Further validity and reliability evidence for Beck Hopelessness scale scores in a nonclinical sample. *Educational and Psychological Measurement, 61*, 303-316. doi: 10.1177/00131640121971121

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association. doi: 10.1037/10694-010

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56*(2), 197-208.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529-554.

Thurstone L. L. (1947). *Multiple-factor analysis*. Chicago: University of Chicago Press.

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with "well fitting" models. *Journal of Abnormal Psychology, 112*(4), 578-598.

Tomás, J., & Oliver, A. (1999). Rosenberg's self-esteem scale: Two factors or method effects. *Structural Equation Modeling, 6*, 84-98.

Tourangeau, R., Couper, M. P., & Conrad, F. G. (2007). Color, labels, and interpretive heuristics for response scales. *Public Opinion Quarterly, 71*(1), 91-112.

Tourangeau, R., Couper, M. P., & Conrad, F. G. (2013). "Up Means Good" the effect of screen position on evaluative ratings in web surveys. *Public Opinion Quarterly, 77*(1), 69-88.

van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PloS one, 8*(7), e68967. Retrieved from: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0068967#pone-0068967-t004 [Accessed 27 Oct. 2015].

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helmes (Eds.), *Problems and Solutions in Human Assessment: Honoring Douglas N. Jackson at Seventy*. Norwell, MA: Kluwer Academic.

Wang, Y. J., Minor, M. S., & Wei, J. (2011). Aesthetics and the online shopping environment: Understanding consumer responses. *Journal of Retailing, 87*(1), 46-58.

Wang, W. C., Chen, H. F., & Jin, K. Y. (2014). Item response theory models for wording effects in mixed-format scales. *Educational and Psychological Measurement*, doi: 10.1177/0013164414528209.

Warne, R. T., & Larsen, R. (2014). Evaluating a proposed modification of the Guttman rule for determining the number of factors in an exploratory factor analysis. *Psychological Test and Assessment Modeling, 56,* 104-123.

Weiss, D. (1976). Multivariate procedures. In Dunnette, M. D. (Ed.), *Handbook of Industrial/Organizational Psychology*. Chicago, IL: Rand McNally.

Whiteside-Mansell, L., & Corwyn, R. F. (2003). Mean and covariance structures analyses: An examination of the Rosenberg self-esteem scale among adolescents and adults. *Educational and Psychological Measurement, 63*, 163-173.

Wong, N., Rindfleisch, A., & Burroughs, J. E. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? The case of the Material Values Scale. *Journal of Consumer Research, 30*(1), 72-91.

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28*(3), 189-194.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Unplublished doctoral dissertation). University of California Los Angeles, California.

Zimprich, D., Kliegel, M., & Rast, P. (2011). The factorial structure and external validity of the prospective and retrospective memory questionnaire in older adults. *European Journal of Ageing, 8*(1), 39-48.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao and S. Sinharay (Eds.). *Handbook of Statistics, Volume 26: Psychometrics* (pp.45-79). Amsterdam, NL: Elsevier Science B.V.

Zumbo, B. D. (2017). On models and modeling in measurement and validation studies. In B. D. Zumbo and A. M. Hubley (Eds.), *Understanding and Investigating Response Processes: Advances in Validation Research* (pp. 363-370). New York, NY: Springer.

Zumbo, B. D., & Chan, E. K. H. (Eds.). (2014). *Validity and validation in social, behavioral, and health sciences*. New York, NY: Springer.

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods, 6*, 21-29.

Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research, 17,* 253-269.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*(3), 432-442.