Fall 2022

Investigating Bifactor Models and Fit Indices for Unidimensionality: An Illustration With Method Effects Due to Item Wording

Elizabeth A. Leighton

INVESTIGATING BIFACTOR MODELS AND FIT INDICES FOR UNIDIMENSIONALITY: AN ILLUSTRATION WITH METHOD EFFECTS DUE TO ITEM WORDING

by

Elizabeth A. Leighton

Bachelor of Science
College of Charleston, 1998

Master of Social Work
University of South Carolina, 2005

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Educational Psychology and Research

College of Education

University of South Carolina

2022

Accepted by:

Christine DiStefano, Major Professor

Tammiee Dickenson, Committee Member

Jin Liu, Committee Member

Diane M. Monrad, Committee Member

Cheryl L. Addy, Interim Vice Provost and Dean of the Graduate School

DEDICATION

To my Mom for always supporting, encouraging, and loving me. To Charles, Sanders, and Will…do your best, dream hard, and continue to be amazing people. Finally, to my Dad for always telling me to "look it up" instead of giving me the answer. I wish you were here to see this!

ACKNOWLEDGEMENTS

I would like to extend my gratitude and appreciation to all of those who have mentored and helped me in this journey. First, to Dr. Christine DiStefano, I have learned everything from you! Thank you for your mentorship, guidance, and patience in my learning process. To Dr. Diane Monrad for being my mentor both professionally and personally. I cannot begin to express how much your mentorship and friendship has meant to me! To Dr. Tammiee Dickenson for your continued mentorship and support. To Dr. Jin Liu, you first served as a peer and outstanding group member in several classes to an outstanding and supportive member on my committee. I look forward to our future collaborations! I am honored by the “womanpower” of my superb committee! I would also like to acknowledge the excellent professors and teachers who guided me in this process, including Dr. Robert Johnson, Dr. Bethany Bell, Dr. Michelle Bryan, Dr. Erik Drasgow, Dr. Brian Habing, Dr. Thomas Hebert, Dr. Steven Liu, and Dr. Michael Seaman.

Finally, to my friends and fellow co-workers for your support in this process, including Dr. Mihaela Ene, Dr. Yin Burgess, Dr. Kate McGrath, Dr. Dawn Coleman, Dr. Kelvin Pompey, Dr. Xumei Fan, and Dr. Ginger Jiang. And always to Dr. Ashlee Lewis! Finally, a special thank you to my amazing friends for always being there and supporting me in my journey, Trey Jameson, Kristin LaRoche, LaToya Gibson, and Heather Grohn.

ABSTRACT

The use of unidimensional scales that contain both positively and negatively worded items is common in both the educational and psychological fields. However, dimensionality investigations of these instruments often lead to a rejection of the theorized unidimensional model in favor of multidimensional structures, leaving researchers at odds for how best to treat their data. One modeling technique that has gained attention in recent years related to item wording effects is the bifactor model, which specifies a general construct that explains the covariation among all items, and two method factors that explain additional covariation among items with similar wording.

Recent applications of the bifactor model have called for utilizing select model-based indices to establish “essentially unidimensional” models (i.e., Bonifay et al., 2015; Reise, Scheines, et al., 2013; Rodriguez et al., 2016a, 2016b). The purpose of this study was to investigate the performance of the select bifactor model-based indices, explained common variance (ECV), omega hierarchical (omegaH), and the percent of uncontaminated correlations (PUC), in establishing essentially unidimensional models. A Monte Carlo simulation study was conducted to examine the performance of these indices under conditions frequently encountered when item wording method effects are identified within unidimensional scales. Conditions included: 3 conditions related to the number of items per method factor (i.e., PUC values; 6:6, 8:4, 3:6) x 6 item loading values and patterns between the method factors and the general factor (0.7, 0.5 for the general factor; 0.5 or 0.3 for the method factors; four balanced and two unbalanced

method factor conditions) x 2 item-level distributions (all normal; negative items non-normally distributed) x 2 model misspecification (correctly specified bifactor model, misspecified unidimensional model).

Outcomes examined in this study included the relative bias in factor loadings estimated from a misspecified unidimensional model (i.e., ignoring the method factors) when data were simulated from a bifactor model. Results indicated the presence of unbalanced method factors (both in size and magnitude) required a substantially "stronger" general factor, as evidenced by higher ECV (>0.80) and omegaH (>0.85) values, to reach "essentially unidimensional" status. Furthermore, PUC was found to be relatively non-informative as an indicator of essential unidimensionality in the context of item wording method effects. Examination of the relationship between ECV and select model-fit indices indicated that while they are somewhat related (i.e., Reise, Scheines, et al., 2013), ECV, as well as omegaH, are more reliable as indices in identifying when data can be treated as essentially unidimensional in the context of item wording method effects. Recommendations and guidelines are provided for applied researchers utilizing unidimensional scales that may be contaminated by item wording method effects.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC.. Akaike Information Criterion

BIC... Bayesian Information Criterion

CFA...Confirmatory Factor Analysis

CFI .... Comparative Fit Index

ECV..Explained Common Variance

OmegaH .... Omega Hierarchical

PUC..Percent of Uncontaminated Correlations

TLI ....Tucker-Lewis Index

CHAPTER 1

INTRODUCTION

The use of unidimensional scales to measure latent constructs is common in many areas of the social sciences. Typically, a scale is composed of a set of items intended to measure a latent construct, with a composite score (i.e., a sum score or mean score) serving as an empirical indicator of the construct (McIver & Carmines, 1981). When well developed, the items that encompass a scale should: (a) measure the single construct but not be overly redundant, (b) capture the full breadth and depth of the construct, and (c) provide acceptable reliability (Reise et al., 2010). A unidimensional scale is one in which a set of items measures one construct (Gustafsson & Aberg-Bengtsson, 2010; Johnson & Morgan, 2016). This implies that the property of unidimensionality holds for the instrument, as well as the property of local independence. Local independence assumes that all item responses are derived solely from a single latent variable and that all covariation between items is explained by the latent variable (Edwards et al., 2018). Violations of local independence imply that additional variance is present in the data that is not explained by the single construct and that additional factors and/or model specifications are needed to explain the covariation between items.

Unidimensionality is desirable property in measurement given the model's simplicity and parsimony. While it is assumed that the construct of interest accounts for the covariation among item responses, there may be other sources of systematic error encountered by researchers when using a scale that is thought to be unidimensional. For

example, when features related to the measurement method and not the substantive trait measured, create systematic error, this is known as a method effect (Bagozzi, 1993; Maul, 2013). Broadly, a "method" can be defined by the multiple aspects that encompass the measurement process, including item content, response format, instructions provided to respondents, as well as internal aspects concerning the respondents themselves and the context surrounding an assessment's administration (Fiske, 1982; Podsakoff et al., 2012). Podsakoff et al. (2003) characterizes sources of method effects across four areas: common rater effects (e.g., social desirability, acquiescence), measurement context effects (e.g., measuring predictor and criterion variables using the same method), item context effects (e.g., item position, scale length), and item characteristic effects (e.g., item complexity, ambiguity). With the additional systematic variance created by method effects, researchers often find that the theorized unidimensional model does not provide adequate fit to the data, and alternative specifications are needed to fit the data and explain the excess covariation (Brown, 2015).

One specific type of method effect that impacts unidimensional scales involves the use positively and negatively worded items. Data from unidimensional scales are often collected via self-report instruments. In self-report instruments, respondents are asked to directly reflect on their individual appraisal or global self-rating on a construct of interest (Paulhus & Vazire, 2007). Under ideal conditions, respondents read an item and respond accordingly to a response scale (e.g., Likert scale) based on the level of the underlying trait. However, in using self-report data from instruments containing positive and negatively worded items together on the same scale, researchers run the risk of obtaining data that may be biased due to how the examinee is responding and not their

level of the trait in question. For example, acquiescence, or “yea-saying,” is frequently cited as problematic for researchers collecting data from self-report instruments (e.g., Baumgartner & Steenkamp, 2001; Billiet & McClendon, 2000). Acquiescence is said to occur when there is an overuse of the “agreement” categories on the response scale independent of the content (i.e., Agree or Strongly Agree; Baumgartner & Steenkamp, 2001; Van Vaerenbergh & Thomas, 2013; Rammstedt & Farmer, 2013). The inclusion of both positively and negatively worded items, particularly if the scale is balanced (i.e., equal positively and negatively worded items), is thought to control for the effects of acquiescence (Baumgartner & Steenkamp, 2001; Cloud & Vaughan, 1970). Additionally, the inclusion of negatively or reversed worded items is thought to represent a cognitive "speed bump" designed to capture respondents' attention and prevent them from engaging in response sets in which they respond based on general feelings rather than attend to the actual item content (Barnette, 2000; Podsakoff et al., 2003). Furthermore, their inclusion is thought to provide a more complete and comprehensive coverage of the measured construct, thereby improving scale and construct validity (Weijters & Baumgartner, 2012).

However, assuming negatively worded items do all of this while maintaining sufficient psychometric quality has been called a "methodological urban legend" (Dalal & Carter, 2015). There is no consensus in the literature that including negatively worded items reduces the effects of response biases, and in some instances, may introduce additional response biases (van Sonderen et al., 2013). In fact, it is common to find differential response patterns between items when both positively and negatively worded items are contained within an instrument (e.g., Spector et al., 1997).

Negatively Worded Items

Throughout the literature, “negatively worded items” and scales that contain both positively and negatively worded items are referred to in many ways, such as reverse worded (e.g., Brown, 2003; Ebesutani et al., 2012; Kam et al., 2021), reversed/negatively keyed (e.g., Lindwall et al., 2012), mixed-format items/keyed (e.g., Chen & Jin, 2020; Wang et al., 2015), and reversed-polarity items (e.g., Boley et al., 2021; Herche & Engelland, 1996). While some researchers make clear distinctions between the semantics of mixed-format items (e.g., Schriesheim et al., 1991), others simply refer to them as negatively worded items (Weijters & Baumgartner, 2012). Weijters and Baumgartner (2012) offer a clear example of positively and negatively worded item formats in measuring the construct of extroversion. A positively worded item designed to measure extroversion is *I am talkative*, whereas a negatively or reverse worded item is *I am not talkative*, or *I am quiet*. The use of negatively worded items typically implies the need for the item to be recoded or reverse coded prior to analysis so that all items have the same directional relationship with the measured construct (Chyung et al., 2018; Weijters et al., 2013).

The assumption of negatively worded items is that they measure the same construct to the same degree as the positively worded items (Marsh, 1996). If this assumption holds, the expected response patterns between these items may look like this: a respondent with a high level of the construct would be expected to "agree" with a positively worded item, while, conversely, they would be expected to "disagree" with a negatively worded item, and vice versa for respondents with a low level of the construct. However, this is not always the case, and those with high levels of the construct do not

always "disagree" when encountering a negatively worded item (e.g., Benson & Hocevar, 1985; Kam et al., 2021; Marsh, 1986; Weems et al., 2006). This type of differential response pattern creates response bias that can impact the psychometric properties of the instrument.

Considering prior studies, researchers have found biased reliability estimates, statistically different means between the positive and negative items, and questions about the dimensional structure underlying the instrument when both positively and negatively worded items are included (e.g., Barnette, 2000; Marsh, 1996; Schriesheim et al., 1991; Weems et al., 2006). Conclusions from studies investigating the functioning of positively and negatively worded items suggest that respondents are often confused or that there is a misunderstanding when respondents encounter a negatively worded item which leads to careless responding and differential response patterns (e.g., Schriesheim et al., 1991; Swain et al., 2008; Woods, 2006). However, other studies have found differential response patterns related to the inclusion of mixed worded items to be more complex than carelessly responding to an item. These studies indicate that factors such as response styles related to personality traits (e.g., DiStefano & Motl, 2006; Quilty et al., 2006), verbal or cognitive ability (e.g., Corwyn, 2000; Gnambs & Schroeder, 2020; Marsh et al., 2010), and social desirability (e.g., Rauch et al., 2007) contribute to the differential response patterns. In other words, these patterns are related to factors beyond that of the intended measured construct. Such findings have led researchers to conclude that the inclusion of both positively and negatively worded items leads to method effects, which in turn, distort an instrument's psychometric properties.

Dimensionality Concerns with Unidimensional Scales

There are many examples of self-report, unidimensional scales that include both positively and negatively worded items in the literature. The Rosenberg Self Esteem Scale (RSES; Rosenberg, 1965), General Heath Questionnaire-12 (GHQ-12; Goldberg & Williams, 1988), Consideration of Future Consequences Scale (CFC; Strathman et al., 1994), Core Self-Evaluation Scale (CSES; Judge et al., 2003), Life Orientation Test-Revised (LOT-R; Scheier et al., 1994), and the Penn State Worry Questionnaire (PSWQ; Meyer et al., 1990) are but a few examples. Each scale has a similar design with respondents asked to indicate their level of agreement across a series of items via a four or five-point Likert response scale. Users of a scale are instructed to reverse code negatively worded items and sum item responses to obtain a composite score.

Each of the referenced unidimensional scales were designed to measure a single construct, however, with the inclusion of both positively and negatively worded items, dimensionality investigations of these instruments result in similar findings: a rejection of the theorized unidimensional model in favor of alternative model specifications. When multidimensionality is identified as the result of a potential method effect, such as the case with item wording effects, researchers must decide on its potential substantive meaning and how best to model the latent construct the instrument was designed to measure (Marsh, 1996). These findings have implications for how to treat the data in subsequent analyses. In general, findings of multidimensionality typically imply the use of subscale scores in place of an overall composite score (Reise et al., 2000). This assumes that the dimensions identified are construct relevant and represent reliable substantive facets, sub-traits, or components related to the overall construct the

instrument was designed to measure. However, studies have indicated that method effects in the context of item wording directionality are often related to construct irrelevant factors and interpretation of substantive dimensions defined by item wording may not be valid (e.g., Brown, 2003; Hazlett-Stevens et al., 2004; Huang & Dong, 2012; Greenberger et al., 2003; Spector et al., 1997). Furthermore, Rauch et al. (2007) noted that "…deviation from unidimensionality of observed scores does not imply deviation from unidimensionality of [the construct]" (p. 1597).

Bifactor Models

In recent years, the bifactor model (Holzinger & Swineford, 1937) has gained attention as an alternative for accommodating method effects. A bifactor model represents a special case of the correlated-traits correlated-methods model (CTCM), a model frequently specified within the confirmatory factor analysis (CFA) multitrait-multimethod (MTMM) framework used to model and examine method effects (Bollen & Hoyle, 2012; Gu et al., 2017). This model specifies that the covariance among as set of items is explained by both a general factor that explains the common variance among all items, and group or specific factors that explain common variance among a cluster of items that are similar in content (Reise, 2012; Reise, Scheines, et al., 2013). Figure 1.1 depicts a CFA path diagram for a bifactor model in the context of modeling item wording method effects.

As shown in Figure 1.1, the instrument modeled is a 12-item, balanced scale containing six positively worded items (x1 – x6) and six negatively worded items (x7 – x12). As depicted in this specification, all items load on the general factor (i.e., the construct the instrument was designed to measure) and one of two method factors (i.e.,

specific factors) representing the positively worded item method factor and the negatively worded item method factor.

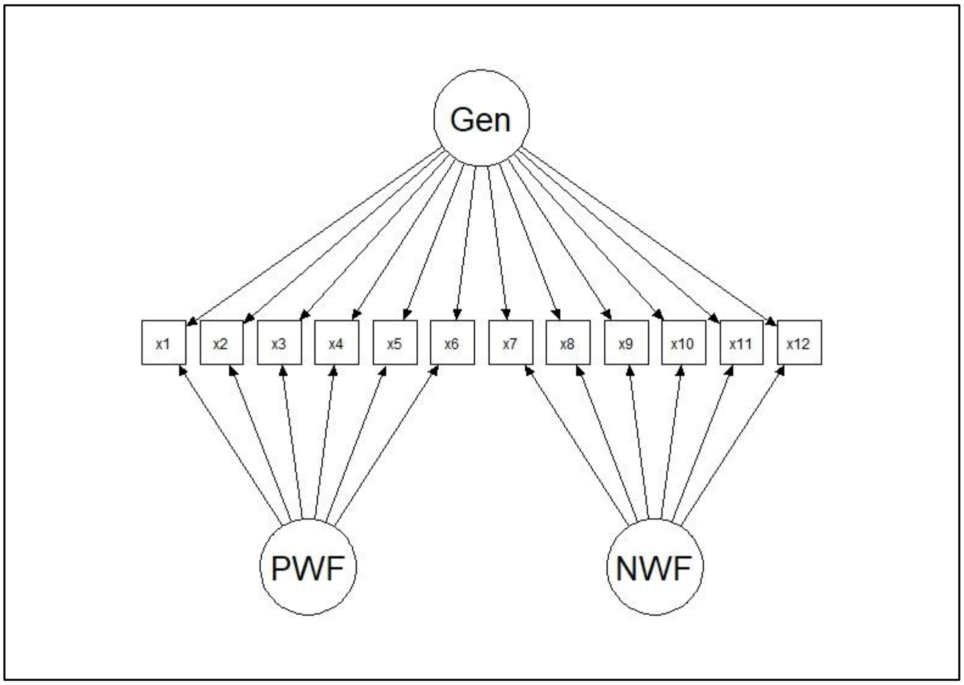

Figure 1.1 *Path Diagram for a Bifactor Model with a General Factor and Two Method Factors*

With the variance decomposition provided by the bifactor model, researchers can assess the degree to which both the general and specific factors influence item-level variance, allowing researchers to determine the "strength" of the general factor in relation to the specific factors (Reise et al., 2010). More specifically, researchers can (a) examine the partitioning of variance across general and group factors, (b) control for multidimensionality in the presence of nuisance dimensions, (c) assess the degree to which data have a strong enough general factor to fit a unidimensional measurement model, and (d) evaluate the adequacy of both total scores and subscale scores (Rodriguez

et al., 2016b). In other words, researchers can assess the degree to which scales and scale scores are perhaps "essentially unidimensional" versus multidimensional.

Essential Unidimensionality and Bifactor Models

The concept of "essential unidimensionality" is frequently applied within the item response theory (IRT) literature and IRT applications, acknowledging the fact that data are rarely, if ever, strictly unidimensional (Reise et al., 2007). Essential unidimensionality (i.e., Stout, 1987) implies that the data predominantly reflect one dimension or that the instrument *essentially* measures one dimension, even when other dimensions are identified (Nandakumar & Ackerman, 2004). Reise (2012) notes that it is well established within IRT literature that model parameters are robust to violations of unidimensionality, provided that the general factor is considered "strong."

In CFA, unidimensionality is typically evaluated by comparing model-data fit to a unidimensional model via conventional fit statistics such as the Comparative Fit Index (CFI), Tucker Lewis Index (TLI), Root Mean Squared Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR). Through comparisons of the fit statistics from a tested model to established guidelines (i.e., CFI $\geq .95$), researchers deem the data unidimensional or not and, if not, proceed to fit alternative multidimensional models that meet the "best fitting" criteria. Given the complexity involved in measuring many psychological constructs, and the need to include mixed item formats, it is not surprising that when the internal structure of instruments designed to be unidimensional is investigated, multidimensionality is found instead. In CFA, there is less focus on how other statistics, outside of model fit statistics, can provide a direct index to assess the degree of dimensionality or, more specifically, the departure from

unidimensionality, much in the way certain IRT based statistics, such as DETECT (Zhang & Stout, 1999), are used (Bonifay et al., 2015). Given the variance decomposition provided by the bifactor model, this model offers a promising technique for researchers to assess the degree of unidimensionality present in the data. In fact, Reise et al. (2010) argue that the bifactor model provides a clearer or more realistic conceptualization of unidimensionality by specifying a general factor that influences all items, but which may be contaminated by other, perhaps nuisance, factors (i.e., method effects).

As part of a comprehensive psychometric evaluation, in addition to traditional model fit statistics, bifactor models can be evaluated using several model-derived indices. These indices help researchers assess the strength of the general and specific factors, the degree to which total scores reflect the intended construct, and how reliable and applicable subscale scores are from specific factors (Rodriguez et al., 2016a, 2016b). These indices include the omega reliability indices (i.e., omega total, omega hierarchical), factor determinacy (FD), construct replicability (*H* index), explained common variance (ECV), and the percent of uncontaminated correlations (PUC). Briefly, the omega reliability indices can be used to assess the reliable variance in both unit-weighted, observed total and subscale scores (Rodriguez et al., 2016a, 2016b). Factor determinacy (*FD*) and the *H* index are used to assess the reliability of general and specific factors at the latent level (i.e., the reliability of factor scores or optimally weighted scores; Rodriguez et al., 2016a, 2016b). Finally, ECV and PUC are considered statistics of unidimensionality in that they can be used to assess the general factor's strength in relation to specific factors (Rodriguez et al., 2016a, 2016b). Taken together, an evaluation of ECV, omega hierarchical (omegaH), and PUC derived from a bifactor

model can be used to indicate when data can be represented as essentially unidimensional.

Rodriguez et al. (2016a) reviewed and evaluated 50 studies across the psychological, personality, and assessment literature in which a bifactor model was identified as the "best fitting" structure. In their review, they calculated the bifactor model-based indices for all 50 studies using the standardized factor loadings as presented in the articles. One key finding from Rodriguez et al. was that most of the studies reviewed claimed multidimensionality, mostly due to model fit of the bifactor, however, after evaluating the studies with the additional bifactor model indices, they found that in many cases, total scores primarily reflected variance due to the general factor. Such findings also relate to several recent studies suggesting the bifactor model may overfit the data by capturing additional model complexity and noise (Bonifay & Cai, 2017; Greene et al., 2019; Morgan et al., 2015; Murry & Johnson, 2013; Reise et al., 2016).

Results from Rodriguez et al.'s (2016a) review suggest the data in many instances was most likely essentially unidimensional. Specifically, they noted that in many cases total scores can be "interpreted as univocal indicators of a single latent variable, despite multidimensionality" (p. 232). Furthermore, Rodriguez et al. note that item-level multidimensionality can impose unnecessary complexity into the model in cases where (a) the general factor is strong, and (b) group factors are poorly defined. Such cases may be present when data are contaminated by item wording method effects.

Statement of the Problem

When data from unidimensional scales including both positively and negatively worded items is examined via CFA, dimensionality investigations often result in similar

conclusions: that is, a rejection of the strict unidimensional model in favor of a multidimensional model that explicitly models method effects related to item wording. These findings have implications for both the construct validity and psychometric properties of the instrument. Since the bifactor model has been included as part of an alternative model evaluation strategy for item wording method effects (i.e., Marsh et al., 2010), it is common to find researchers selecting this model based on the model fit indices familiar to the structural equation modeling (SEM) literature (i.e., CFI, TLI, RMSEA, SRMR, AIC, BIC) in combination with substantive conclusions and utility of this model. Most investigations have found the bifactor model, modeling both the positively and negatively worded method factors (e.g., Alessandri et al., 2015; Hyland et al., 2014; Hystad & Johnsen, 2020; Marsh et al., 2010; McKay, Morgan, et al., 2015; Salerno et al., 2017) or a modified version modeling only a negatively worded method factor (e.g., DiStefano & Motl, 2006; Gnambs & Schroeders, 2020; Gu et al., 2015; Ye, 2008) to provide the best fit to the data. While the bifactor model preserves the unidimensional nature of the general construct, it also indicates that the structure and data are contaminated by these method effects. These findings have led some to question the psychometric properties of the instrument, particularly the use of an overall composite score. For example, Marsh et al. (2010) in declaring the bifactor model as the appropriate modeling structure for the Rosenberg Self-Esteem Scale noted the unidimensional model was inappropriate and furthermore questioned the validity of applied research that previously utilized the RSES, noting that failing to control for the method effects results in biased interpretations. Others have echoed Marsh et al.'s concerns (i.e., Alessandri et al., 2015; Gana et al., 2013). However, Reise, Bonifay, et al. (2013) contend that while

such conclusions are valid, they may be unnecessarily concerning. Clarity concerning the unidimensionality of such instruments can be provided using additional statistics derived from a bifactor model.

More recent applications of the bifactor model to model item wording effects have applied the recommended model-based indices to make psychometric conclusions regarding the influence of item wording effects on the factor structure and scoring procedures (e.g., Arias & Arias, 2017; Hystad & Johnsen, 2020; McKay et al., 2015; Salerno et al., 2017; Urban et al., 2014). As outlined by Reise, Bonifay, et al. (2013) and Rodriguez et al. (2016a, 2016b), ECV, omega hierarchical, and PUC, can be used to assess the strength of the general factor and provide evidence for essential unidimensionality, effectively evaluating the extent to which the method effects related to item wording are contaminating the unidimensional property and structure. The information assists researchers in making decisions for how best to model the data in scoring procedures and subsequent statistical analyses.

A number of recommended benchmarks and guidelines in utilizing the bifactor model-based indices in establishing essential unidimensionality have been suggested, including omega hierarchical values greater than .80, ECV values greater than .70 to .80, and PUC and ECV values greater than .70 (Rodriguez et al., 2016a). A key finding from simulation studies examining these indices is that PUC, or the extent of data's "unidimensional" structure, plays an important role in moderating the effectiveness of ECV as an indicator of essential unidimensionality (Bonifay et al., 2015; Reise, Scheines, et al., 2013). As such, Reise, Scheines, et al. (2013) established the following tentative general benchmark: when PUC is < .80, ECV >.60 and omegaH >.70, bias in parameter

estimates should be minimal, and a unidimensional measurement model can be used. They note, however, that additional research is needed to assess the generalizability of this benchmark across other data conditions.

The purpose of the current study was to investigate the use of bifactor models and the recommended model-based indices in reaching decisions regarding essential unidimensionality in the context of item wording method effects. This study utilized a Monte Carlo simulation to evaluate the effectiveness in applying ECV, omega hierarchical, and PUC in determining when data can be treated as essentially unidimensional in the context of item wording method effects. This study expands on previous simulation studies investigating these indices (i.e., Bonifay et al., 2015; Gu et al., 2017; Reise, Scheines, et al., 2013) by examining how they perform under the context of various item and specific factor (i.e., method factor) characteristics frequently found in relation to item wording method effects, including the role of varying degrees of PUC, item distributional patterns, and unbalanced specific factors. As model selection within CFA is traditionally based on the assessment of model-data fit via fit indices, this study also included an evaluation of select model fit indices and their relationship between the model-based indices (i.e., ECV, PUC). The following research questions were addressed:

1. What is the relationship between the general factor strength indices (ECV, omega hierarchical, and PUC) and parameter bias in factor loading estimates when data are fit to a unidimensional measurement model?
2. What is the relationship between ECV and PUC in the context of item wording method effects?
3. What is the relationship between ECV, PUC, and model fit indices?

In summary, this study informs both methodological and applied researchers. First, the use of omegaH, ECV, and PUC as proposed indicators of essential unidimensionality, has received limited attention in simulation research. Examination of the functioning of these indices in more diverse data conditions is needed. Furthermore, given this study's focus on model-data fit in the CFA framework, this study provides an examination on how the bifactor model-based indices function either against or in combination with conventional model-fit indices in model section.

Second, in the context of item wording method effects, the use of omegaH, ECV, and PUC can provide researchers with information for when method effects can be ignored, and the data treated as unidimensional (i.e., the intended structure). Additionally, the use of a composite or total score that primarily reflects the target construct is typically desired in this context (e.g., global self-esteem). In this sense, the use of omgeaH, ECV, and PUC may be helpful in establishing the general or target construct's strength in the presence of the method factors (i.e., nuisance factors).

CHAPTER 2

LITERATURE REVIEW

The inclusion of both positively and negatively worded items is a common design practice in developing self-report instruments. While including items of varying wording direction is thought to prevent response styles such as acquiescence, their inclusion adds additional error to the data in the form of method effects, which can lead to questions about the scale's dimensionality, particularly if the scale was designed to be unidimensional. Unexpected findings regarding dimensionality can lead researchers to question how best to model the data, and subsequently, the appropriate use of a composite or total score obtained from these instruments. The use of bifactor modeling and the application of model-based indices derived from a bifactor model offer researchers with a promising technique to evaluate data collected from instruments that may be contaminated by item wording method effects.

This chapter begins with a detailed description of the bifactor model, including the model's assumptions and its relationship to the unidimensional model. Next, procedures for evaluating model-data fit, including the alternative model-based indices recommended for bifactor models are detailed, as well as issues encountered when evaluating model-data fit with bifactor models. Finally, a review of simulation studies investigating both the bifactor model-based indices and item wording method effects are reviewed. This chapter concludes with an overview of the current study.

Bifactor Models

Bifactor models, also known as nested or hierarchical models (Gignac & Watkins, 2013), are a type of measurement model that can be specified in confirmatory factor analysis (CFA) and employed within a full SEM model. A bifactor model can also be estimated within the exploratory factor analysis (EFA) framework via alternative factor rotation procedures such as a Schmid-Leiman transformation or the Jennrich-Bentler analytic rotations (e.g., Jennrich & Bentler, 2011; Jennrich & Bentler, 2012; Manslof & Reise, 2016; Markon, 2019). Additionally, bifactor models can also be utilized within the item response theory (IRT) framework (Gibbons & Hedeker, 1992; Gibbons et al., 2007).

In a bifactor model (Holzinger & Swineford, 1937), item responses are modeled as a function of a general factor, and no more than one group or specific factor (Reise, 2012). Each item loads simultaneously onto two latent factors, a general factor that explains the common variance among all items, and a specific factor that explains additional variance among a subset of items (Reise, 2012). Thus, two factor loadings will be estimated, one indicating the relationship to the general factor and the other indicating the relationship to the specific factor (Dunn & McCray, 2020). Additionally, since the general and specific factors are fit simultaneously in a bifactor model, specific factors provide an estimate of shared variance among a set of items, once common variance from the general factor has been removed (Dunn & McCray, 2020). This indicates the specific factors should be interpreted as residual factors derived from the general factor, rather than separate or independent factors as is typically modeled via a correlated-traits model (DeMars, 2013; Dunn & McCray, 2020).

Modeling Psychometric Multidimensionality

Bonifay et al. (2017) describe two primary applications of the bifactor model, (1) as a tool for evaluating the psychometric properties of an assessment or scale, and (2) as the representation of an entire domain of a construct, to which they refer to the use of the bifactor model in this context as "ambitious." However, as a tool for psychometric evaluation, Bonifay et al. note the bifactor model's applicability in assessing the reliability of both total scores and potential subscale scores. Within this context, Reise (2012) summarizes the key applications of the bifactor model as the ability to: (1) partition item variance, (2) provide a unidimensionality statistic, (3) provide evidence for the interpretation of a scale score, and (4) assess the viability of subscales. From this perspective, the bifactor model and its specification permits researchers to engage in a comprehensive psychometric evaluation of an instrument's properties (Rodriguez et al., 2016a; 2016b).

Morin et al. (2016) distinguish between substantive multidimensionality versus psychometric multidimensionality. They describe substantive multidimensionality as an instrument designed to assess multiple dimensions, with separate items aligned to each dimension. Substantive multidimensionality is what is typically modeled via a correlated-traits factor model. Alternatively, psychometric multidimensionality implies that the items within an instrument are potentially associated with more than one source of variance (Morin et al., 2016). They further note that psychometric multidimensionality may be construct relevant or irrelevant. The bifactor model's specification of a general factor that explains covariation among all items, and specific factors providing an estimate of shared variance among a subset of items once common variance from the

general factor has been removed, lends itself as particularly useful for modeling psychometric multidimensionality.

From a psychometric perspective, many educational and psychological instruments are designed to measure constructs that are hierarchical in nature, in that the structure is such that there may be conceptually narrow subdomains encompassed by a general or global construct (Morin et al., 2016; Reise, 2012). In the development of instruments designed to capture such constructs, researchers often engage in a two-stage sampling process, identifying domains within a broad or general construct and then developing or sampling items within the domains for inclusion in the complete instrument (Gibbons et al., 2007). While the instrument is theoretically targeted to measure the overall construct, the item sampling process leads to multidimensionality (Gibbons et al, 2007; Morin et al., 2016). When this occurs, researchers must decide on the construct relevance of the multidimensionality and its implications for constructing composite scores and potentially subscale scores. For example, an instrument designed to measure school engagement may include items that relate to different aspects of engagement such as the behavioral, emotional, and cognitive components (Stefansson et al., 2016). Modeling the obtained data via a bifactor model would indicate the degree to which each item reflects the overall factor of school engagement as well as each potential subdomain, providing information regarding the use of an overall composite score and subscale scores. In this context, the potential psychometric multidimensionality is construct relevant, prompting the potential need to evaluate the use of subscale scores.

Alternatively, psychometric multidimensionality may be identified due to construct-irrelevant features. In this context, specific factors are often referred to as

nuisance dimensions (Cho et al., 2014). Reise et al. (2010) define nuisance dimensions as "factors arising because of content parcels that potentially interfere with the measurement of the main target construct" (p. 546). Within the IRT literature, such dimensions are often referred to as testlets. For example, reading comprehension exams often consists of reading passages with students asked to respond to a series of items based on the passage. While these items capture reading comprehension, they may also capture other information unrelated to reading comprehension due to their shared content, introducing additional sources of variance (DeMars, 2012). Cho et al. (2014) also note that differential response patterns may lead to nuisance dimensions, such as the case with timed exams, which can cause items towards the end of an exam to inaccurately appear more difficult. Other examples of differential response patterns that can potentially lead to nuisance dimensions is when both positively and negatively worded items are included in an instrument. Item wording effects caused by the inclusion of positively and negatively worded items are one of the most frequently encountered sources of construct irrelevant psychometric multidimensionality (Morin et al., 2020). In this context, bifactor models can be applied to model a general construct of interest in the presence of method effects caused by item wording direction (Bollen & Hoyle, 2012; Markon, 2019; Schmitt et al., 2018).

Model Assumptions and Specification

In a restricted bifactor model (Gibbons & Hedeker, 1992; Reise et al., 2010; Stucky & Edelen, 2015) two primary assumptions are required. One, it is assumed there is a pattern of zero and non-zero loadings between the indicators, the general factor, and the specific factors (Canivez, 2016; Markon, 2019). All indicators are assumed to have

non-zero loadings on the general factor, and one specific factor, with zero loadings set between the items and specific factors in which these items do not load. This loading pattern also holds important implications for the meaning of the latent factors in the bifactor model, in that the general factor is assumed to broadly characterize the relationships between all the indicators, while specific factors represent more defined patterns among the indicators that may be the result of subsets of related content or item features (DeMars, 2013; Reise, 2012).

Second, it is assumed that the latent variables in a bifactor model are orthogonal. The assumption of orthogonal relationships between the latent variables is critical. Constraining the general and specific factors to be orthogonal allows the item variance to be partitioned into these two sources, and for the specific factors to be interpreted as residual factors (Reise, 2012). While it is preferred for the specific factors to also be orthogonal, this is not a requirement. However, Reise (2012) cautions that correlations between specific factors suggests that additional unmodeled general factors are present, which may lead to issues with model interpretability and applicability.

Both assumptions are clearly seen considering the matrices contained within the CFA measurement model and model-implied variance-covariance matrix required to estimate model parameters. To facilitate interpretation of both the measurement model and the model-implied variance-covariance matrix, consider a 12-item instrument with a balanced scale (i.e., six positively worded and six negatively worded items). A bifactor measurement model specification would imply that all items load on the general factor, all positively worded items (i.e., items 1-6) simultaneously load on a specific factor representing the positively worded item method factor, and all negatively worded items

(i.e., items 7-12) simultaneously load on a specific factor representing the negatively worded item method factor. Observed item responses can be decomposed according to Equations 1 and 2 (adapted from Chen et al., 2006; Liu & Thompson, 2021).

$$X = \Lambda_x \xi + \delta, \tag{1}$$

where X is 12 x 1 vector of observed variables, Λ_x is a 12 x 3 matrix representing the factor loadings on the general and specific factors, ξ is a 3 x 1 vector representing the general and specific latent factors, and δ is a 12 x1 vector representing residual variance (i.e., uniqueness).

$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \\ x_{11} \\ x_{12} \end{bmatrix} \; \Lambda_x = \begin{bmatrix} \lambda_{G1,1} & \lambda_{Sp1,1} & 0 \\ \lambda_{G2,1} & \lambda_{Sp2,1} & 0 \\ \lambda_{G3,1} & \lambda_{Sp3,1} & 0 \\ \lambda_{G4,1} & \lambda_{Sp4,1} & 0 \\ \lambda_{G5,1} & \lambda_{Sp5,1} & 0 \\ \lambda_{G6,1} & \lambda_{Sp6,1} & 0 \\ \lambda_{G7,1} & 0 & \lambda_{Sn7,2} \\ \lambda_{G8,1} & 0 & \lambda_{Sn8,2} \\ \lambda_{G9,1} & 0 & \lambda_{Sn9,2} \\ \lambda_{G10,1} & 0 & \lambda_{Sn10,2} \\ \lambda_{G11,1} & 0 & \lambda_{Sn11,2} \\ \lambda_{G12,1} & 0 & \lambda_{Sn12,2} \end{bmatrix} \; \xi = \begin{bmatrix} \xi_G \\ \xi_{Sp} \\ \xi_{Sn} \end{bmatrix} \; \delta = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \\ \delta_7 \\ \delta_8 \\ \delta_9 \\ \delta_{10} \\ \delta_{11} \\ \delta_{12} \end{bmatrix}. \tag{2}$$

As shown in Equation 2, the Λ_x matrix specifies that factor loadings will be estimated for the general factor across all items, items 1-6 for the specific positive item wording method factor, and items 7-12 for the specific negative item wording factor. All other factor loadings are constrained to zero.

The model implied covariance matrix for this example is presented in Equation 3 below.

$$\Sigma = \Lambda_x \Phi \Lambda_x' + \Theta_\delta , \tag{3}$$

where the Λ_x matrix is defined the same as above, with Λ_x' the transposed factor loading matrix, and Θ_δ is a 12x12 variance-covariance matrix of the residuals. This matrix is

symmetric, with residual variances on the diagonal and covariances on the off-diagonal (Brown, 2015). When residual correlations are constrained to zero, as they are in this example, the off-diagonal consists of all zeros. The Φ matrix is a 3x3 variance-covariance matrix for the latent factors. In a bifactor model, the Φ matrix is specified according to Equation 4 below.

$$\Phi = \begin{bmatrix} \phi_G & & \\ 0 & \phi_{Sp} & \\ 0 & 0 & \phi_{Sn} \end{bmatrix}. \tag{4}$$

The elements of the Φ matrix are key to a bifactor model given the orthogonality assumption. In the Φ matrix as defined above, the factor covariances are constrained to zero (i.e., elements of the off-diagonal), with only factor variances estimated (i.e., the diagonal). When the factor variance is set to one for identification (or in a completely standardized solution), this diagonal becomes a series of all ones (Brown, 2015). As a result, the elements of the model-implied variance-covariance matrix are estimated with Equations 5-7 below. Continuing with the 12-item instrument example, assume Item 1 and Item 2 both load on the general factor as well as the positively worded item method factor. Item 7 also loads on the general factor but the negatively worded item method factor.

$$\sigma_{x1}^2 = \lambda_{1G}^2 + \lambda_{1Sp}^2 + \sigma_{e1}^2 \,. \tag{5}$$

$$\sigma_{x1,x7} = \lambda_{1G}\lambda_{7G} \,. \tag{6}$$

$$\sigma_{x1,x2} = (\lambda_{1G}\lambda_{2G}) + (\lambda_{1Sp}\lambda_{2Sp}) \,. \tag{7}$$

The variance of each item (Equation 5) is estimated as the sum of its squared factor loadings on both the general factor and its specific factor plus the item's residual

component. In other words, the variance explained by the general factor, the variance explained by the specific factor, and the residual variance (Gu et al., 2017).

Decomposing the covariances between items (or correlations in completely standardized solutions) is somewhat more complex. With the orthogonality constraint placed on all latent factors, item covariances (or correlations) are a function of the item's relationship between the general and its specific factor. In the above example, Item 1 and Item 7 both load on the general factor, but different specific factors. Therefore, the only element that defines their relationship are the item loadings on the general factor (Equation 6). In other words, the correlation between these items reflects variance due to only the general factor (Bonifay et al., 2015; Reise, 2012; Reise, Scheines, et al., 2013). Item 1 and Item 2 both load on the general factor and the same specific factor. In this situation, their estimated covariance is a function of their relationship between *both* the general factor and the specific factor. In situations in which the general factor is of primary interest, and possibly the use of a composite score, the influence of the specific factor on these items is a potentially contaminating factor (Bonifay et al., 2015; Reise, 2012; Rodriguez et al., 2016a, 2016b). In this situation, it is advantageous to have fewer items that are impacted by both the general and specific factor variance (i.e., fewer items contained within each specific factor; Bonifay et al., 2015; Reise, Scheines, et al., 2013).

Relationship to the Unidimensional Model

Reise (2012) specifies that by constraining the specific factor loadings in a bifactor model to zero, the unidimensional model is obtained. In other words, the models are nested, with the unidimensional model nested within the bifactor model. A comparison between a fitted unidimensional model and a bifactor model offers critical

information regarding the relationship between these two models, or more specifically the general construct as defined by these two models. Reise notes that the degree to which factor loadings from the general factor in a bifactor model and loadings estimated from a unidimensional model are consistent provides evidence that the latent variable captured by both models is equivalent. If items load more strongly on the general factor relative to the specific factor, this is one indication that the items primarily reflect the general construct and ignoring the multidimensionality from the specific factors would not lead to severe bias by treating the data as unidimensional (Reise et al., 2010; Reise et al., 2015). Importantly, such comparisons between the bifactor model and the unidimensional model only hold if the data follow a bifactor structure (Reise et al., 2007). Data is best modeled via a bifactor model when it theorized that item responses mostly reflect a general or common trait, but multidimensionality is caused by well-defined item clusters (Reise et al., 2007; Reise, 2012; Rodriguez et al., 2016a). If a bifactor model can reproduce relationships in the data, then the comparisons between the estimated loadings on the general construct between the two models offers information regarding the degree to which parameters may be biased or distorted by fitting a unidimensional model to multidimensional data (Reise et al., 2007).

Evaluating Model-Data Fit

Assessing model-data fit is one of the most important elements in conducting CFA. However, it is also one of the more controversial elements (Brown, 2015; Raykov & Marcoulides, 2006). Model-data fit can be assessed using a variety of indices and information that capitalize on different elements of the model. Overall, fit is assessed at the global and local level in a CFA model. This includes an assessment of the model's

overall goodness of fit, as well as individual parameter estimates (Byrne, 2005). While model fit statistics provide critical information, decisions regarding model fit need to also be made on substantive grounds (Raykov & Marcoulides, 2006). Ultimately, researchers are focused in identifying a model that is, (1) theoretically sound, (2) parsimonious, and (3) corresponds to the data (Kline, 2011).

Assessing Model-Data Fit in Bifactor Models

In assessing model-data fit, researchers determine how well a theorized model fits the data, or in situations in which multiple alterative models are evaluated, which of the models provides the *best* fit to the data (West et al., 2012). Bifactor models are rarely evaluated in a strictly confirmatory (i.e., Jöreskog, 1993) context, in which it is the only model investigated. Instead, they are typically evaluated against several alternative models. Broadly, research investigating an instrument's dimensional structure that includes a bifactor model specification usually includes an evaluation of a unidimensional model, a correlated-factors model, and a higher-order or second-order model (e.g., Bear et al., 2011; Dierendonck et al., 2020; Gordon et al., 2020; Huang et al., 2014; Kopp et al., 2011; Stefansson et al., 2016). Within the context of evaluating item wording method effects, researchers apply a similar approach, although fewer studies have utilized the higher-order model in this context (see McKay et al., 2014, for an example). Instead, a more common approach is to also include a correlated uniqueness (CU) model in which correlated errors between items of similar wording are modeled. Together, the unidimensional model, correlated-factors model, correlated uniqueness model, and the bifactor model form what Marsh et al. (2010) refer to as a "taxonomy of models" (p. 379). Each of these models offers a different conceptualization of the method

effect arising from the item wording. Marsh et al. stress that evaluating each of these models in a taxonomic strategy allows for a more complete assessment of the nature and size of the item wording method effects.

Several procedures exist for evaluating alternative CFA models. In the context of evaluating item wording method effects, the most common approach is to take a complete examination of multiple fit indices and compare between models. In this context the overall fit of each model and the relative fit between models is often used to select the best fitting model structure. In this procedure, researchers typically evaluate fit statistics from different classes of model fit. This typically includes an assessment of each model's chi-square (χ^2) statistic, approximate fit indices representing both absolute (i.e., RMSEA and SRMR) and comparative fit (i.e., CFI and TLI), and an evaluation of model selection indices (i.e., AIC and/or BIC).

Model Fit Indices. The chi-square (χ^2) statistic, under maximum likelihood (ML) estimation, provides the only statistical test of significance for assessing model-data fit (Schumacker & Lomax, 2010). Chi-square is a test of exact fit in that it tests the hypothesis that the population covariance matrix and the reproduced or model-implied covariance matrix are an exact fit (DiStefano, 2016; West et al., 2012). A chi-square value of zero indicates perfect fit between the model implied covariance matrix (Σ) and the sample covariance matrix (S) (Schumaacker & Lomax, 2010). In this context, the goal is to retain the null hypothesis. A non-significant p-value indicates the data fits the model. However, the chi-square statistic is highly sensitive to sample size and is therefore not recommended for use as the sole index to make global model-data fit decisions (Brown, 2015; DiStefano, 2016; Raykov & Marcoulides, 2006).

It is unlikely that any CFA or SEM model specified will provide an exact fit to data structures encountered in real-world applications. As such, several approximate or alternative model fit indices have been developed to aid researchers in reaching decisions regarding model-data fit. Fit indices are often selected and evaluated together because they represent different aspects of model fit (i.e., comparative fit, close fit). Broadly, alternative fit indices can be characterized as absolute or comparative/incremental fit indices (Brown, 2015; DiStefano, 2016; West et al., 2012). Absolute fit indices (i.e., RMSEA, SRMR, AIC, BIC) evaluate the discrepancy between the model input and model implied covariance matrices (DiStefano, 2016). Comparative fit indices (i.e., CFI and TLI) assess model-data fit by comparing fit from the tested model and a baseline model (DiStefano, 2016). More specifically, comparative fit indices represent the proportion of improved fit in the tested model compared to a null model in which only the variances of the variables are estimated (i.e., covariances are not estimated; Brown, 2015; DiStefano, 2016; West et al., 2012). In selecting models that represent the "best fit" to the data, several suggested thresholds and cutoff values for these fit indices have been recommended. Those most frequently applied throughout the CFA/SEM literature come from Hu and Bentler (1998; 1999), including CFI and TLI $\geq$.95, RMSEA $\leq$.06, and SRMR $\leq$.08.

The Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) are considered model selection indices and can be used to compare models that are either nested or non-nested (West et al., 2012). Models are considered nested when a more restricted model (i.e., one with fewer parameters estimated) can be derived from a less restricted model (i.e., more model parameters estimated) by fixing parameters from

the less restricted model (Schermellah-Engel et al., 2003). The individual values of both the AIC and BIC index are non-informative when evaluated on their own (DiStefano, 2016). It is the comparison of values between models that provides information regarding model selection, with lower values of AIC and BIC indicating better model fit (DiStefano, 2016).

The cutoff values for the approximate fit statistics, such as those suggested by Hu and Bentler (1999), are often characterized as "golden rules" within the CFA and SEM literature in the sense that if a series of indices meets the recommended cutoff, the model is selected and declared the "best fitting." However, the conditions in which these cutoffs were set were limited, with several researchers acknowledging that, for example, Hu and Bentler only intended for their guidelines to be considered general recommendations (Kline, 2011; Marsh et al., 2004). A number of researchers have raised concerns about the dichotomy of "good fit"/"bad fit" that exists when evaluating model fit via approximate fit indices (e.g., Barrett, 2007; Marsh et al., 2004). Nevertheless, their application to examining model-data fit are standard practice.

Concerns with Model-Fit Indices and Bifactor Models. In the context of research involving bifactor models more generally, several researchers have noted examining CFI, TLI, RMSEA, AIC, and BIC as they favor model parsimony (e.g., Gignac & Watkins, 2016; Morgan et al., 2015; Murray & Johnson, 2013). These indices include a penalty for model complexity as the model's degrees of freedom or the number of free parameters is considered in its calculation (i.e., models that estimate more parameters are less parsimonious/more complex; Brown, 2015; Kline, 2011). Parsimony is an important feature in measurement, in that a simpler or less complex model, provided

it fits the data and is theoretically sound, is preferred (Kline, 2011). Compared to the unidimensional model, the correlated-factors model, and the higher-order model, the bifactor model represents the most complex, or least parsimonious model. However, recent simulation studies suggest that model fit indices that include a penalty for model complexity may be biased in favor of the bifactor model.

Murray and Johnson (2013) hypothesized that the bifactor model, with its additional parameters, may absorb unmodeled complexities such as cross-loadings and correlated residuals when they are present in data structures. In such cases, the bifactor model has been shown to exhibit better model-data fit compared to either a higher-order or correlated-factors model, despite not being the true underlying structure (Greene et al., 2019; Murray & Johnson, 2013). Reise et al. (2016) further note that "unless data are perfectly consistent with a model…fit indices will be biased toward the model with more parameters" (p. 3). In other words, fit indices tend to favor the bifactor model with its additional parameters. Additionally, in some cases, model fit indices have provided poor discrimination between the correlated-factors, higher-order, and bifactor model when no additional model complexities were included (Morgan et al., 2015). Such findings indicate that selection of a model must be based on substantive evaluation and judgement.

Bonifay and Cai (2017) recently examined the fit propensity of the bifactor model compared to other model structures, including an exploratory factor analytic model and a unidimensional model. Fit propensity refers to a model's average ability to fit diverse patterns of data (Preacher, 2006). They found the bifactor model's fit propensity to be almost as high as an exploratory model, indicating the bifactor model was able to fit multiple types of data structures. Conversely, the unidimensional model's fit propensity

was low, indicating that it rarely fit diverse data patterns. These findings led Bonifay and Cai to conclude that the bifactor model, in some instances, overfits the data. Furthermore, they cautioned researchers against relying solely on standard model fit statistics in making decisions, noting that the bifactor model is likely to be chosen as the best fitting. In such situations, application of additional model information, such as model-based reliability indices, may be informative in evaluating the bifactor model's utility (i.e., Bonifay et al., 2017; Reise et al., 2016).

Alternative Model-Based Indices

A pure inspection of factor loadings between the items and the general factor and the specific factors provides information in relation to both the strength of the general factor and specific factors in a bifactor model (Reise et al., 2010). However, to fully evaluate the structure of these latent factors, researchers recommend the application of several model-based indices to provide a more comprehensive evaluation of the bifactor model. These indices help researchers assess the strength of the general and specific factors, the degree to which total scores reflect the intended construct, and how reliable and applicable subscale scores are from group factors (Rodriguez et al., 2016a, 2016b). The use of recommended model-based indices can assist researchers in evaluating the functioning of the general factor (i.e., is it strong enough to support essential unidimensionality) and provide support for the interpretation (or lack thereof) of potential subscales by providing information concerning the reliability of these factors. These indices include, omega, omega hierarchical, factor determinacy (*FD*), construct replicability (*H* index), explained common variance (ECV), and the percent of uncontaminated correlations (PUC). Each of these statistics are considered model-based

indices in that information from the model (i.e., factor loadings) is used in calculating each index. Additionally, the omega reliability indices and ECV can be computed for both the general factor and the specific factors. However, all descriptions below concern their application to the general factor to assess its strength and reliability in relation to the specific factors, as general factor strength is of primary relevance to the current study.

Each of the recommended model-based indices have been available for use within the CFA and SEM framework for some time, but their use in both methodological and applied contexts is relatively sparse. One reason may be the lack of availability as an option across readily used statistical software (i.e., SAS, SPSS; Gignac, 2016). As noted by Rodriguez et al. (2016b), with the exception of the PUC index, these indices are not necessarily bifactor-specific and can be estimated with other model types. However, their applicability within the context of bifactor modeling yields precise information relevant to the psychometric evaluation of both the general and specific factors. Furthermore, Savalei & Reise (2019) stress that as these indices are model-based, the appropriate latent model structure must be taken into account. For example, while omega total can be estimated for any multidimensional latent model, it is best applied to data that follow a hierarchical structure (Watkins, 2017). Additionally, as a bifactor model partitions item variance into two common sources, it is likely the only viable model in which omega hierarchical can be accurately computed and interpreted (Canivez, 2016; Reise, Bonifay, et al., 2013; Savalei & Reise, 2019).

Omega Reliability Indices (omega total ω_t; omega hierarchical ω_h). In general, internal consistency is among the most frequently types of reliability reported. Internal consistency refers to the degree of consistency, or item homogeneity, across responses to

a set of items designed to measure the same construct (Crocker & Algina, 2008). Coefficient alpha (Cronbach's alpha) is frequently calculated and reported as a measure of internal consistency reliability. The alpha coefficient operates under similar assumptions to classical test theory (CTT): (1) unidimensionality, (2) tau equivalency (i.e., equal factor loadings or assuming each item contributes equally to total scores), (3) items are continuous and normally distributed, and (4) uncorrelated errors (McNeish, 2018). In practice, these assumptions are quite rigid and may be difficult to meet, however, alpha continues to be used rather frequently (Gignac, 2016; McNeish, 2018).

More recently, researchers have argued for using SEM based reliability coefficients in place of coefficient alpha (e.g., McNeish, 2018; Green & Yang, 2009; Yang & Green, 2011). SEM based reliability coefficients recognize the data's structure (i.e., dimensionality, item relationship to the construct) and make use of model-based information, relaxing many of the assumptions of coefficient alpha. Many scales within the educational and psychological fields, although designed to be unidimensional, often do not fit a unidimensional model (Flora, 2020). As described by Morin et al. (2016) this multidimensionality may be the result of psychometric multidimensionality, the hierarchical nature of many constructs, and may or may not be construct relevant. The use of alpha as an estimate of reliability may be inappropriate in this context. When data follow a hierarchical structure, or one in which there is a general factor and multiple specific factors that explain the covariation among items, this violates many of the assumptions of the alpha coefficient, including unidimensionality and uncorrelated errors. In contrast, the omega family of reliability coefficients provides a more accurate estimate

of reliability by considering the data's multidimensional structure and the multiple sources of common and unique variance in obtained composite scores (Watkins, 2017).

Coefficient omega (McDonald, 1999), also known as omega total (Revelle & Zinbarg, 2009), is an estimate of internal consistency. However, unlike alpha, omega total does not assume tau equivalence (i.e., equal factor loadings). Instead, items are assumed to be congeneric in that each item measures the same construct, but perhaps to a different degree (i.e., unequal factor loadings; McNeish, 2018). Omega total estimates the proportion of variance in the observed, unit-weighted total score that is attributable to all modeled sources of common variance or multiple common factors (Rodriguez et al., 2016a, 2016b). A unit-weighted score refers to a total score that is calculated by summing the raw scores for each item, where each item receives an equal weight (McNeish & Wolf, 2020).

There are multiple ways to calculate omega total which differ depending on if a researcher is utilizing a bifactor model in an exploratory or confirmatory framework, and if the factor pattern analyzed is based on the analysis of covariances or correlations (Reise, Bonifay, et al., 2013). In a completely standardized solution, using standardized factor loadings estimated from a bifactor model, omega total is calculated according to Equation 8 (Reise, Bonifay, et al., 2013; Rodriguez et al., 2016a).

$$\omega_t = \frac{\sum \lambda_{iG})^2 + \sum_{k=1}^{K}(\sum \lambda_{iSk})^2}{\sum \lambda_{iG})^2 + \sum_{k=1}^{K}(\sum \lambda_{iSk})^2 + \sum(1-h^2)}, \tag{8}$$

where $\sum \lambda_{iG})^2$ is the squared sum of the standardized factor loadings on the general factor, $\sum_{k=1}^{K}(\sum \lambda_{iSk})^2$ is the squared sum of the standardized factor loadings for each specific factor, and $\sum(1 - h^2)$ is the sum of the item uniqueness values.

Omega values range 0 to 1.0, with higher omega total values (i.e., closer to 1.0) indicating a highly reliable multidimensional composite; however, as this value represents a mixture of both general and specific factor variance, the precision of total and subscale scores as unique representations of the general or specific factors cannot be separated (Canivez, 2016; Watkins, 2017). Furthermore, if the primary interest is in the reliability of the general factor, taking specific factors into account is unnecessary (Kelley & Pornprasertmanit, 2016; Savalei & Reise, 2019). If data follow a bifactor structure, omega hierarchical may be a more informative reliability coefficient (Canivez, 2016; Savalei & Reise, 2019).

Contrasted with omega total, omega hierarchical (omegaH; McDonald, 1999; Zinbarg et al., 2005) estimates the proportion of variance in unit-weighted total scores attributable only to the general factor (Rodriguez et al., 2016a, 2016b). OmegaH is a valuable index in understanding if a composite score's reliability is mostly due to a single construct in the context of multidimensionality (Reise, Bonifay, et al., 2013; Reise, Moore, et al., 2010). In the calculation of omegaH, variance in total scores from specific factors is treated as measurement error (Reise, Moore, et al., 2010; Rodriguez et al., 2016b). Unlike omega total, omegaH allows researchers to assess the strength of the general factor independent of specific factors (Gignac, 2016; Rodriguez et al., 2016a, 2016b).

Similar to omega total, there are several ways to calculate omegaH which also differ depending on if a researcher is utilizing a bifactor model in an exploratory or confirmatory framework, and if the factor pattern analyzed is based on the analysis of covariances or correlations (Reise, Bonifay, et al., 2013). In a completely standardized

solution, using standardized factor loadings estimated from a bifactor model, omegaH can be calculated according to Equation 9 (Gignac, 2016; Reise, Bonifay, et al., 2013; Rodriguez et al., 2016a).

$$\omega_h = \frac{(\sum \lambda_{iG})^2}{(\sum \lambda_{iG})^2 + \sum_{k=1}^{K}(\sum \lambda_{iSk})^2 + \sum(1 - h_i^2)}, \tag{9}$$

where $(\sum \lambda_{iG})^2$is the squared sum of the standardized factor loadings on the general factor, $\sum_{k=1}^{K}(\sum \lambda_{iSk})^2$ is the squared sum of the standardized factor loadings for each specific factor, and $\sum(1 - h_i^2)$ is the sum of the item uniqueness values. The numerator in the calculation of omegaH only contains the squared sum of the loadings for the general factor, hence the specific factors are treated as measurement error. OmegaH values range 0 to 1.0, with higher omegaH values indicating that the general factor is the primary source of variance in unit-weighted scores (Rodriguez et al., 2016a, 2016b).

Factor Determinacy (*FD*) and Construct Replicability (*H* index). A bifactor model specifies that all items load on a general factor, with subsets of items simultaneously loading on a specific factor. As specific factors are often characterized as factors with substantially fewer items and potentially lower loadings, the functioning of specific factors and the reliability of subscale scores obtained from these factors is critical (Rodriguez et al, 2016a, 2016b). Rodriguez et al. (2016a, 2016b) propose the use of factor determinacy (*FD*) and construct reliability or replicability (*H* index) to assess the functioning and stability of the specific factors. Unlike the omega coefficients, which provide evidence for the reliability of unit-weighted (observed) composite scores, *FD* and the *H* index provide evidence for the reliability of optimally weighted latent composite scores and the stability of the specific factors for use in the SEM framework (Rodriguez et al., 2016a, 2016b).

For any given set of items, there exists multiple ways to obtain scores relating an individual's placement on the construct(s) of interest. Broadly, methods of obtaining factor score estimates fall into one of two categories, refined and non-refined methods, which differ based on the technical methods used to obtain the scores (DiStefano et al, 2009). For example, unit-weighted summed scores are considered non-refined or coarse factor score estimates, while scores obtained using regression methods that take into account factor loadings are considered refined (DiStefano et al., 2009; Grice, 2001). A concern with using refined methods to obtain factor scores is that, in theory, there exists an infinite number of ways of obtaining scores that are consistent with the factor loadings, a problem known as indeterminacy (Grice, 2001). The lack of a unique solution means the adequacy of the obtained scores can vary. The proposed *FD* index assesses the degree to which factor score estimates are determinate and provide an adequate representation of individuals on the construct (Rodriguez et al, 2016a, 2016b). The *FD* index is calculated according to Equation 10 (Rodriguez et al., 2016a, 2016b).

$$FD = diag(\Phi\Lambda^{\mathrm{T}}\Sigma^{-1}\Lambda\Phi)^{1/2}, \qquad (10)$$

where Φ is a *m* x *m* matrix of factor intercorrelations. Recall from Equation 4 presented earlier that in a bifactor model, the Φ matrix consists of a series of ones on the diagonal, and zeros on the off-diagonal. The Λ is a *k* x *m* matrix of the standardized factor loadings with *k* the number of items and *m* the number of factors. Finally, the element Σ is the model-implied variance-covariance, defined as previously described in Equation 3 ($\Sigma = \Lambda_x \Phi \Lambda_x' + \Theta_\delta$).

The *FD* index provides the correlation of the factor scores with the factors, with values ranging 0 to 1.0 (Rodriguez et al, 2016a, 2016b). Higher values, or those closer to

1.0, suggest better determinacy (Rodriguez et al, 2016a, 2016b). Rodriguez et al. recommend the cutoff suggested by Gorsuch (1983) of .90 or greater as indicating when factor scores should be considered from factors obtained in the bifactor model.

Rodriguez et al. (2016a, 2016b) propose the use of the *H* index (Hancock & Mueller, 2001) to assess the stability and replicability of the latent factors in the bifactor model. Again, as the specific factors are associated with fewer items and typically lower factor loadings, this index is particularly useful in evaluating these factors. The *H* index provides a way to evaluate how well a set of items represents the latent variable (Rodriguez et al., 2016a). Specifically, Rodriguez et al. note that the *H* index is useful for establishing which measurement model, and which factors, should be included in subsequent latent analyses via a full SEM model. The *H* index is calculated according to Equation 11 (Rodriguez et al, 2016a, 2016b).

$$H = 1/\left[1 + \frac{1}{\sum_{i=1}^{k}\frac{\lambda_i^2}{1-\lambda_i^2}}\right], \tag{11}$$

where λ_i^2 is the squared factor loading for the items on the factor being evaluated, and $1 - \lambda_i^2$ is the item's residual or uniqueness value. The *H* index ranges from 0 to 1, with higher values (closer to 1) indicating the latent variable is well defined by the item set and is likely to be stable and replicated across further studies (Rodriguez et al., 2016a, 2016b). Rodriguez et al. (2016a, 2016b) recommend a minimum *H* index value of 0.70 (as suggested by Hancock & Mueller, 2001) as a minimum cutoff indicating the latent construct is stable enough for inclusion in a full SEM model.

Explained Common Variance (ECV). Technically, ECV is the ratio of the first eigenvalue (i.e., the common factor) to the sum of all eigenvalues (Bentler, 2009; Ten

Berge & Socan, 2004). Ten Berge & Socan (2004) further describe ECV as a coefficient that represents "closeness to unidimensionality" (p. 621). In the context of a bifactor model, ECV assesses the strength of the general factor relative to the specific factors and is considered a direct indicator of unidimensionality (Rodriguez et al., 2016a, 2016b). As such, ECV is considered a unidimensionality statistic (Reise, 2012). When the orthogonality constraint (i.e., uncorrelated factors) is placed on the general and specific factors, ECV provides the variance explained by the general factor by taking the ratio of variance explained by the general factor (i.e., the eigenvalue of the general factor), divided by the variance explained by both the general and specific factors (Reise, 2012; Rodriguez et al., 2016b).

Using the standardized factor loadings estimated from a bifactor model, ECV is calculated according to Equation 12 (Rodriguez et al., 2016a).

$$ECV = \frac{(\sum \lambda_{iG}^2)}{(\sum \lambda_{iG}^2) + \sum_{k=1}^{K}(\sum \lambda_{iSk}^2)}\ , \tag{12}$$

where $(\sum \lambda_{iG}^2)$ are the sum of the squared loadings for the general factor (i.e., the eigenvalue of the general factor) and $\sum_{k=1}^{K}(\sum \lambda_{iSk}^2)$ are the sum of the squared factor loadings for each of the specific factors. For interpretation, ECV values range 0 to 1.0, with values closer to 1 indicating the general factor is likely "strong" enough to be treated as unidimensional in subsequent analyses (Reise, 2012). ECV is a valuable index for evaluating what model can be specified as the measurement model in a full SEM analysis (Rodriguez et al., 2016a).

Contrasted with omegaH, ECV is a direct indicator of dimensionality, while omegaH indicates that observed scores can be considered essentially unidimensional if the reliable variance is mostly due to the general factor (Rodriguez et al., 2016b). In this

sense, ECV may be thought of as an index of general factor strength, while omegaH may be considered an index of general factor saturation (Reise, Scheines, et al., 2013; Zinburg et al., 2006).

Percent of Uncontaminated Correlations (PUC). PUCs are the percent of item correlations (i.e., covariances) that only reflect general factor variance. More specifically, these are the correlations between items from different specific factors, as item correlations *within* each specific factor are contaminated by both specific and general factor variance (Rodriguez et al, 2016b). In a bifactor model, a general factor is specified to explain the covariation among all items, while smaller specific factors explain additional variance among a subset of items. Therefore, items from different specific factors are only influenced by their shared relationship to the general factor. As was detailed earlier in this chapter via Equations 6 and 7, the correlations between items are derived from only their factor loadings on the general factor. Conversely, items within specific factors are influenced by both their relationship to the general and the specific factor. Correlations between these items are derived from their factor loadings on both the general and specific factors. As such, these correlations are inflated due to being influenced by variance from both the general and specific factors (Bonifay et al., 2015).

PUC is calculated using Equation 13 (adapted from Bonifay et al., 2015; Rodriguez et al, 2016a).

$$PUC = 1 - \frac{\sum_{k=1}^{K}\left(\frac{p_{Sk}(p_{Sk}-1)}{2}\right)}{\left(\frac{p_{Total}(p_{Total}-1)}{2}\right)}, \tag{13}$$

where p is equal to the number of items; the numerator is the sum of the number of correlations within each specific factor and the denominator is the total number of correlations.

Notably, compared to omega hierarchical and ECV, PUCs are not influenced by the magnitude of the factor loadings on either the general factor or the specific factor (Liu et al., 2022). Instead, PUCs are influenced by model size, specifically the number of total items, number of specific factors, and number of items within each specific factor (Reise, Scheines, et al., 2013). As the number of specific factors increases and the number of items within each factor decreases, PUC increases, as fewer item correlations are contaminated by both the specific and general factor. For example, a model with 12 items loading on a general factor, and six items each across two specific factors represents a PUC of 0.54, indicating that a slight majority of item correlations, 54%, reflect general factor variance only. Alternatively, a model with 12 items loading on a general factor, and three items each across four specific factors represents a PUC of 0.82, indicating that strong majority, 82%, of items correlations reflect only the general factor variance. As described by Rodriguez et al. (2016b) "…as PUC increases, the general trait in the bifactor model becomes more and more similar to the single trait estimated in a unidimensional model…" (p. 145). As such, PUC, much like ECV, is one way of quantifying the unidimensionality of the data's structure.

Bifactor Model-Based Indices and Applications to Essential Unidimensionality

Given the model's parsimony and simplicity, unidimensionality is a desirable property in measurement. By definition, unidimensionality exists when all covariation among items contained within a scale is explained by the latent construct, with no correlated residuals, and all measurement error is random (Brown, 2015; Ziegler & Hagemann, 2015). The simplicity of the unidimensional model is also its downfall. Results from Bonifay & Cai (2017) highlight the rigidity of the unidimensional model

and show that the assumption of strict unidimensionality may be difficult to meet with real data. Any minor violations or deviations in unidimensional structure (e.g., correlated residuals) will lead to poor fitting models when data, which are theorized to follow a unidimensional structure, are fit to a unidimensional measurement model. In this regard, Reise et al. (2010) argue that the bifactor model may provide a more accurate or realistic conceptualization of unidimensionality by specifying a general factor that influences all items, but which may be contaminated by other, perhaps nuisance, factors (i.e., method effects). A more appropriate approach may not be to assess if the data are strictly unidimensional, but to assess to what degree data are unidimensional (Reise et al., 2016).

Several of the model-based indices recommended for use with bifactor models have been proposed for use in establishing essentially unidimensional models. In the context of a bifactor model, the degree to which the data can be considered as "essentially unidimensional" is related to the extent to which the general factor is considered "strong" in relation to the specific factors and accounts for the majority of covariation in item responses (Donnellan et al., 2016; Raykov & Pohl, 2012, 2013). As indices of general factor strength and unidimensionality, omega hierarchical (omegaH), explained common variance (ECV), and the percent of uncontaminated correlations (PUC), provide researchers with evidence for when data can be treated as essentially unidimensional despite multidimensionality (Reise, Bonifay, et al., 2013; Rodriguez et al., 2016a, 2016b).

Several guidelines regarding omegaH, ECV, and PUC have been proposed and recommended for researchers to assess essential unidimensionality, although there appears to be some hesitancy about recommending definitive values. As model-based

indices, their values are dependent on the quality of the underlying model (Reise, Bonifay, et al., 2013). Reise, Bonifay et al. (2013), suggested an omegaH of at least 0.50 but closer to 0.75 is preferred, but stress if an instrument is designed to measure a single construct, omegaH should "be as high as possible" (i.e., closer to 1; p. 138). Following a systematic review of 50 studies across the psychological, personality, and assessment literature in which a bifactor model was identified as the "best fitting" structure, Rodriguez et al. (2016a) offered some preliminary guidelines and potential cutoff values. As part of this review, Rodriguez et al. calculated the model-based indices for each study using the reported standardized factor loadings. In addition to calculating the model-based indices, Rodriguez et al. also used a technique in which they reproduced each study's correlation matrix using the reported standardized factor loadings. They then subsequently fit a unidimensional model and examined the relative bias in factor loadings from the unidimensional model and reported loadings from the general factor in the bifactor model. Using this technique, they found when both ECV and PUC were greater than 0.70, relative bias between factor loadings estimated from a unidimensional model and the bifactor model was acceptable, indicating that the structure could be treated as unidimensional. More broadly, they recommended ECV in excess of 0.70 to 0.80 likely indicates bias from utilizing a unidimensional model will be minimal. The general consensus regarding benchmarks for when data can be treated as essentially unidimensional is that omegaH, ECV, and PUC should be "high." However, what constitutes "high" is likely be context dependent.

Simulation Studies Examining the Bifactor Model-Based Indices

Reise, Scheines, et al. (2013) were among the first to systematically examine the relationship among general factor strength indices (ECV, PUC, and omegaH) and the degree of structural coefficient bias (i.e., validity estimates) present when bifactor/multidimensional data are fit to a unidimensional model. In their study, Reise, Scheines, et al. varied the following conditions: relative strength of the general and specific factor loadings (i.e., the factor loading values), number of specific factors, and number of total items, as each of these conditions directly impacts the values of ECV, omega hierarchical (omegaH), and percent of uncontaminated correlations (PUCs). As such their study included a range of ECV (0.20 - 0.85), omegaH (0.33 - 0.86) and PUC (0.69 - 0.94) values. As expected, the results overall indicated that as ECV and omegaH increased, the bias in estimated structural coefficients (defined as bias of less than 10%) from utilizing a unidimensional measurement model became less severe. Additional findings from their study included, (1) ECV was more influential in influencing structural parameter bias (i.e., validity) than omegaH, and (2) PUC played an important role in moderating the effectiveness of ECV as an indicator of essential unidimensionality. For example, in their study, when PUC was high (e.g., 0.94, indicating that 94% of item correlations reflect only the general factor variance), ECV values as low as 0.20 were associated with minimal bias in estimated validity coefficients obtained from the unidimensional model. However, when PUC was low (e.g., 0.69), ECV needed to be approximately 0.60 for bias less than 10% in estimated validity. From this, Reise, Scheines, et al. established the following tentative benchmark, when PUC is greater than 0.80, the value of the strength indices becomes less important; however, when PUC is <

0.80, ECV >0.60, and omegaH >0.70, bias should be minimal, and a unidimensional model can likely be used. They note, however, that additional research is needed to assess the generalizability of this benchmark across other data conditions.

The findings from Reise, Scheines, et al. (2013) provide insight into the relationship between PUC and ECV, noting that when PUC is "low", ECV values need to be "high." As indicators of unidimensionality, it seems logical that both indices are strongly related and work in tandem to establish essential unidimensionality. As described by Rodriguez et al. (2016b), as PUC increases, the data structure becomes more unidimensional. As the structure takes a more unidimensional form, it would follow that the strength of the general factor (i.e., ECV) becomes less of a factor in establishing essential unidimensionality. Conversely, as PUC decreases, and the data structure takes a more multidimensional form, and the strength of the general factor becomes more critical in establishing essential unidimensionality.

In Reise, Scheines, et al.'s (2013) study, the lowest PUC value included was 0.69. In the context of item wording method effects, in which a bifactor model will have one general factor and two specific factors related to item wording direction, researchers will encounter PUC values that will be substantially lower than 0.69. In a related study, Bonifay et al. (2015) examined the relationship between ECV and PUC in the context of establishing essentially unidimensional models. Although one of the purposes of their study was to compare the effectiveness of ECV and PUC in this context to the DETECT statistic (Dimensionality Evaluation to Enumerate Contributing Traits; Zhang & Stout, 1999), their study offers a comprehensive examination of ECV and PUC in the context of fitting a unidimensional measurement model to bifactor data. Bonifay et al. included a

condition in which the PUC was 0.52, a similar value that might be expected in the context of item wording method effects. Similar to Reise, Scheines, et al., Bonifay et al. varied the strength of the general and specific factors by altering the factor loading patterns, the number of specific factors, and the number of total items to generate a variety of ECV and PUC values (they did not investigate omegaH in this context). Unlike Reise, Scheines, et al. who investigated structural coefficient bias (i.e., the structural model in SEM), Bonifay et al. investigated bias from a pure measurement model perspective. Specifically, they looked at the bias in factor loading estimates when data simulated from a bifactor model was fit to a unidimensional measurement model. Here, Bonifay et al. investigated how well the misspecified unidimensional model recovered the factor loadings from the general factor in a bifactor model. Conclusions from Bonifay et al. (2015) were similar to Reise, Scheines, et al., in that, as ECV increased, relative bias observed in estimated factor loadings when data was fit to a unidimensional measurement model, decreased (i.e., bias less than 10%), and this was moderated by PUC. As PUC increased, the strength of the value of ECV became less important in indicating relative bias in factor loading estimates. As noted, the lowest PUC condition in their study was 0.52, a similar PUC that might be expected in the context of item wording effects. In this condition, ECV needed to be approximately 0.70 for bias right at 10% in estimated factor loading values and 0.75 for bias less than 10%. This minimum threshold is slightly higher than what was set by Reise, Scheines, et al.'s tentative benchmark of when PUC is less than 0.80, ECV values of at least 0.60 are likely to indicate essential unidimensionality. It is unclear if this is perhaps related to differences in outcomes investigated (i.e., structural coefficients vs. factor loadings) or perhaps differences in low

PUC values investigated. Bonifay et al. do not offer a benchmark based on their results, only to state that "when PUC is low…relative strength of the first factor is a critical determinant of bias." (p. 509).

Simulation Studies Examining Item Wording Method Effects

The use of bifactor models in the context of item wording method effects has received limited attention within simulation research. Gu et al. (2017) simulated an item wording effect using a modified bifactor model, in which all items loaded on a general factor, but where only the negatively worded items loaded on a specific factor (i.e., only one method factor was included). Recent applications of this modified bifactor model have referred to it as the bifactor-(S-1) model (Eid et al., 2017; Gnambs & Schroeders, 2020). Similar to Reise, Scheines, et al. (2013) and Bonifay et al. (2015), Gu et al. investigated the impact of fitting a unidimensional model to data generated from a bifactor model (i.e., ignoring the item wording method effect). Specifically, Gu et al. investigated the impact on both reliability and validity estimates. Reliability investigated in this study included both omega total and omega hierarchical (omegaH), however unlike Reise, Scheines, et al.'s study, Gu et al. did not investigate the relationship between omegaH and parameter bias (i.e., bias in structural coefficients, factor loadings) from a fitting a unidimensional model. Instead, they focused on bias in reliability indices estimated from the unidimensional model. Design factors in Gu et al.'s study included variations in sample size, total number of items, items per specific factor (balanced vs. unbalanced), and factor loading patterns on the general and specific factor. The only model-based index from the bifactor model considered in their study in relation to its efficiency in indicating parameter bias (i.e., structural or validity coefficients) was ECV,

with population values ranging 0.33 to 0.92. Across conditions, they found an ECV of 0.75 or higher was associated with minimal bias (i.e., less than 10%) in estimated validity and reliability coefficients, indicating the item wording effect is likely not a concern in this situation, and utilizing a unidimensional model (i.e., ignoring the method effect) is sufficient. While this result is consistent with the broad recommendation provided by Rodriguez et al. (i.e., ECV in excess of 0.70 to 0.80 likely indicates essential unidimensionality; 2016a), both Reise, Scheines, et al. and Bonifay et al. concluded that PUC moderates the relationship between ECV and bias in validity. The role of PUC was not considered in Gu et al.'s study. Relatedly, ECV was only examined between, and not within conditions. Considering the total number of items, number of group factors, and items contained within each group factor, PUC values in Gu et al.'s study were calculated to be 0.76, 0.77, 0.90, and 0.91. Given the high PUC values of 0.90, and based on the findings from Reise, Scheines, et al., bifactor models with ECV related to the general factor as low as 0.20 should still produce minimal bias in validity estimates when fit to a unidimensional model. But this was not the case in Gu et al.'s study. For example, in one condition in which the ECV was 0.75, and the PUC was 0.91, Gu et al. reported negative bias in the estimated validity coefficient of -16.5%, well above the acceptable threshold of |10|%. As PUC was not included as part of Gu et al.'s analysis, this discrepancy in findings from Reise, Scheines, et al. was not examined. It is unclear if this is perhaps due to the unbalanced nature of the number of items contained within the group factors, or perhaps the modified bifactor model (i.e., only one specific factor modeled) used to generate the data.

As noted, Gu et al. also investigated bias related to omega hierarchical (omegaH). However, the study investigated the bias related to estimating omegaH from a unidimensional model, concluding that when ECV is less than 0.75, omegaH as estimated from a unidimensional model is severely biased. Furthermore, they offer the recommendation that if omegaH is sufficiently "large enough" as estimated from the unidimensional model, then a total observed score can be utilized as an estimate of the construct; no index is stated to quantify "large enough." In the context of Gu et al.'s study, a more accurate use of omegaH would have been similar to that of Reise, Schienes, et al. (2013), in using it as an indicator of essential unidimensionality when derived from the bifactor model.

Xia (2018) examined item wording method effects by simulating data from three separate models representing different conceptualizations of item wording method effects, (1) a correlated-factors model with two latent variables representing a positive wording factor and a negative wording factor, (2) a traditional bifactor model with one general construct and two specific factors representing item wording method factors, and (3) a modified bifactor model in which one specific factor was included related to the negatively worded items (i.e., bifactor-(S-1)). The data was then subsequently fit to each model as well as a unidimensional model. Xia did not conduct a direct analysis of the model-based indices derived from the bifactor model (ECV, omgeaH, and PUC) and their relationship to parameter bias when fit to a unidimensional model. Most relevant to the current study is Xia's findings related to data simulated from a traditional bifactor model with two specific method factors and fit to a unidimensional model. Findings related to bias in estimated validity from the misspecified unidimensional model are similar to

those of Bonifay et al. (2015), Gu et al. (2017), and Reise, Scheines, et al. (2013), in that at "high" ECV values, bias in in validity fell within an acceptable range. Although not directly stated, "high" in this study was an ECV of 0.80.

Overall, simulation studies examining the model-based indices ECV, omega hierarchical, and PUC agree that as these indices increase, data that follows a bifactor structure can be treated as unidimensional with minimal bias expected (Bonifay et al., 2015; Gu et al., 2017; Reise, Scheines et al., 2013). This finding is relevant both for modeling and scoring purposes, indicating that multidimensional bifactor data can be subsequently examined via a unidimensional measurement model and the use of a composite score is likely unbiased in certain contexts (i.e., Reise et al., 2010; Reise, Bonifay et al., 2013). However, what is also clear is that the threshold for what constitutes "high enough," particularly that of ECV, is context dependent and highly related to PUC, as was shown in both Bonifay et al. (2015) and Reise, Scheines et al. (2013). Furthermore, the relationship between PUC and ECV within the context of item wording effects has not been thoroughly evaluated, as PUC was not a factor considered in either Gu et al. (2017) or Xia (2018).

Relationship to Model Fit Indices

In addition to investigating the bifactor model-based indices, Reise, Scheines et al. (2013) included an assessment of the approximate model fit indices CFI, RMSEA, and SRMR. Prior to including a structural coefficient into the model, Reise, Scheines, et al., fit the unidimensional model to the data to obtain model-data fit specifically for the measurement model, acknowledging that fit of the measurement model is often evaluated prior to including structural paths or a full SEM model. Recognizing that in their study

the unidimensional model was misspecified, their purpose in investigating model-data fit was not so much to examine if the unidimensional model fit the data, but to examine if model fit can be used as a proxy for structural parameter bias in the same way as ECV, omegaH, and PUC. They concluded that CFI and SRMR showed some evidence as "unidimensional enough" indices, while RMSEA was not effective in this context. CFI was strongly related to ECV (r = 0.80) and moderately related to omegaH (r = 0.49) but was effectively unrelated to PUC (r = 0.09). SRMR was moderately related to all three indices, ECV (r = -0.57), omegaH (r = -0.41), and PUC (r = -0.40). But their relationship to predicted bias in estimated structural parameters was not as effective an indicator as ECV and PUC. Reise, Scheines, et al. ultimately concluded that while CFI and SRMR are somewhat related to the general factor strength indices, they are not as effective as indicators of resulting parameter bias when data are fit to a misspecified unidimensional model. In other words, model fit alone is likely not an adequate indicator of potential parameter bias in a misspecified fitted model. Although Reise, Scheines, et al. reported a strong relationship between ECV and CFI, indicating that as ECV increased, or as data became more "unidimensional," CFI also increased. What they did not examine was the relationship between ECV and when CFI reached acceptable levels of fit (i.e., CFI $\geq$ 0.95). Additionally, they did not examine the model fit of the bifactor model, enabling an evaluation of the discrepancy in fit between the bifactor and unidimensional model.

Unlike Reise, Scheines, et al. (2013), Gu et al. (2017) fit both the (true) bifactor and the (misspecified) unidimensional model in their study and examined the performance of CFI and RMSEA. The goal was to investigate if these indices would suggest appropriate fit for the misspecified unidimensional model. It should be noted that

model fit in their study pertains to the fit of the structural model, as a measurement model was not utilized in their study. Not surprisingly, fit of the bifactor model was acceptable in all conditions. However, there were conditions in which fit of the unidimensional model reached acceptable levels, based on CFI and RMSEA values. For example, in one condition in which ECV was very high at 0.89, RMSEA for the bifactor and unidimensional models was 0.009 and 0.033, respectively, and CFI values were 0.998 and 0.983, respectively. While these fit indices favor the bifactor model, the values are within an acceptable range (i.e., CFI $\geq$ 0.95; RMSEA $\leq$ 0.06) for the unidimensional model. Applying the parsimony principle, the unidimensional model is likely preferred in this condition (Kline, 2011). In another condition in which ECV was 0.33, RMSEA for the bifactor and unidimensional models was 0.009 and 0.034, respectively, and CFI values were 0.995 and 0.965, respectively. Again, while the bifactor model showed better model-data fit, the unidimensional model reached an acceptable level of fit, however, this condition was associated with structural parameter bias greater than 10%. A comprehensive evaluation of model fit was not conducted in their study. Gu et al. only cautioned that under certain conditions, model fit indices might suggest a unidimensional model, but bias may be substantial.

Finally, relevant from Xia's (2018) study include findings related to an examination of model fit, particularly concerning the performance of the model selection indices AIC and BIC in relation to selecting the true bifactor model. Similar to Gu et al. (2017), model-data fit in Xia's study was assessed for the structural model and not the measurement model. When the bifactor model was the true model, the percentages of AIC and BIC values were at or close to 0% in selecting the bifactor model, indicating

these indices rarely if ever selected the correct bifactor model. Xia did not report which model these indices favored. Similar findings were reported from Morgan et al., (2015) in their examination of model-data fit between the correlated-factors models, higher-order model, and bifactor model. Under certain conditions, they reported BIC performed poorly in selecting the true bifactor model. Additionally, in Xia's study, when the unidimensional model was fit to data generated from the bifactor model, in certain conditions, model fit indices indicated adequate fit for the unidimensional model. In an unbalanced condition where the ratio of positively worded items to negatively worded items was 7:3, and an ECV of 0.80, Xia reported CFI values were satisfactory in 98% of solutions, TLI was satisfactory in 85% of solutions, RMSEA was satisfactory in 95% of solutions, and SRMR was 100% satisfactory. Furthermore, Xia reported the chi-square statistic performed satisfactorily across all conditions in which the unidimensional model was fit to data generated from a bifactor model, rejecting this model in 100% of conditions.

Results from Reise, Scheines, et al. (2013), Gu et al. (2017), and Xia (2018) suggest that the relationship between conventional model fit indices and indices of general factor strength in the bifactor model is inconsistent. Results from these studies regarding the general factor strength indices and their potential relationship to standard CFA/SEM model fit indices suggests that model fit indices, in the traditional sense, are unreliable as a direct indicator of parameter bias when a misspecified unidimensional model is fit to the data. Additionally, while Reise, Scheines et al. reported a strong, positive correlation between ECV and CFI, Gu et al. reported a more inconsistent relationship between ECV and CFI, as ECV values of 0.89 and 0.33 in their study were

both associated with CFI values that exceeded 0.95. Finally, results from Xia, indicated that AIC and BIC are perhaps not reliable indices in selecting a true bifactor model in the context of item wording method effects. However, both Gu et al. and Xia only assessed model-data fit from the structural model. A more thorough investigation is needed, particularly concerning comparisons between the bifactor model and unidimensional model in fit indices obtained for the CFA measurement model, as this is where decisions are made regarding an instrument's potential unidimensional structure.

Context of the Current Study

In recent years, the application of bifactor models to model item wording method effects appears to have become standard in practice (e.g., e.g., Alessandri et al., 2015; Hyland et al., 2014; Hystad & Johnsen, 2020; Marsh et al., 2010; McKay, Morgan, et al., 2015; Salerno et al., 2017). However, as indicated by several simulation studies, model fit indices may be biased towards selecting a bifactor model, indicating the bifactor model may overfit the data (e.g., Bonifay & Cai, 2017; Greene et al., 2019; Morgan et al., 2015; Murry & Johnson, 2013). In a recent applied study, Reise et al. (2016) extended similar findings in the context of item wording method effects. Specifically, Reise et al. examined item responses obtained from the Rosenberg Self-Esteem Scale in a large national database. In recognizing that the inclusion of both positively and negatively worded items often leads to questions regarding the instrument's dimensionality, Reise et al. examined the relationship between response patterns and model fit. Specifically, they applied iteratively reweighted least squares (IRLS) estimation in order to estimate an individual's response pattern congruence with (or distance to) a proposed model structure. The authors examined the unidimensional model, the correlated-factors model, and the

bifactor model. Results suggested the bifactor model, with one general factor and two method factors, provided the best fit to the overall data compared to a unidimensional model and a correlated-factors models with factors defined by the item wording direction. However, in evaluating individual response pattern congruence with the various model structures, Reise et al. found that for 86% of individuals, their response patterns were adequately modeled by the unidimensional model, 3% were adequately modeled by the bifactor model, and 11% had response patterns that were determined to not follow a specified model. Reise et al. described these 11% of responses as "unmodelable." Despite the fact that the bifactor model showed the best fit overall, only 3% of data patterns were congruent with this model structure, while an overwhelming majority of response patterns followed a unidimensional structure. The study results confirm Bonifay & Cai's (2017) findings that the bifactor model is better able to accommodate more diverse, and perhaps random, response patterns. However, in Reise et al.'s study, the application of model-based indices derived from the bifactor model provided critical information. Reise et al. found an ECV of 0.80 and an omega hierarchical of 0.84, indicating that a strong majority of the variance was attributable to the general factor, and the data were likely essentially unidimensional. Furthermore, they concluded that a unidimensional scoring strategy can be employed, and the unidimensional model can be used as the measurement model in a full SEM analysis. In other words, the intended use of the RSES as a unidimensional scale with total scores as representations of global self-esteem was confirmed. Such conclusions cannot be made based solely on standard model fit statistics, as model fit in Reise et al.'s study favored the bifactor model.

More recent applications of the bifactor model in the context of item wording method effects have seen an increased use of the bifactor model-based indices to reach decisions regarding how best to model and treat the data from instruments potentially contaminated by item wording method effects. However, the degree to which these indices are applied has varied. For example, some researchers have only utilized ECV (e.g., Rodrigo et al., 2019; Urban et al., 2014), while others have taken a more comprehensive approach, similar to Reise et al. (2016), and considered ECV and the omega reliability indices (e.g., Arias & Arias, 2017; Gnambs & Staufenbiel, 2018; Hystad & Johnsen, 2020; McKay, Cole, et al., 2015; McKay, Morgan, et al., 2015; Salerno et al., 2017). Furthermore, PUC has not been a factor considered across most studies utilizing ECV and/or omegaH in the context of item wording method effects.

In an investigation of the structure of the General Heath Questionnaire (GHQ-12), a 12-item balanced scale, Rodrigo et al. (2019) relied on model fit of the bifactor model, inspection of factor loadings, and an obtained ECV of 0.524 to arrive at the conclusion that the method factors associated with the positively and negatively worded items were severe enough that the bifactor model was needed as the measurement model in subsequent SEM analyses, as opposed to treating the data as essentially unidimensional and fitting a unidimensional CFA model. In contrast, Hystad and Johnsen (2020) in another study of the GHQ-12, reported an ECV of 0.547 and an omega hierarchical for the general factor of 0.598. Despite obtaining similar ECVs, Hystad and Johnsen reached a different conclusion than Rodrigo et al. They concluded that while the GHQ-12 is not strictly unidimensional, ignoring the multidimensionality introduced by the method

factors would likely to lead to minimal bias in a global composite score (i.e., the scale may be treated as unidimensional).

Another example comes from a study examining the dimensionality of the Consideration for Future Consequences Scale (CFCS), a 12-item instrument containing five positively worded items and seven negatively worded items. McKay, Morgan, et al. (2015) administered the CFCS across two samples. They concluded the bifactor model with two method factors representing the item directionality best fit the data. In Sample 1, they reported an ECV of 0.58 and an omegaH of 0.57 for the general factor. In Sample 2, they reported an ECV of 0.73 and an omegaH of 0.69 for the general factor. Despite the difference in magnitude of the reported model-based indices across the samples, McKay, Morgan, et al. concluded that the CFCS is essentially unidimensional, and that scores on the CFCS can be obtained through a unidimensional scoring scheme by summing the 12-items.

Across the three studies reviewed, PUC was not considered as part of the analysis. As a balanced scale with 12 items, the GHQ-12 has a PUC of 0.54, while the CFCS is an unbalanced scale with five positively worded items and seven negatively worded items. This represents a PUC of 0.53. Both scales have what may be considered low PUC (i.e., Bonifay et al., 2015; Reise, Scheines, et al., 2013). In Bonifay et al.'s (2015) study, ECV needed to be at least 0.70 for acceptable bias when PUC was 0.52. Based on this threshold, neither study investigating the GHQ-12 met the threshold, despite claims from Hystad and Johnsen (2020) that a unidimensional model could be utilized. The results from Sample 2 in McKay, Morgan, et al.'s (2015) investigation of the CFCS met the 0.70 threshold. Clearly, more research is needed to help provide guidance to applied

researchers utilizing bifactor models in the context of item wording method effects, particularly in contexts of low PUC.

Purpose of the Current Study

As the use of bifactor models and the model-based indices derived from this model becomes more utilized in practice, an examination of the functioning of the indices proposed as indicators of essential unidimensionality across a variety of conditions in needed. Specifically, an extension of Reise, Scheines, et al.'s (2013) study examining the relationship between omegaH, ECV, and PUC and parameter bias in the context of item wording method effects is needed.

Thus far, simulation studies examining omegaH, ECV, and PUC have done so under limited and/or ideal conditions. The purpose of the current study was to conduct a comprehensive simulation study that included many of the data characteristics encountered by researchers investigating item wording method effects. For example, Reise, Scheines, et al. (2013) and Bonifay et al. (2015) only considered balanced specific factors in their studies. Balanced, in this context, refers the number of items as well as the size of the factor loadings on the specific factors. Gu et al. (2017) did consider unbalanced specific factors in that the ratio of positively to negatively worded items varied such that in some conditions there were more positively worded items (i.e., the negatively worded item method factor was smaller). However, Gu et al. only considered one specification of the bifactor model, a modified bifactor or the bifactor-(S-1), where only one specific factor is included. In practice, when researchers identify a bifactor model as fitting data containing item wording method effects related to both positively and negatively worded items, it is common to find varying degrees of strength related to

the method factors. Table 2.1 below highlights the average standardized loadings for the general factor, the positive item method factor, and the negative item method factor across a sample of studies that have utilized the bifactor model. As shown in this table, the strength of the wording factors can vary both between and within instruments. For example, Michaelides et al. (2016), in a study investigating the RSES, reported relatively balanced, moderate average loadings between the positive and negative wording method factors. In contrast, Urban et al. (2014), also in a study investigating the RSES reported average loadings across the negative item wording method factor that were more substantial compared to the positive item wording factor.

Table 2.1 *Average Standardized Factor Loadings from Select Studies Utilizing a Bifactor Model in the Context of Item Wording Method Effects*

Instrument/ Scale	Study Reference	General Factor	Positive Wording Method Factor	Negative Wording Method Factor
RSES	Michaelides et al. (2016)	.627	.463	.425
RSES	McKay et al. (2014)	.582	.196	.376
RSES	Urban et al. (2014)	.562	.323	.524
RSES	Xu & Leung (2016)	.533	.245	.393
GHQ-12	Hystad & Johnsen (2020)	.408	.275	.398

Note. RSES = Rosenberg Self-Esteem Scale; GHQ-12 = General Health Questionnaire-12

In a balanced condition, in which the standardized loading values are equivalent between the specific factors, the variance explained by these factors is evenly partitioned between the two factors. When the loading patterns become unbalanced, the variance

explained by the specific factors also becomes unbalanced, with one specific factor explaining more variance in a subset of items compared to the other. This uneven partitioning of the variance within the specific factors could impact the values of both ECV and omegaH that suggest parameter bias is minimal when data is fit to a unidimensional model is fit.

A second consideration for the current study included item distributional characteristics. In describing the differential response patterns that can arise when both positively and negatively items are included in an instrument, Spector et al. (1997) noted that this can manifest in the data as skewed item distributions, which in turn, can lead to the finding that a two-factor model with the factors defined by the item wording, fits the data. Spector et al. cautioned against the substantive interpretation of these so-called "pseudo-factors," implying that they are artifacts of the data. Since the application of bifactor models in the context of item wording effects has become more standard, the (mis)interpretation of method effects as separate, substantive factors has become less of a concern, however, researchers still encounter non-normally distributed data. For example, several researchers have reported violations of normality across items (e.g., Alessandri et al., 2015; Hazlett-Stevens et al., 2004; Salerno et al., 2017), with some reporting more severe violations of normality for the negatively worded items (e.g., Michaelides et al., 2016; Rodrigo et al., 2019). How this additional model complexity impacts the efficiency of ECV, omegaH, and PUC, and the subsequent parameter bias associated with fitting a unidimensional model has not been examined.

A third consideration for the current study was the role of the percentage of uncontaminated correlations (PUC) in the context of item wording method effects. The

PUC index refers to the percent of item correlations that only reflect variance from the general factor (i.e., the correlations between items from different specific factors). PUC is a way of quantifying the data's structure, as a structure with more specific factors, and fewer items contained within these specific factors, will have a higher PUC. In other words, fewer items are contaminated by both general and specific factor variance. The findings from Reise, Scheines, et al. (2013) and Bonifay et al. (2015) indicated that PUC plays an important role in relation to ECV. Both studies found that in conditions in which PUC was high, the value of ECV becomes less important, as low ECV values (as low as .20) were associated with minimal bias in parameter estimates. This indicates that as data becomes more "unidimensional" with fewer contaminated item correlations, the strength of the general factor in relation to the specific factors becomes less important. Findings from both studies suggested that when PUC is "low," ECV needs to be "high." How "low" is low for PUC and how "high" is high for ECV is unclear. "High" PUC in both studies were PUC values in excess of 0.90, a condition unlikely to be met in the context of a bifactor model with two item wording method effects. There are, however, two situations that alter the PUC in the context of item wording method effects. One concerns the number of negatively worded items included in an instrument. While the use of balanced scales (i.e., scales containing an equal number of positively and negatively worded items) is common, one of the recommendations from Gu et al. (2017) for researchers concerned with the impact of item wording method effects was to reduce the number of negatively worded items included in an instrument. This type of unbalanced item composition will alter the PUC. A second situation concerns the composition of the item wording method factors. It is common for researchers to find items that do not have

significant loadings on the specific factors. This is often found with some items related to the positively worded item method factor. In this situation, researchers often remove these items from loading on the specific factor associated with the positive item wording factor (e.g., Alessandri et al., 2015; Michaelides et al., 2016; Reise et al., 2016; Urban et al., 2014). This modification also alters the data structure, and, in turn, the PUC.

Finally, a fourth consideration is that of model-data fit. Results from Reise, Scheines, et al. (2013), Gu et al. (2017), and Xia (2018) indicated that model fit indices obtained from the unidimensional model will reach adequate levels (based on suggested cutoff values of Hu & Bentler, 1998,1999) when fit to data generated from a bifactor model. A bifactor model will always show better fit compared to a unidimensional model, given its model complexity (Reise et al., 2016). However, as shown by Gu et al. there are certain conditions in which both the bifactor and unidimensional model may reach acceptable levels of fit (per the recommendations of Hu and Bentler, 1998, 1999). Perhaps the better question is *to what degree* does the bifactor model fit better under certain conditions? One way to examine this is by recognizing the nested structure of the unidimensional and bifactor model. As noted by Reise (2012), the bifactor model and unidimensional model form a nested structure, with the unidimensional model nested within the bifactor model. In an investigation between the higher-order model and bifactor model (models that also form a nested structure), Gignac (2016) reported that examining the practical change in TLI, AIC, and BIC was informative, as these model fit indices include a greater penalty for model complexity compared to either CFI or RMSEA. Gignac found that the degree to which a bifactor model would show better fit compared to a higher-order model was related to the degree to which the proportionality

constraint imposed by the higher-order model was violated. The proportionality constraint refers to a condition in which the factor loading ratios (*g*/*s*) between the general factor and specific factors are constrained to be equal within each group-level factor in a higher-order model (Gignac, 2016). Put differently, the general and specific factor variance associated with indicators must be proportional within each group-level factor in a higher-order model (Giganc, 2016). The bifactor model does not impose such constraints between the general and specific factors. Gignac found strong, linear relationships related to the degree with which the proportionality constraint was violated and the magnitude in difference between TLI, AIC, and BIC (favoring the bifactor model). Applying this concept within the context of comparing unidimensional and bifactor models, the degree of misspecification between these models is a function of ECV (Reise, Scheines, et al., 2013). That is, as the explained common variance associated with the general factor becomes stronger, the more "unidimensional" the data. Examining this factor per the procedures of Gignac might provide more information regarding the relationship between approximate model fit indices and general factor strength indices, specifically ECV and PUC as these indices are considered statistics of unidimensionality. Additionally, Xia's (2018) findings indicated the model selection indices AIC and BIC did not perform well in identifying the (true) bifactor model, although no information was provided about which model was preferred in this context. Therefore, a more thorough examination of these indices is needed.

Recognizing that it is standard practice for researchers utilizing SEM to first assess the fit and specification of the measurement model prior to the structural model, the current study examined the bias in fitting a unidimensional measurement model to

multidimensional data generated from a bifactor model. Specifically, this study replicated similar procedures to Bonifay et al. (2015) in investigating the bias in estimated factor loadings of the general factor in a bifactor model when data is fit to a misspecified unidimensional measurement model. To examine the use of ECV, omegaH, and PUC in the context of a bifactor model with specific factors related to item wording method effects, this study incorporated several conditions frequently encountered by researchers including unbalanced specific factors, varying item distributional characteristics, and the role of varying degrees of PUC in this context. This study also included an evaluation of model fit statistics, comparing model fit values of the true bifactor model and the misspecified unidimensional model. As ECV and PUC are considered more direct indices of unidimensionality, the relationship between model fit indices, ECV, and PUC was examined. Results from this study were designed to assist applied researchers utilizing unidimensional scales that may be impacted by item wording method effects by providing guidance on how bifactor models, and select model-based indices, can be used to reach decisions on when the item wording method effects can be ignored with negligible bias.

CHAPTER 3

METHOD

The use of bifactor models to model data obtained from unidimensional scales potentially contaminated by method effects due to item wording has been frequently found in practice (e.g., Alessandri et al., 2015; Gana et al., 2013; Hyland et al., 2014; Hystad & Johnsen, 2020; Marsh et al., 2010; McKay, Morgan, et al., 2015; Salerno et al., 2017). With a bifactor model, a researcher can conduct a psychometric evaluation of an instrument's properties by applying several recommended model-based indices. The indices evaluate the strength of the general factor in relation to the specific (method) factors and assess the applicability for utilizing a unidimensional measurement model and an overall composite score in the context of multidimensionality. In this context, bifactor models, and subsequent model-derived indices, can be used to establish when data can be treated as "essentially unidimensional."

Following the procedures of Reise, Scheines, et al. (2013), Bonifay et al. (2015), and Gu et al. (2017), a simulation study was conducted to investigate data generated from a bifactor model but fit to a unidimensional model. This purposefully modeled a misspecified structure that ignored the method factors. The purpose of this study was twofold. First, this study examined the relationship between the general factor strength model-based indices derived from a bifactor model, explained common variance (ECV), omega hierarchical (omegaH), and the percent of uncontaminated correlations (PUC), and parameter bias when data were fit to a misspecified unidimensional model. Taken

together, ECV, omegaH, and PUC can be used to assess when bifactor data can be treated as essentially unidimensional. Both ECV and PUC are considered unidimensional statistics in the sense that ECV provides an assessment of the general factor strength by providing the percentage of variance explained by the general factor compared to the specific factors (Reise, 2012; Rodriguez et al., 2016b). PUC is a way of quantifying a data's structure that follows a bifactor model. Higher PUC values indicate fewer items are contaminated by both general and specific factor variance, and that the data follow a more "unidimensional" structure (Reise, Scheines, et al., 2013; Rodriguez et al., 2016a, 2016b). While omegaH is not considered a direct indicator of unidimensionality, it is useful in the sense that it assesses the reliable variance in observed, unit-weighted composite scores that is due to only the general factor (i.e., the construct of interest; Rodriguez et al., 2016a, 2016b). The higher the omegaH, the more likely scores represent variation in the general construct. Traditionally, "unidimensionality," within the CFA framework is assessed by evaluating a unidimensional measurement model by examining its model-data fit via model fit statistics. Therefore, the second purpose of this study was to examine the relationship between select model fit indices and ECV and PUC, as these model-based indices are thought to represent unidimensionality.

Specifically, this study utilized a Monte Carlo simulation study. Monte Carlo simulation studies are frequently used within SEM to investigate bias in estimated parameters, standard errors, and model fit under a variety of controlled conditions in which assumptions may be violated, including sample size, normality, model size, categorical vs. continuous variables, and model misspecification (Bandalos & Leite, 2013; Bandalos & Gagne, 2012; Paxton et al., 2001). In a Monte Carlo simulation,

repeated samples are estimated from a known population model, thereby creating a known sampling distribution (Paxton et al., 2001). Results are then aggregated across the samples to investigate potential bias in parameters, standard errors, and/or model fit (Bandalos & Leite, 2013).

Monte Carlo simulation studies may be preferred over other simulation techniques, such as a population study, in that the impact of sampling variability and the stability of parameter estimates can be investigated (Bandalos & Leite, 2013). Further, simulations are useful to the extent that the conditions included are representative of the conditions encountered with real data (Bandalos & Leite, 2013). This requires a researcher to balance generalizability and feasibility regarding the conditions to include in a study (Bandalos & Gagne, 2012). Overall, Monte Carlo simulation studies provide both methodological and applied researchers with guidance for how to treat data and interpret results across a variety of conditions, which may often be complex and/or "suboptimal" (Bandalos & Gagne, 2012).

Population Model

The population model used in this study was a bifactor model with one general factor and two specific factors. The model was specified such that the general factor represented the construct of interest and the specific factors represented method factors or nuisance factors unrelated to the general factor, but which were set to vary in strength and contaminating effect with measuring the general factor. The population bifactor model met both assumptions of a restricted bifactor model. Specifically, all items had non-zero loadings on the general factor and only one specific factor, either a specific factor

representing the positively worded item method factor or the negatively worded item method factor. All other loadings were to zero (i.e., no cross loadings).

All latent factors were specified to be orthogonal. While orthogonality is required between the general and specific factors in a bifactor model, it is only recommended (and not required) between specific factors (Reise, 2012). Some researchers utilizing bifactor models in the context of item wording method effects have considered correlations between the specific factors (e.g., positively and negatively worded item method factors; Donnellan et al., 2016; Lindwall et al., 2012; Quilty et al., 2006; Rodrigo et al., 2019). However, researchers have cautioned against this modeling strategy noting both issues in model interpretability and applicability (i.e., Reise, 2012). Furthermore, interpretation of the bifactor model-derived indices (ECV, omegaH, and PUC) require these factors to be orthogonal (Canivez, 2016; Bonifay et al., 2015; Gignac & Watkins, 2013).

The population model consisted of 12 items. A 12-item model was selected because this represents, approximately, the typical length found with instruments designed to be unidimensional scales that contain both positively and negatively worded items. While the Rosenberg-Self Esteem Scale (RSES; Rosenberg, 1965) is a 10-item instrument, the General Heath Questionnaire-12 (GHQ-12; Goldberg & Williams, 1988), Consideration of Future Consequences Scale (CFC; Strathman et al., 1994), and the Core Self-Evaluation Scale (CSES; Judge et al., 2003) each consist of 12 items. Additionally, Gu et al.'s (2017) simulation study contained both a 12-item and an 18-item condition, with results consistent across both item-lengths. In the current study, while all 12 items loaded on the general factor across all conditions, the number of items loading on each

method (specific) factor varied as a function of the number of positive to negative items (N_p:N_n) specified to load on their respective method factor (i.e., 6:6, 8:4, 3:6).

Data Generation

All data were generated and analyzed using the *Mplus* software package (v8.7; Muthen & Muthen, 1998-2017). Sample code is presented in Appendix A. Simulated conditions consisted of a sample size of 500. The bifactor model represents a highly complex model, with multiple factor loadings estimated per item. In the current study, the most complex model estimated was a bifactor model in which each item loaded on both a general factor and one of two method factors. In this model, 36 parameters were estimated (12 factor loadings on the general factor, 12 factor loadings across two method factors, 12 error components). In considering the appropriate sample size, Kline (2011) has noted a *N:q* (i.e., sample to size to number of estimated parameters) ratio of 20:1 is ideal, with a ratio of 10:1 considered as a minimum threshold. Based on this recommendation, a sample size of 360 to 720 is appropriate for the current study. A sample size of 500 was selected because it is representative of the sample size frequently used with applied studies investigating item wording method effects. Additionally, recent studies have suggested that a minimum sample size of 500 is sufficient for model convergence and parameter estimates for a bifactor model (Bader et al., 2022). Finally, Gu et al. (2017) reported similar bias levels in estimated structural coefficients from a misspecified unidimensional model across sample sizes of 500 and 1,000.

Five-point ordered categorical data were generated and categorized through the application of threshold values. Five categories were chosen because this is representative of the response scale (i.e., Likert scale) employed by many of the unidimensional scales

that include both positively and negatively worded items. Additionally, when categorical data contain five categories or more, data can be treated as continuous and estimated via maximum likelihood (ML; Rhemtulla et al., 2012).

In categorical variable methodology (CVM), it is assumed each observed categorical indicator is associated with a latent response variable that is continuous and normally distributed (Finney & DiStefano, 2013; Kline, 2011). This underlying latent response variable (y*) is then linked to the observed categorical variable (y) by way of a threshold. Thresholds are the point at which the underlying continuous latent response variable is divided into categories (Finney & DiStefano, 2013). A total of c-1 thresholds are needed, where c is the number of response categories. As this study utilized a five-point response scale, a total of four thresholds were needed. Discretizing or cutting the continuous underlying latent distribution by applying thresholds to generate desired observed item level distributions is a frequently applied technique in simulation studies (e.g., DiStefano, 2002; DiStefano & Morgan, 2014; Flora & Curran, 2004; Maydeu-Olivares, 2017; Rhemtulla et al., 2012).

In this study, item distributions were varied to include both normal and non-normal distributions. Reports of normality violations across empirical studies investigating item wording method effects are mixed. While some have reported more general violations of normality and multivariate normality (e.g., Alessandri et al., 2015; Hazlett-Stevens et al., 2004; Salerno et al., 2017), others have reported normality violations to a larger degree across the negatively worded items (e.g., Michaelides et al., 2016; Rodrigo et al., 2019). To examine the impact of varying item distributional characteristics, this study incorporated two conditions, one in which all items were

normally distributed, and another in which the negatively worded items were non-normally distributed. Introducing non-normality into the model adds an additional layer of model complexity, in addition to misspecification with the unidimensional model. Some simulation studies have found that even under robust estimation methods, model fit, parameter estimates, and standard errors can be biased in certain conditions if both model misspecification and non-normality are present (e.g., Curran et al., 1996; Lai, 2018). Within CFA research, researchers report that when univariate skewness and kurtosis approach extreme levels (e.g., values absolute values in excess of 2.0 and 7.0, respectively), issues may arise when using ML-based estimation (Curran et al., 1996; Finney & DiStefano, 2013; Rhemtulla et al., 2012). To simulate data from both normal and non-normal item distributions, threshold values were applied to generate normally distributed data with approximate skewness and kurtosis values of zero (i.e., Rhemtulla et al., 2012), and non-normally distributed data across the negatively worded items with approximate skewness of -3.0 and kurtosis of 7.0. Similar values have been used or reported to demonstrate extreme levels of non-normality across a number of studies (e.g., DiStefano & Morgan, 2014; Hoogland & Boomsma, 1998)

Number of Items per Method Factor

Both Reise, Scheines, et al. (2013) and Bonifay et al. (2015) noted the importance of PUC in moderating the effect of ECV in indicating when bifactor data can be treated as unidimensional. PUC is impacted by both the number of specific factors and the number of items contained within each specific factor. Three conditions were included related to the number of items specified to load across the two method factors. These conditions were included to represent both the design of the instruments applied in the

research and model irregularities found in applied studies using bifactor models to model item wording method effects. Each of these conditions generated different PUC values. One condition was simulated to represent a balanced scale. The balanced scale included an equal number of positively and negatively worded items (6:6), with each item loading on their respective method factor. Utilizing Equation 13 presented in Chapter 2, this condition represented a PUC value 0.54, indicating that 54% of the item correlations were not contaminated by the method factors.

A second design related condition that directly impacts the PUC occurs when fewer negatively worded items are included in an instrument. In this condition, eight items represented the positively worded items and four items represented the negatively worded items (8:4). This condition represented a PUC value of 0.48, indicating that 48% of the item correlations were not contaminated by the method factors.

A third condition was selected to represent a common finding across studies analyzing bifactor models in the context of item wording method effects. Several studies have reported irregularities in the composition of method factors in that not all items load significantly on their respective method factor. This is most frequently associated with the positively worded item method factor (e.g., Alessandri et al., 2015; Michaelides et al., 2016; Reise et al., 2016; Urban et al., 2014). As such, a third condition was simulated to replicate this finding. In this condition, three items loaded on the positively worded item method factor and six items loaded on the negatively worded item method factor (3:6). This resulted in a PUC value of 0.73, indicating that 73% of item correlations were not contaminated by the method factors. In summary, three conditions were included that represented the composition of specific method factors encountered in applied research.

In turn, these conditions generated the following PUC values for investigation: 0.54, 0.48, and 0.73.

Loading Values and Patterns

Item loading values and patterns were selected to reflect values typically found when bifactor models are applied in the context of item wording method effects. Loading values were also varied as these values directly influence the values of ECV and omega hierarchical. In both the Reise, Schienes et al. (2013) and Bonifay et al. (2015) research studies, factor loadings were specified to be equal across all specific factors. This condition does not generalize to situations frequently found when investigating item wording method effects. As shown in Table 2.1 in Chapter 2, researchers investigating item wording method effects often find variations or unbalanced patterns of item loadings between the method factors. As such, this study incorporated an unbalanced factor loading pattern to examine its effect.

In the current study, standardized item loading values of 0.7, 0.5, and 0.3 were varied across loading patterns. These values were selected to represent strong, moderate, and weak loading values. These loading values have also been used to establish varying degrees of unidimensionality (or multidimensionality) in related studies examining bifactor models (i.e., Liu & Thompson, 2021). Standardized loading values of 0.7 and 0.5 were varied across the general factor, while 0.5 and 0.3 were varied across the method factors. The reason for this was twofold. First, 0.7 is not included as a method factor loading following the recommendation of Reise, Scheines, et al. (2013) who noted computational issues when communalities become too high. Second, loading values of 0.5 and 0.3 were selected for the method factors as these approximate typical values

found in applied studies. Additionally, Reise, Scheines, et al. noted that any item set with average loadings less than 0.3 is not worth considering. In the balanced condition, the factor loading values were equal across both method factors. This was similar to the method employed by both Bonifay et al. (2015) and Reise, Scheines, et al. The unbalanced conditions were simulated such that the loadings on the method factors were unbalanced, with the positively worded items loading 0.3 on its associated method factor and the negatively worded items loading 0.5 on its associated method factor. This condition was selected recognizing that when this unbalanced pattern occurs, the "stronger" method factor is most often associated with negatively worded items (e.g., Hystad & Johnsen, 2020; McKay et al., 2014; Urban et al., 2014; Xu et al., 2016).

Tables 3.1 and 3.2 below detail the loading value patterns and the population ECV (Table 3.1) and omegaH (Table 3.2) values associated with these patterns. ECV and omegaH were calculated using the population standardized loading values via Equations 9 and 12 presented in Chapter 2. Also shown in Tables 3.1 and 3.2 is the effect of PUC, in that the values may change as a result of the data's structure (i.e., the number of items per method factor).

Table 3.1 *Factor Loading Patterns and Explained Common Variance (ECV) Values*

Condition	General Factor Loading	Positive Factor Loading	Negative Factor Loading	ECV (6:6)	ECV (8:4)	ECV (3:6)
Balanced	0.7	0.3	0.3	0.84	0.84	0.88
Balanced	0.7	0.5	0.5	0.66	0.66	0.72
Balanced	0.5	0.3	0.3	0.74	0.74	0.79
Balanced	0.5	0.5	0.5	0.50	0.50	0.57
Unbalanced	0.7	0.3	0.5	0.74	0.77	0.77
Unbalanced	0.5	0.3	0.5	0.60	0.64	0.63

Table 3.2 *Factor Loading Patterns and Omega Hierarchical (OmegaH) Values*

Condition	General Factor Loading	Positive Factor Loading	Negative Factor Loading	OmegaH (6:6)	OmegaH (8:4)	OmegaH (3:6)
Balanced	0.7	0.3	0.3	0.86	0.85	0.88
Balanced	0.7	0.5	0.5	0.77	0.75	0.82
Balanced	0.5	0.3	0.3	0.71	0.70	0.75
Balanced	0.5	0.5	0.5	0.60	0.58	0.67
Unbalanced	0.7	0.3	0.5	0.81	0.83	0.83
Unbalanced	0.5	0.3	0.5	0.65	0.68	0.68

In summary, this study included six factor loading value patterns, four patterns associated with a balanced condition and two patterns associated with an unbalanced condition. The combination of both the general and method factor loadings represents the degree of unidimensionality present in the data. The "strongest" unidimensional structure for the bifactor model occurred when general factor loadings were 0.7 and group loadings were 0.3, while the weakest unidimensional structure (or more bifactor/multidimensional), occurred when general factor loadings were 0.5 and method factor loadings were 0.5.

Estimated Models

Two levels of model misspecification (a correctly specified bifactor model and a misspecified unidimensional model) were implemented to assess the utility of using the model-based indices in identifying essentially unidimensional models in the context of item wording method effects. When scales designed to be unidimensional contain both positively and negatively worded items, researchers nearly always find that a unidimensional measurement model does not fit the data. The bifactor model, with a general factor that influences all items, but which may contain smaller, additional specific factors that influence a subset of items related by content or item features, may provide a

more realistic or accurate view of unidimensionality in this context (Reise et al., 2010). In finding the bifactor model best models data containing item wording method effects, some researchers have concluded that a unidimensional representation and treatment of the data obtained from such instrument is inappropriate and would lead to potentially biased conclusions (e.g., Alessandri et al., 2015; Gana et al., 2013; Marsh et al., 2010). However, clarity regarding such conclusions can be provided by applying the recommended model-based indices from a bifactor model. Findings from Reise, Scheines, et al. (2013), Bonifay et al. (2015), and Gu et al. (2017) have shown that under certain conditions, utilizing a unidimensional model, and subsequent composite scores, results in minimal bias. This is related to the strength of the general factor in relation to the specific factors in a bifactor model. Therefore, following the procedures of these studies, the current study simulated data from a bifactor model (i.e., the true model) and then fit the data to a misspecified unidimensional model to examine the biasing effect. Effectively, this view of the data assessed the impact of ignoring the item wording method effect. Specifically, bias in estimated factor loadings when a misspecified unidimensional model was fit to the data were examined. Additionally, an evaluation of model fit indices and their relationship to ECV and PUC were investigated.

Estimator

All models were estimated using the MLMV (maximum likelihood with mean and variance adjusted) estimator available in the *Mplus* software package (v8.7; Muthen & Muthen, 1998-2017). Maximum likelihood (ML) is a popular estimator for many researchers, however its assumptions of continuous, multivariate normal data present challenges as data analyzed in the CFA/SEM framework are often categorical in nature.

Under conditions of non-normality, the ML-based chi-square statistic (χ^2) will be biased, even when the model is correctly specified, however the use of ML with a robust correction will yield unbiased standard errors and fit statistics (Curran et al., 1996; Finney & DiStefano, 2013).

As such, all models in this study were estimated using the MLMV estimator, as this produces a chi-square statistic that is both mean and variance adjusted. While several ML-based estimators with a robust correction are available within *Mplus* (e.g., MLM, MLR), MLMV has been shown to perform best under conditions of non-normality, yielding more accurate standard errors and Type I error rates (Maydeu-Olivares, 2017). The MLMV estimator uses a corrected chi-square statistic that is a variant of the Satorra-Bentler (SB) scaled chi-square statistic (Maydeu-Olivares, 2017). Consequently, approximate fit indices that include the model chi-square statistic as part of its calculation also incorporate the adjusted chi-square (i.e., CFI, TLI; Finney & DiStefano, 2013).

Summary of Study Conditions

The population bifactor model and design factors were selected to replicate typical models and conditions found with unidimensional scales that contain both positively and negatively worded items. Design factors held constant included the total number of items, sample size, and response scale format. Manipulated factors included the number of items per specific (method) factor (3 conditions), item loading values and patterns between the method factors and the general factor (6 conditions), and item-level distributions (2 conditions). Varying the number of items per method factor and the item loading values is directly related to the values of ECV, omegaH, and PUC, as this

information is used to calculate these statistics. Varying these design factors in turn generated a variety of ECV, omegaH, and PUC values used for investigation. Table 3.3 below summarizes the conditions used in this study. Factor variances of latent variables were set to 1.0 for identification.

Table 3.3 *Summary of Design Factors*

Design Factor	Condition
Number of Items per Method Factor (N_p:N_n)	6:6; 8:4; 3:6
Factor Loading Patterns	
General Factor	0.7 or 0.5
Method Factors (Balanced/Unbalanced)	0.5 or 0.3
Item Distribution	All items normal (Sk = 0, Ku = 0) Negative items non-normal (Sk = -3, Ku = 7)
Model Misspecification	Correctly specified bifactor model, misspecified unidimensional model

Note. Sk = skewness; Ku = kurtosis

This simulation study consisted of a crossed design with 72 cells: 3 conditions related to the number of items per method factor (6:6, 8:4, 3:6) x 6 item loading values and patterns between the method factors and the general factor (0.7, 0.5 for the general factor; 0.5 or 0.3 for the method factors; four balanced and two unbalanced method factor conditions) x 2 item-level distributions (all normal; negative items non-normally distributed) x 2 model misspecification (correctly specified bifactor model, misspecified unidimensional model). One thousand replications were run for each cell.

Study Outcomes and Analysis

Parameter Bias

In their study examining ECV as an indicator for when data generated from a bifactor model can be treated as unidimensional, Bonifay et al. (2015) described the

population bifactor model as one in which only the general factor was of interest, noting that this allowed them to precisely define parameter bias as bias in factor loading estimates when data is fit to a unidimensional model. The current study employed a similar conceptualization. Parameter bias was defined as bias in factor loading estimates between the general factor in the bifactor model and the misspecified unidimensional model, essentially examining how well the unidimensional model recovered the general factor's structure from a bifactor model. To examine bias in factor loading estimates, the average factor loading estimates per replication were computed and compared to the population values. Next, the relative bias (RB) was assessed as the difference between the average values and population values (i.e., Bandalos & Leite, 2013; Rhemtulla et al., 2012).

$$RB = \left(\frac{\bar{\theta}_{est} - \theta}{\theta}\right) * 100, \quad (14)$$

where $\bar{\theta}_{est}$ is the average estimated value of the parameter (i.e., standardized factor loading) across each replication of a cell, and θ is the population or true value of the parameter. Positive relative bias indicates the parameter estimate has been overestimated, while negative relative bias indicates underestimation. Relative bias of |5%| to |10%| may be considered acceptable, while bias greater than |10%| is considered more severe (Flora & Curran, 2004; Hoogland & Boomsma, 1998). The 10% relative bias threshold was also used in similar studies (i.e., Bonifay et al., 2015; Gu et al., 2017; Reise, Scheines, et al., 2013). Relative bias of the factor loading estimates from the unidimensional model were then compared to ECV, omegaH, and PUC values calculated from the population bifactor model to identify thresholds for when bias reached unacceptable levels relative to the values of the model-based indices.

Model Fit

In recognizing that CFA model selection, and therefore, dimensionality decisions in the context of item wording method effects, are based on examination of model fit indices, this study included an examination of model fit indices and the bifactor model-based indices considered as unidimensional statistics, ECV and PUC. Reise, Scheines, et al. (2013) offered the most comprehensive, although still limited, evaluation, finding that CFI was more strongly related to ECV (r = 0.80), compared to either RMSEA or SRMR. Findings from Gu et al. (2015) indicated that under certain conditions, both the bifactor model and unidimensional model showed acceptable levels of fit. Recognizing that the bifactor model will always fit better compared to a unidimensional model, this prompted the question, *to what degree does the bifactor model fit better?* To examine model-data fit between the bifactor model and the unidimensional model, this study included an examination of CFI, TLI, AIC, and BIC by assessing the degree of difference in fit between the two models (i.e., ΔCFI, ΔTLI, ΔAIC, and ΔBIC). The examination allowed the difference to be a function of both ECV and PUC, as both indices are considered statistics of unidimensionality.

As noted by Reise (2012), the bifactor model and unidimensional model form a nested structure, with the unidimensional model nested within the bifactor model. Therefore, the bifactor model and unidimensional model can be compared using nested model comparisons. One way to compare nested models is a likelihood ratio (LR) test comparing the obtained chi-square statistics between models. However, this test is also impacted by the same factors that influence the model-based chi-square (i.e., sensitive to sample size; Schermelleh-Engel et al., 2003; West et al., 2012). Additionally, when using

a ML-based estimator with a robust correction (i.e., MLMV), the traditional chi-square difference or LR test cannot be utilized (Schermelleh-Engel et al., 2003). As an alternative, an evaluation of change in approximate fit indices compared to a recommended cutoff criterion can be used (West et al., 2012). Gignac (2016) utilized such a procedure in a simulation study examining model-data fit between the higher-order model and the bifactor model.

Gignac (2016) examined practical change in TLI, AIC, and BIC between the higher-order model and the bifactor model as a function of violating the proportionality constraint imposed by the higher-order model. Gignac selected these fit indices as these indices impose a greater penalty for model complexity compared to either CFI or RMSEA. In other words, TLI, AIC, and BIC reward model parsimony to a greater extent than either CFI or RMSEA. As the unidimensional model is a more parsimonious model compared to the bifactor model, examination of these fit indices in this context is relevant.

The Tucker-Lewis Index (TLI), like CFI, is an incremental fit index, in that it assesses the improvement in model fit compared to a baseline model which assumes no relationship between indicators (DiStefano, 2016). TLI values $\geq .95$ are indicative of acceptable model fit (Hu & Bentler, 1999). Also known as the nonnormed fit index (NNFI), TLI considers both the degrees of freedom from the hypothesized and baseline model in its calculation, therefore more complex models are penalized by a downward adjustment (Schermelleh-Engel et al., 2003). This is seen in Equation 15, where the subscripts 0 and k represent the baseline and tested model (West et al., 2012).

$$TLI = 1 - \frac{(\chi_k^2 - df_k)/\,(\chi_0^2 - df_0)}{df_k/df_0}. \tag{15}$$

AIC and BIC are considered model selection indices and can be used to compare models that are either nested or non-nested, with lower values between models representing improved fit (DiStefano, 2016; West et al., 2012). AIC and BIC both reward model parsimony (i.e., fewer estimated parameter; lower degrees of freedom) by incorporating the number of free parameters into the formula. Compared to AIC, BIC places a harsher penalty on more complex models (DiStefano, 2016). Equations 16 and 17 are the formulas for AIC and BIC as used in *Mplus* (Muthen, 1998-2004). In each equation *r* is the number of free model parameters and *n* is the sample size.

$$AIC = -2logL + 2r \ . \tag{16}$$

$$BIC = -2logL + (r)(ln)(n) \ . \tag{17}$$

In addition to TLI, AIC, and BIC as utilized in Gignac (2016), this study also included an evaluation of the comparative fit index (CFI). Reise, Scheines, et al. (2013) found that CFI displayed the strongest relationship to ECV. Findings from Gu et al. (2017) suggested an inconsistent relationship between CFI and ECV. CFI, like TLI, is also an incremental fit index. While CFI also includes a penalty for model complexity, Gignac (2016) notes it is only a minor penalty compared to TLI. The chi-square statistic for both the tested model and baseline model included in the formula for CFI (Equation 18 below) follows a noncentral chi-square distribution (DiStefano, 2016). The numerator and denominator of CFI's calculation provide an estimate of the noncentrality parameter for the tested model and baseline model (Kline, 2011). The noncentrality parameter provides a measure of discrepancy in fit between the population and tested model (DiStefano, 2016). The CFI provides a measure of the relative reduction in the

noncentrality parameter between the tested and baseline model (DiStefano, 2016). CFI values ≥ .95 are recommended for acceptable model fit (Hu & Bentler, 1999).

$$CFI = 1 - \frac{(\chi_k^2 - df_k)}{(\chi_0^2 - df_0)}, \tag{18}$$

where χ_k^2 is the chi-square statistic for the tested model, df_k is the degrees of freedom for the tested model, χ_0^2 is the chi-square for the baseline model, and df_0 is the degrees of freedom for the baseline model.

To examine model fit between the bifactor model and unidimensional model, descriptive statistics (i.e., *M* and *SD*) for each fit index were calculated by study condition for both models. To assess practical improvement in model fit between the bifactor and unidimensional model, the procedures of Gignac (2016) were replicated. Gignac considered a difference in TLI between the higher-order and bifactor model (i.e., ΔTLI) of ≥ 0.01 as an indication of practical improvement in fit and followed the recommended practical differences for AIC and BIC as proposed by Raftery (1995) of AIC/BIC differences between models less than 10 as indicating practical improvement in model fit. To assess practical improvement for CFI, a CFI difference (i.e., ΔCFI) of ≥ 0.01 was considered as evidence of improved model-data fit, as this threshold is frequently applied within the measurement invariance testing framework to assess practical differences in model fit (Cheung & Rensvold, 2002). The average difference in these fit indices were then examined by ECV and PUC to investigate the relationship between model-data fit and these general factor strength indices.

Summary of the Current Study

This study examined the use of model-based indices derived from a bifactor model in establishing essential unidimensionality in the context of item wording method

effects. As outlined by Reise, Bonifay, et al. (2013) and Rodriguez et al. (2016a, 2016b), these indices, specifically ECV, omega hierarchical, and PUC, can be used to assess the strength of the general factor and provide evidence for essential unidimensionality, effectively evaluating the extent to which the method effects related to item wording are contaminating an instrument's unidimensional property and structure. Use of these indices was evaluated in the context of data and model features frequently encountered by researchers utilizing unidimensional scales that contain item wording method effects including, number of items per specific (method) factor (3 conditions), item loading values and patterns between the method factors and the general factor (6 conditions), and item-level distributions (2 conditions). Following the procedures of related studies, a population bifactor model was used to generate data reflecting varying degrees of general and specific (method) factor strengths. A unidimensional model, ignoring the method factors, was then fit to the data to examine the biasing effect on factor loading estimates. Values of ECV, omegaH, and PUC derived from the population bifactor model were then compared to the relative bias in factor loading estimates from the unidimensional model to examine threshold points in which data can be treated as essentially unidimensional under conditions of minimal bias. Additionally, recognizing dimensionality decisions within the CFA framework are typically assessed by evaluating approximate model fit indices, this study also examined the relationship between select model fit indices and ECV and PUC, as these model-based indices are thought to represent statistics of unidimensionality. Specifically, the discrepancy in fit between the bifactor model and unidimensional model was examined by evaluating the difference in CFI, TLI, AIC, and BIC, and its relationship to ECV and PUC.

CHAPTER 4

RESULTS

The purpose of the current study was to investigate the use of bifactor models and the subsequent model-derived reliability indices in the context of item wording method effects. Previous studies have shown that the indices explained common variance (ECV), omega hierarchical (omegaH), and the percent of uncontaminated correlations (PUC) can be used to indicate when data that follows a bifactor structure can be treated as essentially unidimensional (e.g., Bonifay et al., 2015; Reise, Scheines, et al., 2013; Rodriguez et al., 2016a; 2016b). Specifically, these studies indicated that the use of a unidimensional model, when the underlying data structure followed a bifactor model, would result in minimal bias in either the estimated structural coefficient (e.g., Reise, Scheines, et al., 2013) or factor loadings on the general factor (e.g., Bonifay et al., 2015; Rodriguez et al., 2016a) under certain conditions. This study extends previous research by examining the unique data conditions faced by researchers utilizing bifactor models in the context of item wording method effects. A total of 72 conditions were included: 3 conditions related to the number of items per method factor (i.e., PUC values; 6:6, 8:4, 3:6) x 6 item loading values and patterns between the method factors and the general factor (0.7, 0.5 for the general factor; 0.5 or 0.3 for the method factors; four balanced and two unbalanced method factor conditions) x 2 item-level distributions (all normal; negative items non-normally distributed) x 2 model misspecification (correctly specified bifactor model, misspecified unidimensional model).

The population model used in this study was specified to reflect models typically encountered when investigating item wording method effects using bifactor models. The population model was specified such that the general factor represented the construct of interest and the two specific factors represented method factors or nuisance factors unrelated to the general factor but varying in strength and contaminating effect. Each method factor included in this study was defined by item wording (i.e., a positively worded method factor and a negatively worded method factor).

The research questions (RQ) addressed in this study were categorized around two key areas: (1) investigating the relationship between ECV, omegaH, and PUC and bias in the general factor loadings when estimated from a unidimensional model (RQ 1 and 2), and (2) examining the relationship between select model fit indices and ECV and PUC (RQ3). This chapter presents the findings organized around these areas. However, before presenting the results related to the research questions, issues related to model convergence are addressed.

Model Convergence

For each of the 72 cells, 1,000 replications were requested for each cell. Prior to analysis, replications were screened to examine model non-convergence and improper solutions. Improper solutions refer to cases in which the model produces implausible results such as negative variances or parameter values outside of possible bounds (i.e., Heywood cases); leaving such cases in a dataset could result in parameter bias (Chen et al., 2001). As part of its Monte Carlo simulation procedure, *Mplus* removes cases that fail to converge from the final dataset used for subsequent analyses. Additional data

screening, including examination of means and ranges of all estimated factor loadings and error variances, was conducted to identify and remove Heywood cases.

The datasets analyzed using a unidimensional model converged in 100% cases and no improper solutions were identified. Data analyzed with the bifactor model, however, were somewhat problematic. Table 4.1 presents the percent of cases that converged/contained proper solutions for the bifactor model for both normally and non-normally distributed data, presented by PUC, ECV, and omegaH values. As seen in Table 4.1, convergence/proper solution rates differed across different values of PUC but were similar for both the normally and non-normally distributed data. Convergence/proper solution rates were highest in conditions in which the true underlying model was associated with a PUC value of 0.73. In this condition, the method factors were specified such that only three of the six positive items loaded on the positive wording method factor, while all six negatively worded items loaded on the negative wording method factor. In this condition, convergence/proper solutions rates ranged 87.9% to 100.0% for normally distributed data and 88.1% to 100.0% for non-normally distributed data. Convergence/proper solution rates were notably smaller when the true underlying model was associated with PUC values of 0.54 and 0.48. When PUC was 0.54, convergence/proper solution rates ranged from 60.5% to 68.7% for normally distributed data and from 54.6% to 62.1% for non-normally distributed data. When PUC was 0.48, convergence/proper solution rates ranged from 57.9% to 67.2% for normally distributed data and 54.7% to 61.0% for non-normally distributed data.

In both conditions in which the PUC values were 0.54 and 0.48, all items loaded on both the general factor and their respective method factors. Green and Yang (2018)

refer to such models as “bifactor pure-clustered models” and note that model convergence may be problematic with these models, particularly if factor loadings are similar across the general factor and within each specific factor. This was the case in the current study, where general factor loadings were either 0.7 or 0.5 across all items and the method factor loadings were either 0.5 or 0.3 across all items. In pure-clustered bifactor models with uniform loadings across the general and specific factors, Green & Yang reported model convergence rates less than 70%, similar to what was found in the current study.

It is recognized that model convergence in certain conditions was problematic for the bifactor model in this study. However, the bifactor model was the correctly specified model (i.e., the true population model) and serves as a comparison to the misspecified unidimensional model. As such, all conditions in which the bifactor model achieved convergence and proper solutions were analyzed. This resulted in substantially fewer analyzed datasets for the bifactor model in the conditions in which PUC was 0.54 and 0.48.

The Relationship between Factor Loading Bias and ECV, OmegaH, and PUC

One of the outcomes of interest in this study was bias in factor loading estimates when data simulated from a bifactor model were then subsequently fit to a unidimensional model. Results for normally distributed data are presented first, followed by results for non-normally distributed data. Although it was expected that minimal parameter bias would result in models estimated from the correctly specified bifactor model, it was included for comparison to the misspecified unidimensional model. Similar to Reise, Scheines, et al. (2013), results are first considered collapsed across conditions as

Table 4.1 *Model Convergence and Proper Solution Rates for the Bifactor Model*

PUC (N_p:N_n)	ECV	OmegaH	Normally Distributed Data %	Normally Distributed Data N	Non-Normally Distributed Data %	Non-Normally Distributed Data N
0.54 (6:6)	0.84	0.86	61.7	617	56.1	561
	0.74	0.71	60.5	605	54.6	546
	0.74[a]	0.81[a]	68.7	687	55.3	553
	0.66	0.77	68.5	685	55.4	554
	0.60[a]	0.65[a]	66.4	664	56.3	563
	0.50	0.60	67.4	674	62.1	621
0.48 (8:4)	0.84	0.85	58.7	587	54.7	547
	0.77[a]	0.83[a]	64.9	649	59.4	594
	0.74	0.70	57.9	579	56.0	560
	0.66	0.75	67.2	672	61.0	610
	0.64[a]	0.68[a]	63.6	636	57.4	574
	0.50	0.58	65.4	654	60.5	605
0.73 (3:6)	0.88	0.88	97.6	976	97.6	976
	0.79	0.75	87.9	879	88.1	881
	0.77[a]	0.83[a]	98.4	984	96.3	963
	0.72	0.82	100.0	1000	100.0	1000
	0.63[a]	0.68[a]	90.2	902	89.5	895
	0.57	0.67	100.0	1000	100.0	1000

Note. PUC=percent of uncontaminated correlations; ECV=explained common variance; OmegaH=omega hierarchical.
[a]Conditions have unbalanced loadings across the positive and negative wording method factors (PWF, NWF); in these conditions the positive items loaded 0.3 on the PWF and the negative items loaded 0.5 on the NWF.

a function of ECV, omegaH, and PUC, followed by an examination of factor loading bias within each PUC condition.

As uniform factor loadings were used across the general factor in the population model, relative bias was calculated using the average estimated loading on the general factor (or the single factor in the unidimensional model) in relation to the true population

factor loading value. This study included two unbalanced conditions, one condition in which the number of items per method factor was unbalanced (i.e., 8 positively worded items to 4 negatively worded items) and one condition in which the magnitude of the factor loadings across the method factors differed (i.e., loadings of 0.3 on the positively worded item method factor and 0.5 on the negatively worded item method factor). Given the unbalanced nature of these conditions, the relative bias of the estimated general factor loadings was calculated and evaluated across all 12 items, as well as separately across the positive and negative items.

Following the procedures used in similar studies, relative bias greater than |10%| was considered severe or non-negligible in this study (Bonifay et al., 2015; Gu et al., 2017; Reise, Scheines, et al., 2013). Two general factor loading values were included in this study, 0.7 and 0.5. Average estimated factor loadings greater than 0.77 or less than 0.63 (when the population value was 0.7) or greater than 0.55 or less than 0.45 (when the population value was 0.5) were considered as non-negligible bias.

Normally Distributed Data

The Relationship between ECV, OmegaH, and PUC Collapsed across Conditions. Figures 4.1 to 4.3 depict the relationship between ECV (Figure 4.1), omegaH (Figure 4.2), PUC (Figure 4.3) and the relative bias in estimated factor loadings for the general factor for both the misspecified unidimensional model and the true bifactor model collapsed across conditions. As seen in Figures 4.1-4.3, reference lines are placed at bias levels of ±10% to indicate the threshold of acceptable bias levels.

It was expected that the general factor loadings, as estimated from the bifactor model, would illustrate minimal bias given it is the true underlying structure of the data.

Overall, when examining Figures 4.1-4.3, the boxplots of the relative bias in estimated general factor loadings from the bifactor model compared to the unidimensional model, bias was more variable when examining the factor loading bias by item type (i.e., positive and negative items), and also slightly exceeds the |10%| in some conditions. The slight instability of the bifactor model is likely due to the reduced number of datasets analyzed as a result of model non-convergence and improper solutions.

Figure 4.1 shows the distribution (i.e., boxplots) of the relative bias of the estimated general factor loadings as a function of explained common variance (ECV) collapsed across conditions for both the unidimensional model and the bifactor model. In the context of this study, ECV provides an assessment of the strength of the general factor relative to the method factors. Higher ECV values (closer to 1.0) indicate a "strong" general factor and one that accounts for a majority of the variance across all items relative to the method factors. ECV was calculated based on the factor loadings in the population model and ranged 0.50 to 0.88 across conditions.

In comparing the average relative bias in estimated factor loadings across all 12 items to the average bias separately for the positive and negative items, there are conditions in which the factor loadings as estimated from the unidimensional model are over/underestimated and exceed the |10%| threshold for bias. However, this bias is masked when considering all 12 items together. For example, at ECV values of 0.57 – 0.63, the average relative bias of the factor loadings across all items estimated from the unidimensional model falls within acceptable levels. However, in looking separately at the items by wording type within this ECV range, the estimated factor loadings associated with the positively worded items are, on average, severely underestimated,

while the factor loadings associated with the negatively worded items are, on average, severely overestimated. This is likely due to the unbalanced nature of the method factors included in this study and is further examined in the next section, as bias is considered within each PUC condition. In general, however, as ECV increased across all conditions, parameter bias in the factor loadings estimated from the unidimensional model tended to fall within acceptable levels. More specifically, at ECV values ≥ 0.79, the average relative bias across both the positive items and the negative items fell within acceptable levels across all conditions.

Figure 4.2 shows the distribution (i.e., boxplots) of the relative bias of the estimated general factor loadings as a function of omega hierarchical (omegaH), collapsed across conditions for both the unidimensional model and the bifactor model. OmegaH estimates the proportion of variance in unit-weighted total scores attributable only to the general factor (Rodriguez et al., 2016a, 2016b). OmegaH was calculated for each study condition using the population factor loading values and ranged from 0.58 to 0.88 across conditions. Similar to ECV, as omegaH increased, average relative bias in factor loadings estimated from the unidimensional model tended to fall within acceptable levels, however, this relationship is more inconsistent than that of ECV. As noted by Reise, Scheines, et al. (2013), this inconsistency is likely due to the fact that omegaH is more influenced by the magnitude of the general factor loading than ECV. For example, in this study, when PUC was 0.54, and the loading pattern across the general and method factors was 0.7, 0.5, 0.5, this resulted in an ECV of 0.66 and an omegaH of 0.77. However, when the loading pattern across the general and method factors was 0.5, 0.3, 0.3, this resulted in an ECV of 0.74 and an omegaH of 0.71. This also relates to the

notion that ECV is a measure of general factor strength, while omegaH is more a measure of general factor saturation (Reise, Scheines, et al., 2013). As seen in Figure 4.2, omegaH values ≥ 0.85 are associated with acceptable levels of bias, on average, in general factor loadings from both the positive and negative items estimated from the unidimensional model. However, omegaH values of 0.70 – 0.77 are associated with negligible bias, on average, but omegaH values of 0.81 – 0.83 are associated with bias levels that, on average, tended to fall just outside of the acceptable levels.

Figure 4.3 shows the distribution (i.e., boxplots) of the relative bias of the estimated general factor loadings as a function of the percent of uncontaminated correlations (PUC), collapsed across conditions for both the unidimensional model and the bifactor model. PUCs are the percent of item correlations (i.e., covariances) that only reflect general factor variance. Higher PUC values are indicative of a more unidimensional data structure. PUCs are a function of the total number of items, the number of specific factors, and the number of items within each specific factor (Reise, Scheines, et al., 2013). As the current study included a total of 12 items across all conditions, the three PUC values included (0.48, 0.54, and 0.73) were defined by the composition of the method factors. As seen in Figure 4.3, the discrepancy in relative bias in the estimated factor loadings from the unidimensional model when looking at all 12 items combined compared to separately by item wording type continued when examining the data collapsed across PUC values. When PUC was 0.48, an unbalanced condition that included eight items that loaded on the positively worded item method factor and four items that loaded on the negatively worded item method factor, the average relative bias in factor loadings estimated from the unidimensional model fell just outside of the

acceptable level for the positive items but was within acceptable bias levels for the negative items. Conversely, when PUC was 0.54, a balanced condition that included six items each that loaded on their respective method factor, the average relative bias in factor loadings estimated from the unidimensional model were within acceptable bias levels for the positive items but fell just outside of the acceptable bias levels for the negative items. When PUC was 0.73, an unbalanced condition where three of the six positive items loaded on the positively worded item method factor and all six negative items loaded on the negatively worded item method factor, the average relative bias in factor loadings estimated from the unidimensional model fell outside of the acceptable level for both the positive and negative items.

The Relationship between ECV and OmegaH within PUC. Table 4.2 lists the means and standard deviations (presented in parentheses in the table) of the relative bias in general factor loadings estimated from both the unidimensional and bifactor model. The average relative bias is presented within each PUC condition and ordered by highest to lowest ECV value. As noted earlier, although it was expected that the bifactor model, being the correctly specified model, would recover the general factor loadings with minimal bias, inspection of the distribution of the relative bias across conditions indicated that the bifactor model showed signs of instability, as bias was more variable in certain conditions compared to the unidimensional model, and slightly exceeded the |10%| threshold in certain conditions. In examining bias within each PUC condition, this instability occurred within PUC conditions of 0.54 and 0.48. For example, when PUC was 0.48 and ECV was 0.50, the average relative bias in general factor loadings for the

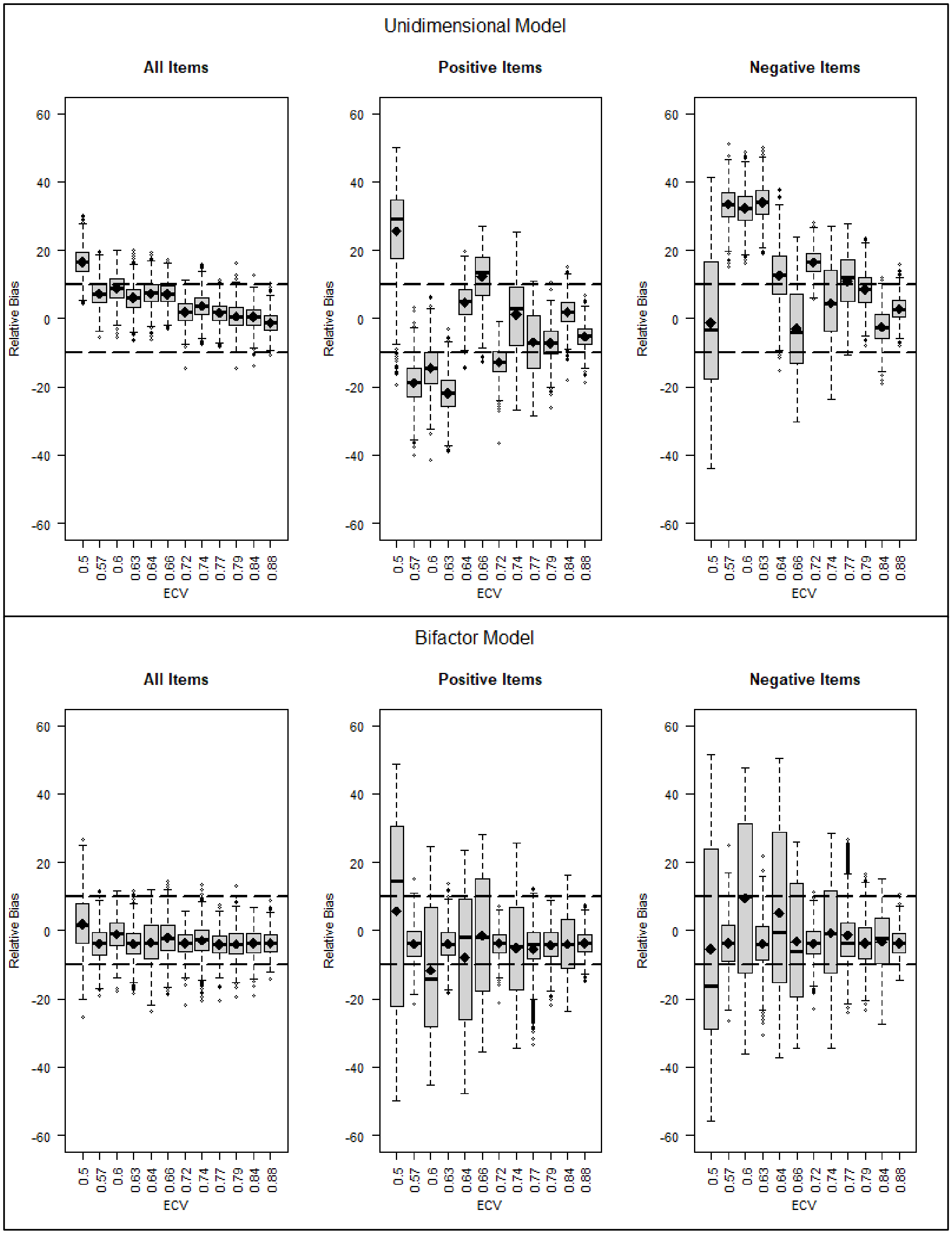

Figure 4.1 *Distribution of the Relative Bias in Estimated General Factor Loadings by Explained Common Variance (ECV), Model Type, and Item Type for Normally Distributed Data.* Average bias is denoted with a black diamond.

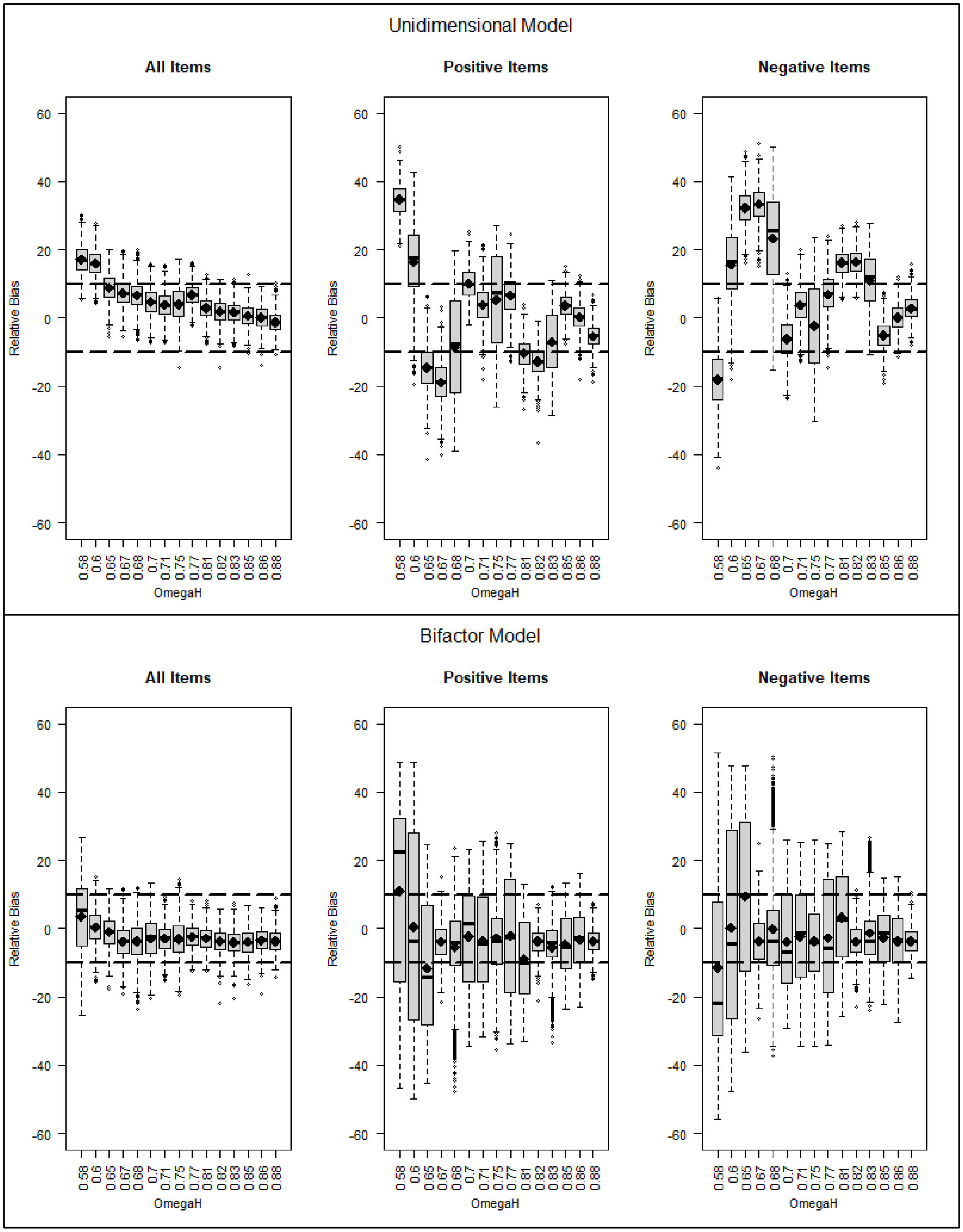

Figure 4.2 *Distribution of the Relative Bias in Estimated General Factor Loadings by Omega Hierarchical (OmegaH), Model Type, and Item Type for Normally Distributed Data.* Average bias is denoted with a black diamond.

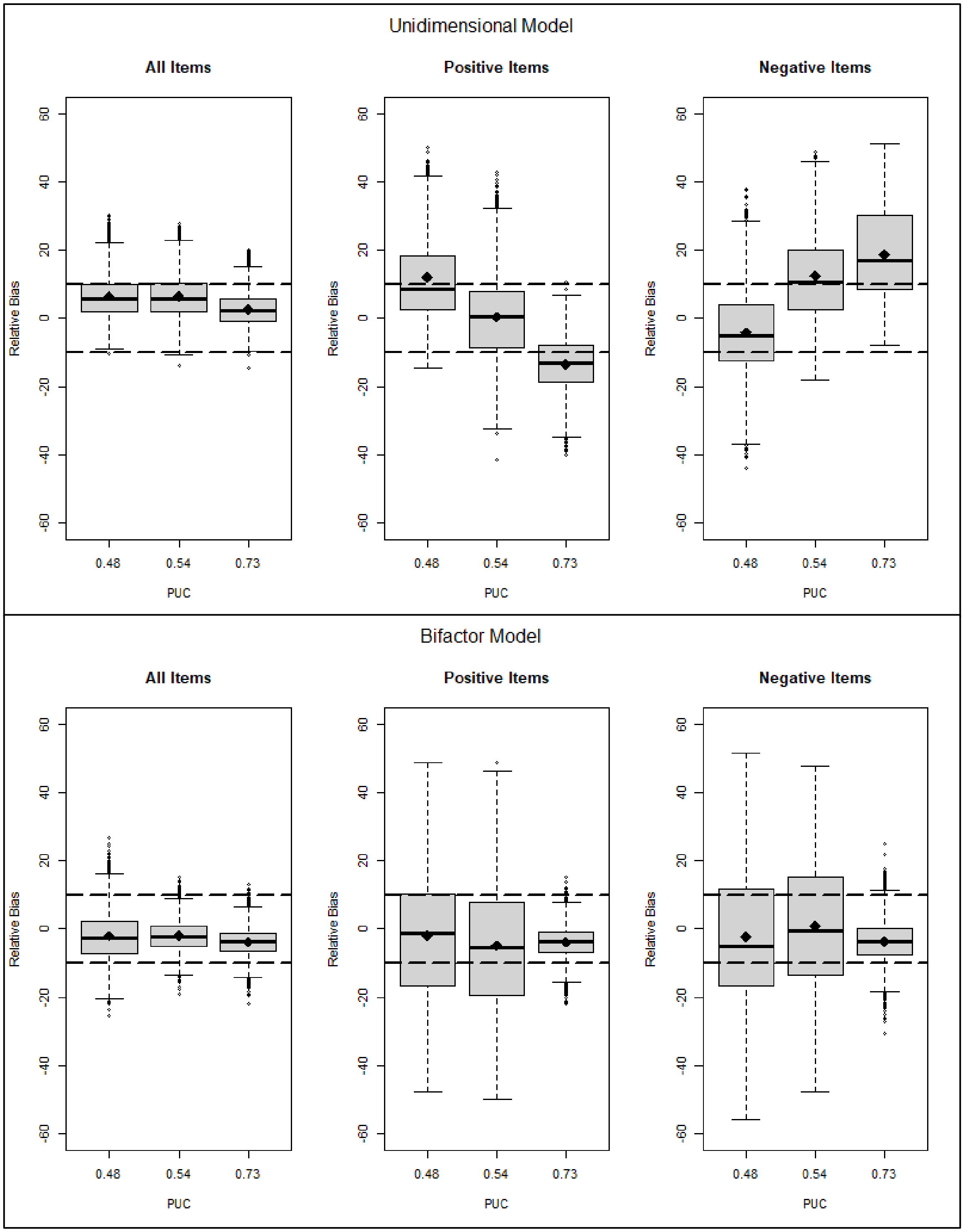

Figure 4.3 *Distribution of the Relative Bias in Estimated General Factor Loadings by the Percent of Uncontaminated Correlations (PUC), Model Type, and Item Type for Normally Distributed Data*. Average bias is denoted with a black diamond.

positively worded items was 10.91% and -11.38% for the negatively worded items, with standard deviations of 25.79 and 25.57, respectively. However, in examining the average relative bias from the bifactor model in the PUC condition of 0.73, all conditions showed minimal bias and less variability. Both the 0.54 and 0.48 PUC conditions showed issues related to model convergence and improper solutions when data was estimated via the bifactor model, whereas the PUC condition of 0.73 did not. Any issues related to the instability of factor loading recovery from the bifactor model is due to the reduction in datasets analyzed in the PUC of 0.54 and 0.48 conditions. In general, general factor loadings estimated from the bifactor model were slightly underestimated, but in most cases, this was within the |10%| threshold for acceptable bias.

When examining the relative bias in general factor loadings as estimated from the unidimensional model, different patterns in bias emerged within each PUC condition. When PUC was 0.54, method factors were balanced in terms of the number of items specified to load on either the positively worded method factor or negatively worded method factor (i.e., 6:6). When factor loadings across the method factors were also balanced, the bias in general factor loadings estimated from the unidimensional model were roughly equivalent across both the positive items and negative items. For example, when ECV, as derived from the population bifactor model, was 0.74, the relative bias associated with factor loadings estimated from the unidimensional model was 3.80%, 3.83%, and 3.76% across all items, the positive items, and the negative items, respectively. However, within this same condition, when unbalanced loadings were present across the method factors (i.e., loadings of 0.3 on the positively worded item method factor, 0.5 on the negatively worded item method factor), estimated general

factor loadings on both the positive and negative items resulted in non-negligible bias, with the factor loadings associated with the negative items being overestimated and the factor loadings associated with the positive items being underestimated.

This differential pattern in factor loading bias was also seen when PUC was 0.48. In this condition, the method factors were unbalanced in that more items loaded on the positively worded method factor than the negatively worded method factor (i.e., 8:4); thus, the positively worded method factor is "stronger" in that it is defined by more items. When balanced loadings were also included across the method factors, general factor loadings on the positive items were overestimated but were underestimated for the negative items. For example, when ECV was 0.74, the average relative bias for the positive items was 10.02% and -6.17% for the negative items. However, when unbalanced loadings were included across the method factors, with higher loadings contained within the negatively worded method factor (i.e., loadings of 0.5), the relative bias in the general factor loadings estimated from the unidimensional model were overestimated for both item types. For example, when ECV was 0.77, and loadings across the method factors were unbalanced, the average relative bias for the positive items was 0.70% and 4.88% for the negative items.

This pattern continued when PUC was 0.73. In this condition, the method factors were defined by including three items on the positively worded method factor and six items on the negatively worded method factor (i.e., a "stronger" negatively worded method factor). Here, relative bias in the general factor loadings estimated from the unidimensional model were overestimated for the negative items and underestimated for the positive items, regardless of the loading magnitude across the method factors (i.e.,

balanced or unbalanced loadings). For example, when ECV was 0.72, the average relative bias in general factor loadings was -12.74% for the positive items and 16.40% for the negative items. Likewise, when ECV was 0.63, and the loadings across the method factors were unbalanced, the average relative bias in the general factor loadings for the positive items was -21.96% and 33.95% for the negative items.

Figure 4.4 plots the average relative bias in the estimated general factor loadings for both the positive items and negative items for the unidimensional and bifactor model by ECV, PUC, and loading type (i.e., balanced or unbalanced loadings on each method factor). Reference lines are placed at the ±10% relative bias threshold. Given the discrepancy in relative bias in estimated factor loadings for the positive and negative items, a condition was considered to have negligible bias when estimated general factor loadings for both the positive and negative items reached acceptable levels.

In their study on the relationship between ECV, omegaH, and PUC as it relates to parameter bias when data that follow a bifactor structure are estimated with a unidimensional model, Reise, Scheines, et al. (2013) offered the following tentative benchmark, when PUC is <0 .80, ECV >0.60 and OmegaH >0.70, fitting data that follow a bifactor structure to a misspecified unidimensional model should result in minimal bias. The parameter bias examined in their study was bias in structural coefficients and not factor loadings. However, Reise, Scheines et al., note that bias in structural coefficients depends, to some degree, on bias in factor loadings (i.e., quality of the measurement model). The current study included PUC conditions that were all less than 0.80, as this is a model structure frequently found with item wording method effects. To examine how

Table 4.2 *Average Percent Relative Bias in Estimated General Factor Loadings: Normally Distributed Data*

Bifactor Model Indices				Average Percent Relative Bias: General Factor Loadings (λ_G)					
				Bifactor Model			Unidimensional Model		
PUC (N_p:N_n)	ECV	OmegaH	True λ_G	All Items	Positive Items	Negative Items	All Items	Positive Items	Negative Items
0.54 (6:6)	0.84	0.86	0.7	-3.50	-3.23	-3.78	0.16	0.24	0.09
				(3.61)	(8.04)	(7.99)	(3.33)	(3.86)	(3.93)
	0.74	0.71	0.5	-2.93	-3.56	-2.29	3.80	3.83	3.76
				(4.13)	(13.91)	(13.58)	(3.83)	(5.57)	(5.50)
	0.74[a]	0.81[a]	0.7	-2.86	-8.95	3.22	2.87	**-10.37**	**16.11**
				(3.47)	(11.49)	(12.59)	(3.23)	**(4.42)**	**(3.66)**
	0.66	0.77	0.7	-2.44	-2.09	-2.80	6.73	6.56	6.90
				(3.68)	(16.67)	(16.91)	(3.32)	(5.86)	(6.03)
	0.60[a]	0.65[a]	0.5	-1.10	**-11.65**	9.46	8.80	**-14.63**	**32.23**
				(4.92)	**(18.58)**	(22.10)	(4.11)	**(6.65)**	**(5.22)**
	0.50	0.60	0.5	0.31	0.41	0.21	**15.98**	**16.39**	**15.58**
				(5.30)	(28.05)	(27.76)	**(4.08)**	**(10.69)**	**(10.23)**
0.48 (8:4)	0.84	0.85	0.7	-3.95	-4.62	-2.60	0.68	3.57	-5.11
				(4.16)	(8.12)	(8.18)	(3.25)	(3.56)	(4.23)
	0.77[a]	0.83[a]	0.7	-4.50	-7.84	2.20	2.09	0.70	4.88
				(4.49)	(11.25)	(13.24)	(3.14)	(3.76)	(4.87)
	0.74	0.70	0.5	-2.86	-2.38	-3.83	4.62	**10.02**	-6.17
				(6.09)	(13.97)	(14.14)	(3.81)	**(4.50)**	(6.17)
	0.66	0.75	0.7	-1.81	-0.82	-3.78	7.51	**17.82**	**-13.12**
				(6.59)	(16.93)	(17.05)	(3.34)	**(3.58)**	**(5.04)**

	0.64[a]	0.68[a]	0.5	-3.61	-8.00	5.17	7.33	4.71	**12.58**
				(6.47)	(18.51)	(22.79)	(3.80)	(5.27)	**(8.20)**
	0.50	0.58	0.5	3.48	**10.91**	**-11.38**	**17.06**	**34.62**	**-18.04**
				(10.16)	(25.79)	(25.57)	**(4.28)**	**(4.62)**	**(8.46)**
	0.88	0.88	0.7	-3.67	-3.68	-3.66	-1.29	-5.38	2.81
				(3.29)	(3.61)	(4.02)	(3.14)	(3.71)	(3.57)
	0.79	0.75	0.5	-3.98	-4.29	-3.67	0.56	-7.30	8.43
				(4.39)	(5.13)	(6.65)	(3.98)	(5.12)	(5.11)
	0.77[a]	0.83[a]	0.7	-3.84	-3.89	-3.80	1.14	**-14.70**	**16.98**
.73				(3.28)	(3.49)	(4.35)	(3.08)	**(3.90)**	**(3.52)**
(3:6)	0.72	0.82	0.7	-3.80	-3.79	-3.82	1.83	**-12.74**	**16.40**
				(3.71)	(3.82)	(4.87)	(3.45)	**(4.35)**	**(3.81)**
	0.63[a]	0.68[a]	0.5	-3.91	-3.95	-3.87	5.99	**-21.96**	**33.95**
				(4.71)	(5.07)	(7.41)	(4.08)	**(5.74)**	**(5.01)**
	0.57	0.67	0.5	-3.84	-3.97	-3.70	7.25	**-18.90**	**33.40**
				(4.69)	(5.35)	(7.65)	(4.04)	**(6.16)**	**(5.12)**

Note. PUC=percent of uncontaminated correlations; ECV=explained common variance; OmegaH=omega hierarchical. Shaded cells indicate cases were bias exceeded |10%|. Standard deviations are in parentheses.

[a]Conditions have unbalanced loadings across the positive and negative wording method factors (PWF, NWF); in these conditions, the positive items loaded 0.3 on the PWF and the negative items loaded 0.5 on the NWF.

the findings from this study relate to the benchmark set by Reise, Scheines, et al., an additional reference line has been placed at ECV of 0.60 across all plots in Figure 4.4.

Figure 4.4a reflects a condition similar to what was included in the Reise, Scheines, et al. (2013) study, in that this condition included both a balanced number of items per method (specific) factor and balanced item loadings across the method factors. As seen in Figure 4.4a, both the positive and negative items' general factor loadings experienced similar bias and the benchmark set by Reise et al. generally holds. In this condition, when ECV was ≥ 0.66 the average relative bias in the general factor loadings fell within acceptable levels as estimated from the unidimensional model. However, as shown in Figure 4.4, the minimum ECV value associated with negligible bias in the estimated factor loadings from the unidimensional model for both item types differs when there is an unbalanced structure in the data. When PUC was 0.48 and loadings were balanced across the method factors (Figure 4.4b), ECV values ≥ 0.74 were associated with acceptable bias levels in factor loadings across both item types estimated from the unidimensional model. When PUC was 0.73 (Figure 4.4c) and balanced loadings were included across the method factors, acceptable bias in factor loadings across both item types estimated from the unidimensional model was achieved at ECV ≥ 0.79.

In conditions in which the loadings on the method factors were unbalanced (i.e., loadings of 0.3 on the positively worded method factor, and 0.5 on the negatively worded method factor), higher ECV values are associated with minimal bias in general factor loadings estimated from the unidimensional model within the same PUC conditions. For example, when PUC was 0.54 (Figure 4.4d), general factor loadings on the negative items did not reach acceptable bias levels when estimated from the unidimensional model

when ECV was 0.74. However, this same ECV value indicated acceptable bias in the factor loadings for both item types estimated from the unidimensional model when the method factor loadings were balanced. Additionally, when PUC was 0.73 (Figure 4.4f), general factor loadings for both the positive items and negative items did not reach acceptable bias levels when estimated from the unidimensional model, at either ECV of 0.63 or 0.77. However, the unidimensional model performed better when PUC was 0.48 (Figure 4.4e) and method factor loadings were unbalanced. In this condition, general factor loadings on both the positive and negative items as estimated from the unidimensional model reached acceptable bias levels when ECV was 0.77.

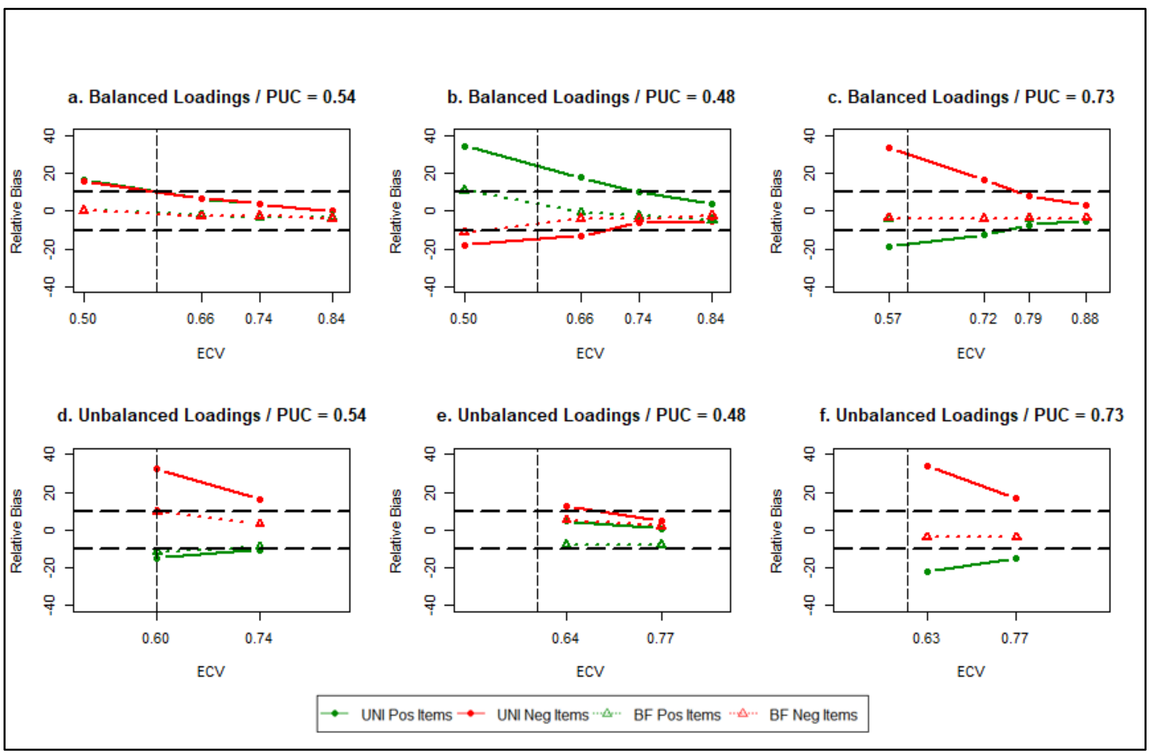

Figure 4.4 *Average Percent Relative Bias in Estimated General Factor Loadings by Explained Common Variance (ECV), Percent of Uncontaminated Correlations (PUC), Loading Type, and Model Type for Normally Distributed Data.* Reference lines placed at ±10% to note acceptable bias threshold. Reference line placed at ECV of 0.60 for comparison to Reise, Scheines, et al. (2013)

Figure 4.5 plots the average relative bias in the estimated general factor loadings for both the positive items and negative items for the unidimensional and bifactor model by omegaH, PUC, and loading type (i.e., balanced or unbalanced loadings on each method factor). Again, reference lines are placed at the ±10% relative bias threshold. An additional reference line is placed at omegaH of 0.70 to reflect the finding by Reise, Scheines, et al. (2013) of PUC $<0\ .80$, ECV >0.60 and omegaH >0.70 as a tentative benchmark for when data that follows a bifactor structure can be treated as essentially unidimensional.

As omegaH is also considered a general factor strength index, the patterns in the average relative bias in general factor loadings estimated from the unidimensional model are similar to those with ECV. When both the number of items and loading patterns were balanced across the method factors (i.e., PUC of 0.54; Figure 4.5a), the benchmark set by Reise, Scheines, et al. (2013) holds in that omegaH values >0.70 were associated with average relative bias in the estimated general factor loadings from the unidimensional model for both item types that were within acceptable levels. Specifically, in this study, when PUC was 0.54, bias in general factor loadings for both the positive and negative items estimated from the unidimensional model reached acceptable levels at omegaH values ≥ 0.71.

Because of the inconsistent relationship between omegaH as an indicator of parameter bias when data is fit to a unidimensional model, due to the fact that omegaH is more effected by the magnitude of the general factor loading, evaluating the relationship between omegaH and general factor loading bias as estimated from the unidimensional model was slightly more complex when PUC was 0.48 (Figure 4.5b) and 0.73 (Figure

4.5c). As shown in Figure 4.5b, when PUC was 0.48 and the method factors contained balanced loadings, the average relative bias in the general factor loadings estimated from the unidimensional model were right at the |10%| threshold for both the positive and negative items. However, the average relative bias slightly exceeded the |10%| threshold when omegaH was 0.75 and then returned to an acceptable level when omegaH was 0.85. A similar pattern was also seen when PUC was 0.73. As shown in Figure 4.5c, average relative bias for both item types reached an acceptable level when omegaH was 0.75, exceed it when omegaH was 0.82, and then returned to an acceptable level when omegaH was 0.88.

Similar to findings related to ECV, when unbalanced loadings were present across the method factors, higher omegaH values are associated with minimal bias in the general factor loadings estimated from the unidimensional model. Specifically, when PUC was 0.54 (Figure 4.5d) and 0.73 (Figure 4.5f) general factor loadings on both the positive and negative items estimated from the unidimensional model did not reach acceptable levels at the highest omegaH values of 0.81 or 0.83. However, when PUC was 0.48 (Figure 4.5e), the average relative bias in the general factor loadings estimated from the unidimensional model for both the positive and negative items were within acceptable levels when omegaH was 0.83.

Non-Normally Distributed Data

In this section, the relative bias in estimated general factor loadings and its relationship to ECV, omegaH, and PUC is examined under conditions of non-normality. In this condition, item distributions on the negatively worded items exhibited extreme

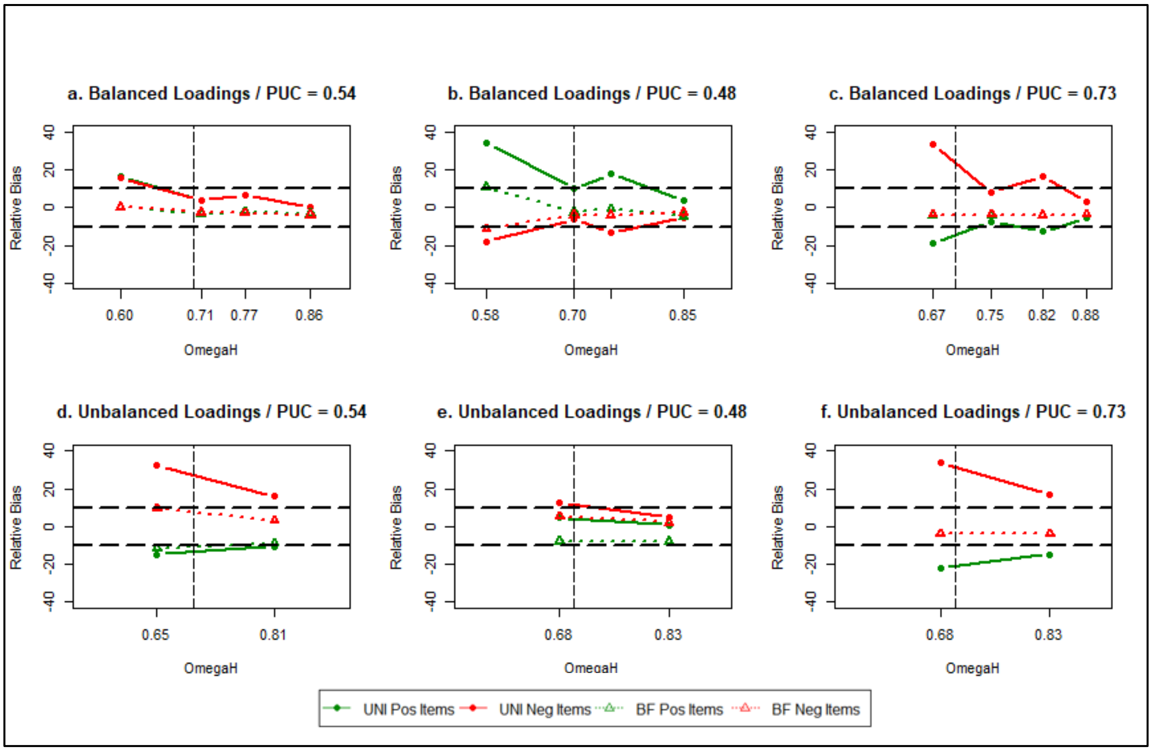

Figure 4.5 *Average Percent Relative Bias in Estimated General Factor Loadings by Omega Hierarchical (OmegaH), Percent of Uncontaminated Correlations (PUC), Loading Type, and Model Type for Normally Distributed Data.* Reference lines placed at ±10% to note acceptable bias threshold. Reference line placed at omegaH of 0.70 for comparison to Reise, Scheines, et al. (2013)

non-normality with approximate skewness and kurtosis values of -3.0 and 7.0, respectively.

The Relationship between ECV, OmegaH, and PUC Collapsed across Conditions. Figures 4.6 to 4.8 depict the relationship between ECV (Figure 4.6), omegaH (Figure 4.7), PUC (Figure 4.8) and the relative bias in estimated factor loadings for the general factor for both the misspecified unidimensional model and the correctly specified bifactor model collapsed across conditions. Figure 4.6 shows the distribution of the relative bias for both the unidimensional and bifactor model for all items and separately by item type (positive and negative) as a function of ECV. Compared to the

normally distributed data condition, the relative bias in the general factor loadings estimated from the unidimensional model is more severe when non-normality is present, and no ECV value is associated with an acceptable level of bias for both the positive and negative items. For example, when ECV was ≥ 0.79, the average relative bias in estimated factor loadings from the unidimensional model were within acceptable bias levels for the positive items but were severely underestimated for the negative items, falling well below the -10% threshold. Unexpectedly, the general factor loadings estimated from the correctly specified bifactor model were also severely biased, and in some cases, showed more severe bias than the unidimensional model. For example, when ECV was 0.63, the average relative bias in estimated factor loadings for the negative items was within acceptable bias levels as estimated for the unidimensional model but fell well below the -10% threshold as estimated from the bifactor model. Similar patterns are also evident when examining bias as a function of omegaH collapsed across conditions (Figure 4.7).

Figure 4.8 shows the distribution of the relative bias in estimated general factors loadings as a function of PUC for both the unidimensional model and bifactor model. Again, similar patterns seen with both ECV and omegaH are evident within each PUC condition in that no PUC value is associated with negligible bias for both the positive items and negative items for either the correctly specified bifactor model or the unidimensional model. For both models, more severe bias is associated with the estimated general factor loadings for the negative items or the items that displayed non-normal distributions. Slightly different patterns were seen within each PUC condition for both models. For example, on average, the relative bias of general factor loadings on the

negative items is slightly more severe for the unidimensional model when PUC was 0.48, similar between both models when PUC was 0.54, and slightly more severe for the bifactor model when PUC was 0.73.

The Relationship between ECV and OmegaH within PUC. Table 4.3 lists the means and standard deviations (presented in parentheses in the table) of the relative bias in general factor loadings estimated from both the unidimensional and bifactor model for the non-normally distributed data. As seen in Table 4.3, most concerning is the severe bias in estimated general factor loadings associated with the bifactor model across all conditions. For example, when PUC was 0.54, the estimated general factor loadings across all items, as well as separately between item type, were severely biased across all ECV and omegaH values contained within this condition for the bifactor model. Within this condition, the average relative bias across all items ranged -24.93% to -19.73% for the bifactor model.

Across all conditions, models were estimated using MLMV (maximum likelihood with mean and variance adjusted). MLMV is considered a robust maximum likelihood (ML) estimator and uses a corrected chi-square statistic to accommodate data that is non-normally distributed (Maydeu-Olivares, 2017). While it has been shown to perform well under conditions of non-normality (i.e., Maydeu-Olivares, 2017), its performance has not been examined with bifactor models. Given the unexpected issues regarding factor loading recovery with the correctly specified bifactor model, a subset of data where PUC was 0.54 was re-analyzed using WLSMV (weighted least squares with mean and variance adjusted) to examine the role of estimation as a potential confounding variable. WLSMV is a robust WLS (weighted least squares) estimator used exclusively with

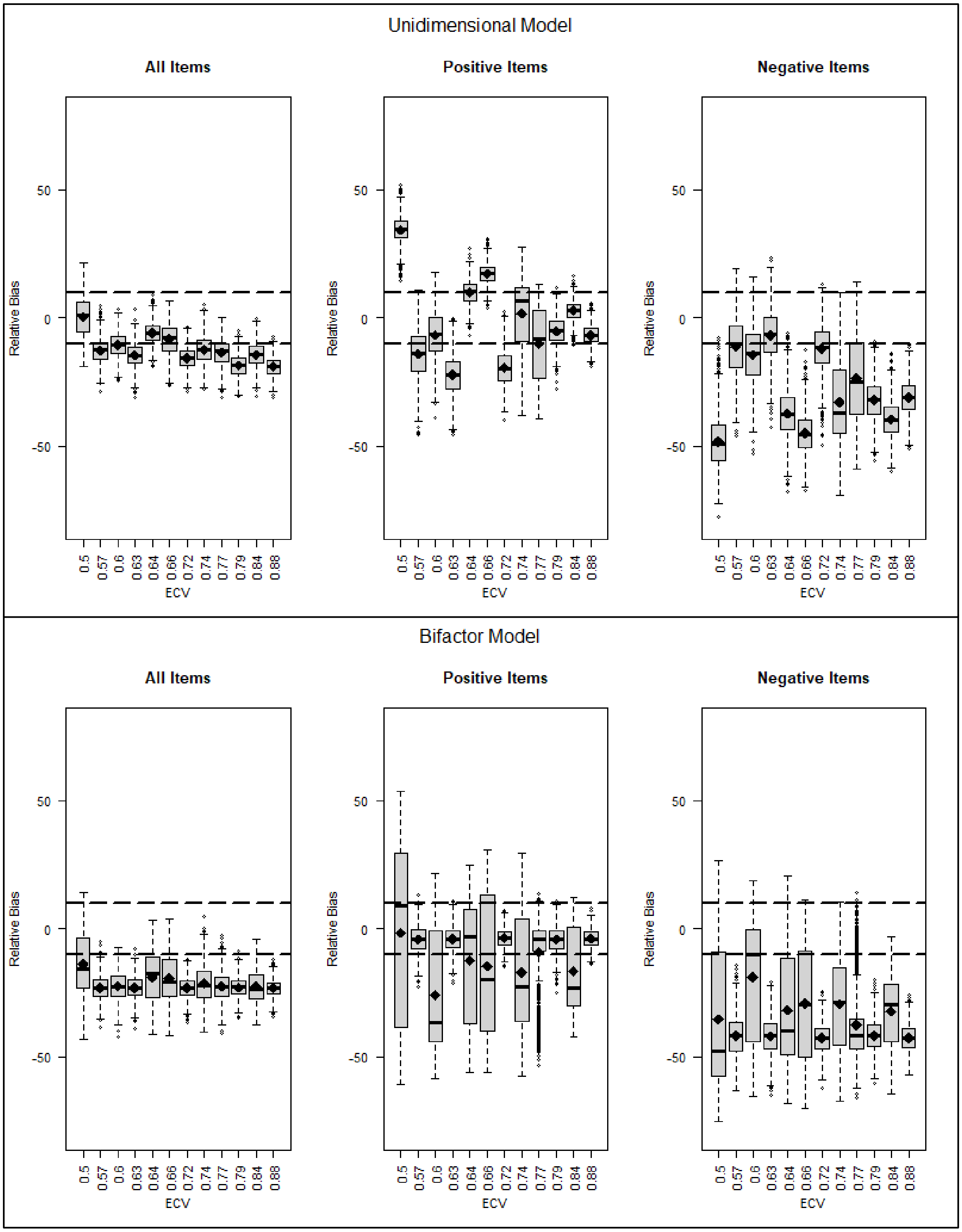

Figure 4.6 *Distribution of the Relative Bias in Estimated General Factor Loadings by Explained Common Variance (ECV), Model Type, and Item Type for Non-Normally Distributed Data*. Average bias is denoted with a black diamond.

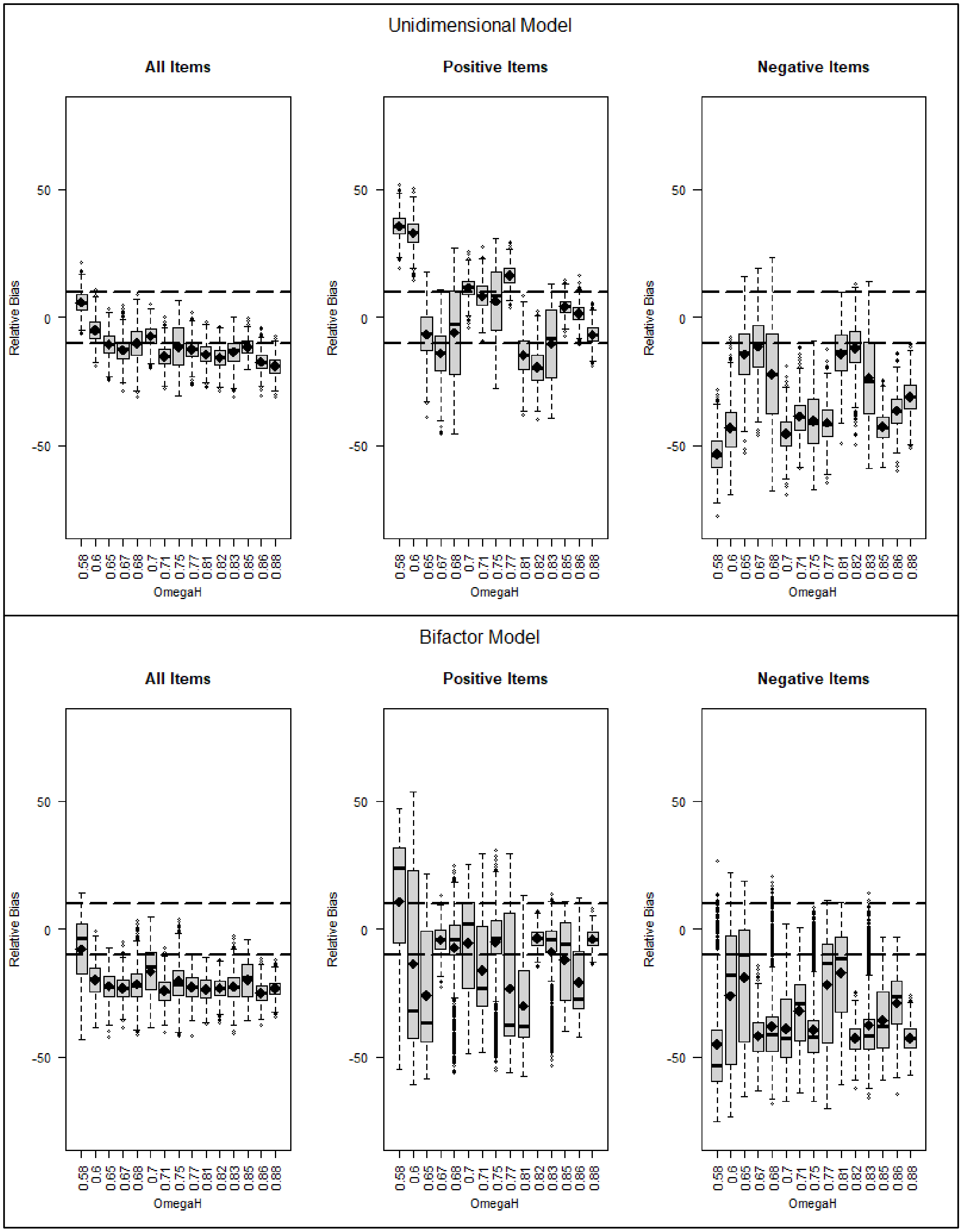

Figure 4.7 *Distribution of the Relative Bias in Estimated General Factor Loadings by Omega Hierarchical (OmegaH), Model Type, and Item Type for Non-Normally Distributed Data.* Average bias is denoted with a black diamond.

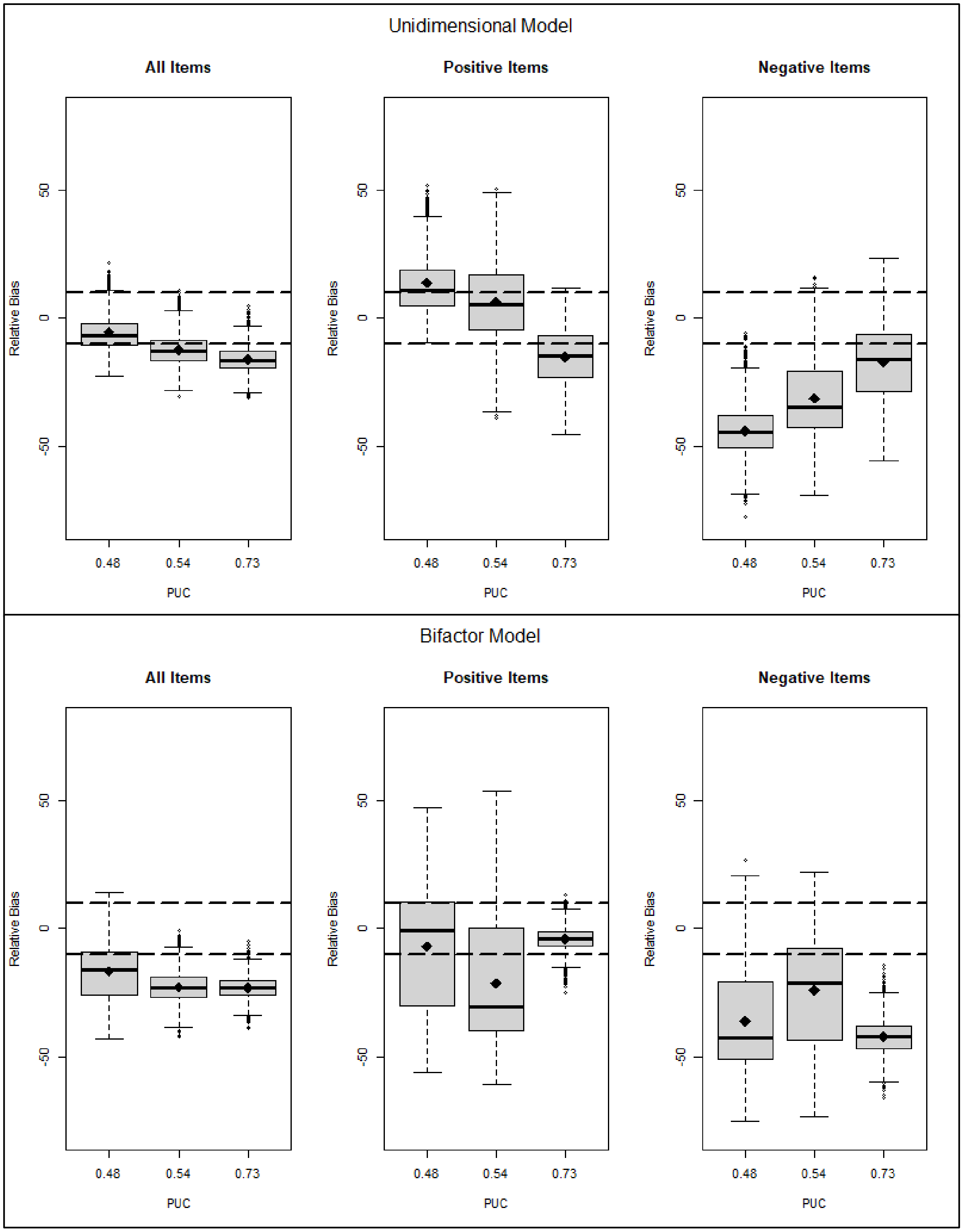

Figure 4.8 *Distribution of the Relative Bias in Estimated General Factor Loadings by Percent of Uncontaminated Correlations (PUC), Model Type, and Item Type for Non-Normally Distributed Data.* Average bias is denoted with a black diamond.

categorial or ordered data and has been shown to produce accurate parameter estimates under conditions of extreme non-normality (e.g., DiStefano & Morgan, 2014; Flora & Curran, 2004; Lei, 2009; Li, 2016).

Table 4.4 lists the means and standard deviations of the relative bias in general factor loadings estimated from both the unidimensional and bifactor model for non-normally distributed data using WLSMV estimation. With WLSMV estimation, factor loading recovery for the bifactor model was more satisfactory, and more closely resembled the patterns with the normally distributed data estimated using MLMV. Given the focus of the study, proper estimation of the general factor loadings from the correctly specified bifactor model is critical in order to reach accurate conclusions regarding factor loadings as estimated from the misspecified unidimensional model. Due to issues found with model estimation and non-normality, subsequent analyses investigating the relationship between ECV, omegaH, PUC, and relative bias in estimated factor loadings from the unidimensional model, as well as the relationship to model fit indices are not reported for the non-normally distributed data in Chapter 4. For completeness, tables and figures examining these relationships for the non-normally distributed data are presented in Appendix B, including plots examining the relationship between the relative bias in estimated factor loadings from the unidimensional model by study conditions and ECV (Figure B.1) and omegaH (Figure B.2). Additionally, results examining the relationship between model fit indices CFI and TLI (Table B.1 and Figure B.3), and AIC and BIC (Table B.2 and Figure B.4) between the bifactor model and the unidimensional model for the non-normally distributed data are also presented in Appendix B.

Table 4.3 *Average Percent Relative Bias in Estimated General Factor Loadings: Non-Normally Distributed Data*

Bifactor Model Indices				Average Percent Relative Bias: General Factor Loadings (λ_G)					
				Bifactor Model			Unidimensional Model		
PUC (N_p:N_n)	ECV	OmegaH	True λ_G	All Items	Positive Items	Negative Items	All Items	Positive Items	Negative Items
0.54 (6:6)	0.84	0.86	0.7	**-24.93** **(4.23)**	**-20.88** **(14.23)**	**-28.98** **(12.05)**	**-17.39** **(3.79)**	1.57 (3.73)	**-36.34** **(6.72)**
	0.74	0.71	0.5	**-24.07** **(5.39)**	**-16.06** **(18.02)**	**-32.08** **(13.44)**	**-15.15** **(4.24)**	8.41 (5.34)	**-38.72** **(7.46)**
	0.74[a]	0.81[a]	0.7	**-23.56** **(4.92)**	**-29.96** **(17.98)**	**-17.15** **(17.77)**	**-14.49** **(4.31)**	**-14.73** **(7.98)**	**-14.25** **(10.34)**
	0.66	0.77	0.7	**-22.56** **(5.15)**	**-23.31** **(24.97)**	**-21.81** **(20.75)**	**-12.33** **(4.15)**	**16.43** **(3.83)**	**-41.09** **(7.63)**
	0.60[a]	0.65[a]	0.5	**-22.36** **(5.57)**	**-25.89** **(23.27)**	**-18.82** **(22.79)**	**-10.54** **(4.71)**	-6.62 (9.51)	**-14.45** **(11.60)**
	0.50	0.60	0.5	**-19.78** **(7.15)**	**-13.51** **(33.77)**	**-26.05** **(26.52)**	-5.08 (4.96)	**33.02** **(5.54)**	**-43.18** **(9.89)**
0.48 (8:4)	0.84	0.85	0.7	**-19.88** **(7.16)**	**-12.01** **(15.55)**	**-35.63** **(12.76)**	**-11.50** **(3.21)**	4.14 (3.31)	**-42.76** **(5.87)**
	0.77[a]	0.83[a]	0.7	**-21.45** **(8.00)**	**-17.40** **(19.78)**	**-29.54** **(19.03)**	**-10.27** **(3.48)**	3.26 (3.55)	**-37.32** **(7.10)**
	0.74	0.70	0.5	**-16.56** **(8.94)**	-5.41 (18.69)	**-38.85** **(14.43)**	-7.41 (3.95)	**11.59** **(4.42)**	**-45.39** **(6.99)**
	0.66	0.75	0.7	**-16.36** **(10.81)**	-6.52 (24.87)	**-36.03** **(20.32)**	-4.23 (3.56)	**18.05** **(3.63)**	**-48.79** **(6.36)**

	0.64[a]	0.68[a]	0.5	**-18.87**	**-12.40**	**-31.81**	-5.79	**10.00**	**-37.37**
				(9.62)	**(23.10)**	**(21.95)**	(4.18)	**(4.71)**	**(9.74)**
	0.50	0.58	0.5	-7.89	**10.72**	**-45.10**	5.98	**35.69**	**-53.45**
				(12.94)	**(28.26)**	**(21.71)**	(4.18)	**(4.59)**	**(7.54)**
0.73 (3:6)	0.88	0.88	0.7	**-23.24**	-3.80	**-42.69**	**-18.95**	-6.80	**-31.09**
				(3.59)	(3.53)	**(5.23)**	**(3.86)**	(3.90)	**(6.82)**
	0.79	0.75	0.5	**-22.96**	-4.21	**-41.71**	**-18.64**	-5.20	**-32.08**
				(4.04)	(5.04)	**(6.33)**	**(4.20)**	(5.32)	**(7.80)**
	0.77[a]	0.83[a]	0.7	**-23.20**	-3.71	**-42.69**	**-16.86**	**-23.79**	-9.93
				(3.87)	(3.48)	**(6.11)**	**(4.24)**	**(5.52)**	(8.10)
	0.72	0.82	0.7	**-23.16**	-3.67	**-42.64**	**-15.67**	**-19.39**	**-11.96**
				(3.96)	(3.69)	**(6.01)**	**(4.21)**	**(7.20)**	**(9.25)**
	0.63[a]	0.68[a]	0.5	**-23.07**	-4.13	**-42.01**	**-14.59**	**-22.24**	-6.94
				(4.46)	(5.13)	**(7.14)**	**(4.68)**	**(8.00)**	(10.00)
	0.57	0.67	0.5	**-23.04**	-4.25	**-41.83**	**-12.69**	**-14.08**	**-11.30**
				(4.75)	(5.32)	**(7.67)**	**(4.81)**	**(9.43)**	**(11.77)**

Note. PUC=percent of uncontaminated correlations; ECV=explained common variance; OmegaH=omega hierarchical. Shaded cells indicate cases were bias exceeded |10%|. Standard deviations are in parentheses.

[a]Conditions have unbalanced loadings across the positive and negative wording method factors (PWF, NWF); in these conditions, the positive items loaded 0.3 on the PWF and the negative items loaded 0.5 on the NWF.

Table 4.4 *Average Percent Relative Bias in Estimated General Factor Loadings: Non-Normally Distributed Data with WLSMV Estimation*

Bifactor Model Indices				Average Percent Relative Bias: General Factor Loadings (λ_G)					
				Bifactor Model			Unidimensional Model		
PUC (N_p:N_n)	ECV	OmegaH	True λ_G	All Items	Positive Items	Negative Items	All Items	Positive Items	Negative Items
0.54 (6:6)	0.84	0.86	0.7	0.39 (2.71)	0.26 (7.11)	0.52 (7.38)	4.77 (2.29)	7.12 (2.02)	2.41 (3.94)
	0.74	0.71	0.5	1.14 (4.65)	0.68 (13.66)	1.61 (13.96)	8.14 (4.07)	**13.30 (4.24)**	2.99 (7.15)
	0.74[a]	0.81[a]	0.7	0.84 (3.25)	-5.24 (9.57)	6.91 (11.16)	**10.50 (2.08)**	3.90 (2.33)	**17.11 (3.44)**
	0.66	0.77	0.7	1.15 (3.38)	-2.27 (13.83)	4.58 (14.22)	**15.81 (2.17)**	**20.04 (1.62)**	**11.58 (4.26)**
	0.60[a]	0.65[a]	0.5	2.30 (5.64)	-7.32 (17.14)	**11.93 (21.60)**	**16.88 (3.95)**	6.47 (4.91)	**27.29 (7.36)**
	0.50	0.60	0.5	3.08 (6.34)	0.48 (24.45)	5.67 (24.56)	**23.03 (4.62)**	**35.71 (3.88)**	**10.35 (9.47)**

Note. PUC=percent of uncontaminated correlations; ECV=explained common variance; OmegaH=omega hierarchical. Shaded cells indicate cases were bias exceeded |10%|. Standard deviations are in parentheses.

[a]Conditions have unbalanced loadings across the positive and negative wording method factors (PWF, NWF); in these conditions, the positive items loaded 0.3 on the PWF and the negative items loaded 0.5 on the NWF.

The Relationship between ECV, PUC, and Model Fit Indices

The second outcome investigated in this study was the performance of select model fit indices when data was misspecified to a unidimensional model. Both ECV and PUC are considered statistics of unidimensionality (Rodriguez et al., 2016a; 2016b). This section examines the relationship between ECV, PUC, and the model fit indices CFI, TLI, AIC, and BIC, as these indices include larger penalties for model complexity, or conversely, favor model parsimony. As noted in the previous section, all results presented in this section are for normally distributed data.

Table 4.5 presents the means and standard deviations for CFI and TLI values for both the bifactor model and the unidimensional model when item distributions were normally distributed. The difference in average CFI and TLI values between the bifactor model and the unidimensional model (i.e., ΔCFI, ΔTLI) is also presented. Average fit values are presented within each PUC level, ordered highest to lowest by ECV. Average CFI and TLI values ≥ 0.95 were considered as indicating acceptable model-data fit (Hu & Bentler, 1999).

As expected, the bifactor model showed exceptional fit, with average CFI and TLI of 1.00 across all conditions. Within each PUC level, average fit reached acceptable levels for the unidimensional model across several conditions, with CFI reaching the ≥ 0.95 threshold slightly more often than TLI. In general, the conditions reaching acceptable fit in the unidimensional model are conditions associated with high ECV values and ECV values, as derived from the bifactor model, that indicated acceptable bias in general factor loadings when estimated from the unidimensional model. For example, within the PUC of 0.54 condition, when ECV was 0.84, the average CFI was 0.95.

Within the PUC of 0.48 condition, the average CFI reached acceptable levels (≥ 0.95) when ECV was 0.84 and 0.74, and TLI ≥ 0.95 when ECV was 0.84. Within the PUC of 0.73 condition, the average CFI and TLI values reached acceptable levels when ECV was 0.79 and 0.88.

More moderate ECV values that were also associated with acceptable bias in general factor loadings estimated from the unidimensional model did not reach acceptable CFI or TLI values. For example, when PUC was 0.54, ECV values of 0.74 and 0.66 are associated with CFI and TLI values that fall outside of the ≥ 0.95 threshold. Similarly, when PUC was 0.48 and ECV was 0.77, the average CFI and TLI values associated with this condition were 0.88 and 0.86, respectively.

The unidimensional and bifactor model form a nested structure, with the unidimensional model nested within the bifactor model. As such, the models can be compared by examining practical change in fit. For CFI and TLI a difference in fit values ≥ .01 indicates improved model-fit fit (Cheung & Rensvold, 2002; Gignac, 2016). In the context of this study, CFI and TLI that exceed ≥ 0.01 indicates a practical improvement in fit between the bifactor model and unidimensional model, favoring the bifactor model. Differences lower than the 0.01 threshold would indicate situations in which the unidimensional model would be retained.

As shown in Table 4.5, the difference in CFI and TLI exceeded the ≥ 0.01 threshold in all cases, indicating practical improvement in fit for the bifactor model. However, at higher ECV values, and when balanced loadings across the method factors were present, the differences in CFI and TLI values between the bifactor and unidimensional model were close.

Table 4.5 *Average CFI and TLI Fit Values for the Bifactor Model and Unidimensional Model: Normally Distributed Data*

		Bifactor Model				Unidimensional Model					
PUC		CFI		TLI		CFI		TLI		Difference[b]	
(N_p:N_n)	ECV	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	ΔCFI	ΔTLI
0.54 (6:6)	**0.84**	**1.00**	0.002	**1.00**	0.005	**0.95**	0.010	0.94	0.013	0.05	0.06
	0.74	**1.00**	0.004	**1.00**	0.012	0.93	0.021	0.91	0.025	0.07	0.09
	0.74[a]	**1.00**	0.001	**1.00**	0.004	0.89	0.015	0.87	0.018	**0.11**	**0.14**
	0.66	**1.00**	0.001	**1.00**	0.003	0.78	0.021	0.73	0.025	**0.22**	**0.27**
	0.60[a]	**1.00**	0.003	**1.00**	0.009	0.87	0.023	0.84	0.029	**0.13**	**0.16**
	0.50	**1.00**	0.002	**1.00**	0.007	0.74	0.029	0.69	0.036	**0.25**	**0.32**
0.48 (8:4)	**0.84**	**1.00**	0.002	**1.00**	0.005	**0.96**	0.009	**0.95**	0.011	0.04	0.05
	0.77[a]	**1.00**	0.002	**1.00**	0.004	0.88	0.016	0.86	0.020	**0.12**	**0.14**
	0.74	**1.00**	0.004	**1.00**	0.012	**0.95**	0.017	0.94	0.021	0.05	0.06
	0.66	**1.00**	0.001	**1.00**	0.003	0.86	0.014	0.83	0.017	**0.14**	**0.17**
	0.64[a]	**1.00**	0.003	**1.00**	0.010	0.85	0.027	0.81	0.033	**0.15**	**0.19**
	0.50	**1.00**	0.002	**1.00**	0.007	0.84	0.022	0.80	0.027	**0.16**	**0.20**
0.73 (3:6)	**0.88**	**1.00**	0.002	**1.00**	0.005	**0.97**	0.008	**0.97**	0.010	0.03	0.03
	0.79	**1.00**	0.006	**1.00**	0.014	**0.96**	0.017	**0.95**	0.021	0.03	0.05
	0.77[a]	**1.00**	0.002	**1.00**	0.004	0.93	0.012	0.91	0.014	0.07	0.09
	0.72	**1.00**	0.002	**1.00**	0.004	0.87	0.014	0.85	0.018	**0.12**	**0.15**
	0.63[a]	**1.00**	0.004	**1.00**	0.009	0.92	0.018	0.90	0.022	0.08	0.09
	0.57	**1.00**	0.004	**1.00**	0.009	0.85	0.023	0.82	0.028	**0.14**	**0.18**

Note. CFI= Comparative Fit Index, TLI = Tucker-Lewis Index. Green shaded ECV values indicate conditions in which the average relative bias for the general factor loadings estimated from the unidimensional model were within acceptable bias levels. Gray shaded boxes indicate cases where average fit was within the recommended value of acceptable fit (≥ .95).

[a]Conditions have unbalanced loadings across the positive and negative wording method factors (PWF, NWF); in these conditions, the positive items loaded 0.3 on the PWF and the negative items loaded 0.5 on the NWF.

[b]Difference in fit is the average bifactor model fit index – the average unidimensional model fit index; differences >.10 are bolded.

The difference in CFI and TLI model fit values between the bifactor model and unidimensional model is further examined in Figure 4.9. Figure 4.9 plots the difference in average CFI (i.e., ΔCFI) and TLI (ΔTLI) by ECV within each PUC condition. Within Figure 4.9, a reference line has been placed at the ΔCFI and ΔTLI mark of 0.01. As shown in Figure 4.9, the pattern in the difference in fit across the three PUC values was similar. As expected, the difference in CFI and TLI values was highest at the lowest ECV values and became substantially smaller at the highest ECV values. In general, as ECV increased, the difference in fit between the unidimensional model became smaller. This pattern was more pronounced for the conditions in which the magnitude in loadings across the method factors was balanced.

In this study, by definition of PUC and ECV, the most "unidimensional structure" occurred when PUC was 0.73 and ECV was 0.88 (Figures 4.9c and 4.9f). Under this condition, the difference between the bifactor and unidimensional models was small, with a difference in average CFI and TLI of 0.03. The least unidimensional structure occurred when PUC was 0.48 and ECV was 0.50 (Figure 4.9b). The difference in fit between the bifactor and unidimensional model at this condition was large, with a difference of 0.16 in CFI and 0.20 in TLI. However, the largest overall difference in both CFI and TLI occurred when PUC was 0.54 and ECV was 0.50 (Figure 4.9a). In this condition, the difference in CFI and TLI between the bifactor and unidimensional model was 0.25 and 0.32, respectively. Across all graphs depicted in Figure 4.9 the pattern in the difference in fit between CFI and TLI between the bifactor model and unidimensional model was largest when PUC was 0.54, narrowed when PUC was 0.48, and was closest when PUC was 0.73.

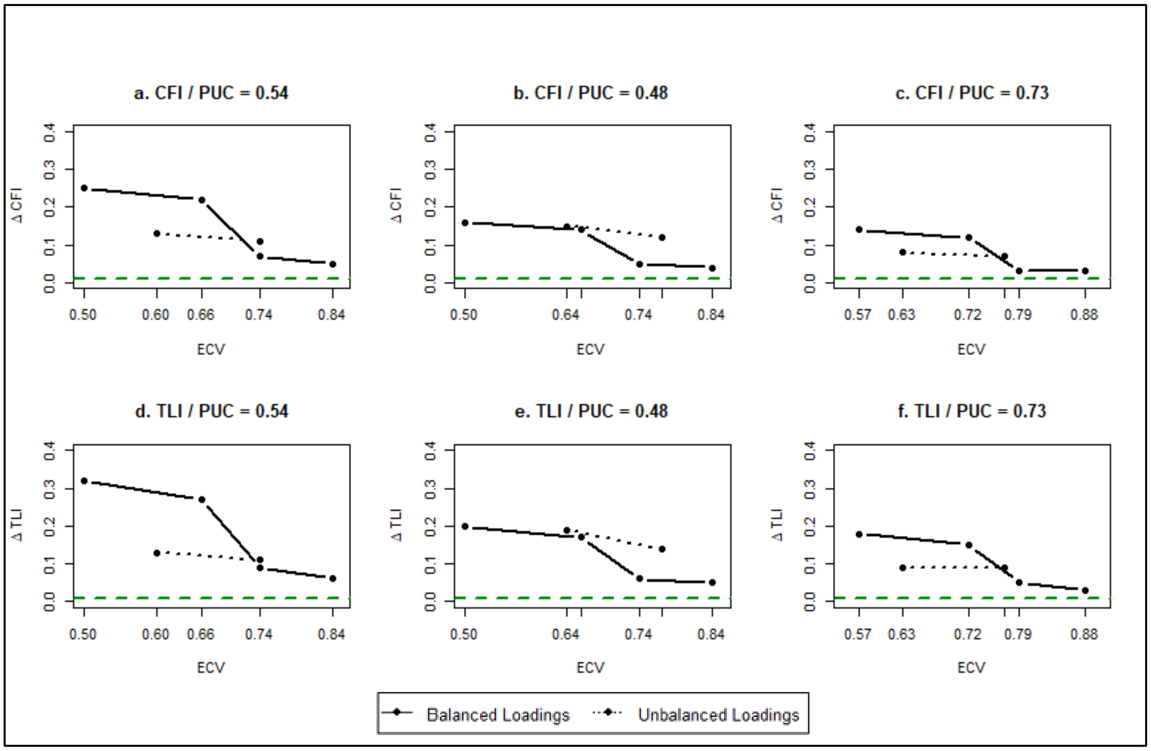

Figure 4.9 *Difference in Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) between the Bifactor Model and the Unidimensional Model by Explained Common Variance (ECV), Percent of Uncontaminated Correlations (PUC), and Loading Type for Normally Distributed Data.* Reference line placed at a difference of 0.01.

Table 4.6 lists means and standard deviations of AIC and BIC values obtained for the bifactor model and unidimensional model. The table is organized by PUC value with ECV ordered highest to lowest within each PUC. The actual value of AIC and BIC is not informative for model fit purposes. Instead, the comparison of AIC and BIC between models offers information regarding model selection, with lower values indicating better model-data fit. More specifically, Raftery (1995) has suggested AIC/BIC differences between models less than 10 (<10) as indicating practical improvement in model fit. Table 4.6 lists the differences in the average AIC and BIC values between the bifactor model and the unidimensional model to compare to this guideline.

In all conditions, the difference in AIC and BIC exceeded the <10 threshold, favoring the bifactor model. There are two conditions in which BIC, on average, was lower for the unidimensional model compared to the bifactor model. This occurred when PUC was 0.48 and ECV was 0.74 and when PUC was 0.73 and ECV was 0.79. Both conditions were associated with ECV values that indicated acceptable bias in general factor loadings estimated from the unidimensional model, with the PUC of 0.48 and ECV of 0.74 being right at the acceptable bias threshold. Although BIC is lower for the unidimensional model in these conditions, the difference does not meet the <10 threshold set by Raftery (1995).

Figure 4.10 shows the difference in AIC and BIC (i.e., ΔAIC and ΔBIC) between the bifactor and unidimensional model across PUC and ECV values. A reference line has been placed at the -10 mark. Patterns between AIC and BIC were similar. However, compared to CFI and TLI, there was a slightly more inconsistent pattern between the values of ECV and the difference in AIC and BIC between the bifactor and unidimensional model. When balanced loadings were present across the method factors, higher ECV values were associated with smaller differences in average AIC and BIC values. However, when unbalanced loadings were included across the method factors, higher ECV values were associated with larger differences in both AIC and BIC. For example, when PUC was 0.54 (Figure 4.10a and Figure 4.10d), and method factor loadings were unbalanced (i.e., loadings of 0.3 on the positively worded item method factor, 0.5 on the negatively worded item method factor), the difference in AIC between the models was -192.72 and -142.15 for BIC when ECV was 0.60. In this same condition

Table 4.6 *Average AIC and BIC Fit Values for the Bifactor Model and Unidimensional Model: Normally Distributed Data*

		Bifactor Model				Unidimensional Model					
PUC		AIC		BIC		AIC		BIC		Difference[b]	
(N_p:N_n)	ECV	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	ΔAIC	ΔBIC
0.54 (6:6)	**0.84**	**14315.54**	103.19	**14517.84**	103.19	14448.61	107.17	14600.34	107.17	-133.07	-82.50
	0.74	**16088.74**	99.69	**16291.05**	99.69	16156.69	99.54	16308.41	99.54	-67.95	-17.36
	0.74[a]	**13551.53**	98.25	**13753.84**	98.25	13946.87	116.23	14098.59	116.23	-395.34	-344.75
	0.66	**12766.68**	104.30	**12968.98**	104.30	13725.71	134.11	13877.43	134.11	-959.03	-908.45
	0.60[a]	**15665.75**	95.17	**15868.05**	95.17	15858.47	100.42	16010.20	100.42	-192.72	-142.15
	0.50	**15223.05**	100.56	**15425.35**	100.56	15720.07	114.99	15871.80	114.99	-497.02	-446.45
0.48 (8:4)	**0.84**	**14287.26**	101.65	**14489.56**	101.65	14382.31	104.01	14534.04	104.01	-95.05	-44.48
	0.77[a]	**13858.98**	101.41	**14061.28**	101.41	14247.98	112.86	14399.70	112.86	-389.00	-338.42
	0.74	**16073.28**	98.22	16275.58	98.22	16118.43	99.96	**16270.16**	99.96	-45.15	5.42
	0.66	**12727.87**	98.59	**12930.17**	98.59	13346.08	114.26	13497.80	114.26	-618.21	-567.63
	0.64[a]	**15856.86**	97.21	**16059.16**	97.21	16047.23	102.56	16198.96	102.56	-190.37	-139.80
	0.50	**15188.28**	99.24	**15390.58**	99.24	15501.04	108.89	15652.77	108.89	-312.76	-262.19
0.73 (3:6)	**0.88**	**14514.72**	103.16	**14704.38**	103.16	14578.59	105.02	14730.32	105.02	-63.87	-25.94
	0.79	**16209.42**	98.09	16399.08	98.09	16237.66	99.25	**16389.39**	99.25	-28.24	9.69
	0.77[a]	**13766.33**	104.14	**13955.99**	104.14	14001.18	111.08	14152.90	111.08	-234.85	-196.91
	0.72	**13532.56**	101.32	**13722.22**	101.32	13986.54	112.32	14138.27	112.32	-453.98	-416.05
	0.63[a]	**15792.44**	97.29	**15982.10**	97.29	15891.38	99.78	16043.11	99.78	-98.94	-61.01
	0.57	**15660.52**	97.99	**15850.18**	97.99	15873.74	101.08	16025.47	101.08	-213.22	-175.29

Note. AIC=Akaike Information Criterion, BIC = Bayesian Information Criterion. Green shaded ECV values indicate conditions in which the average relative bias for the general factor loadings estimated from the unidimensional model were within acceptable bias levels. Gray shaded cells indicate the lower AIC/BIC value.

[a]Conditions have unbalanced loadings across the positive and negative wording method factors (PWF, NWF); in these conditions, the positive items loaded 0.3 on the PWF and the negative items loaded 0.5 on the NWF.

[b]Difference in fit is the average bifactor model fit index – the average unidimensional model fit index

when ECV was 0.74, the difference in AIC was -395.34 and -344.75 in BIC. Similar patterns were seen within each PUC condition. Within each PUC condition, when balanced factor loadings were present across the method factors, more moderate ECV values (i.e., ECV of 0.66 and 0.72), as opposed to substantially smaller ECV values, were associated with the largest differences between AIC and BIC. Overall, at high ECV values, and when balanced loadings were included on both method factors, AIC and BIC differences tended to be smaller and approached the 10-point difference threshold.

Across PUC values, a similar pattern that was seen with CFI and TLI was also evident in AIC and BIC. By definition of PUC, the most "unidimensional structure" included in this study occurred with PUC was 0.73, while the least "unidimensional structure" occurred when PUC was 0.48. However, similar to the pattern seen with both CFI and TLI, the difference in AIC and BIC, overall, was largest when PUC was 0.54, narrowed when PUC was 0.48, and was closest when PUC was 0.73. This pattern was more pronounced for BIC.

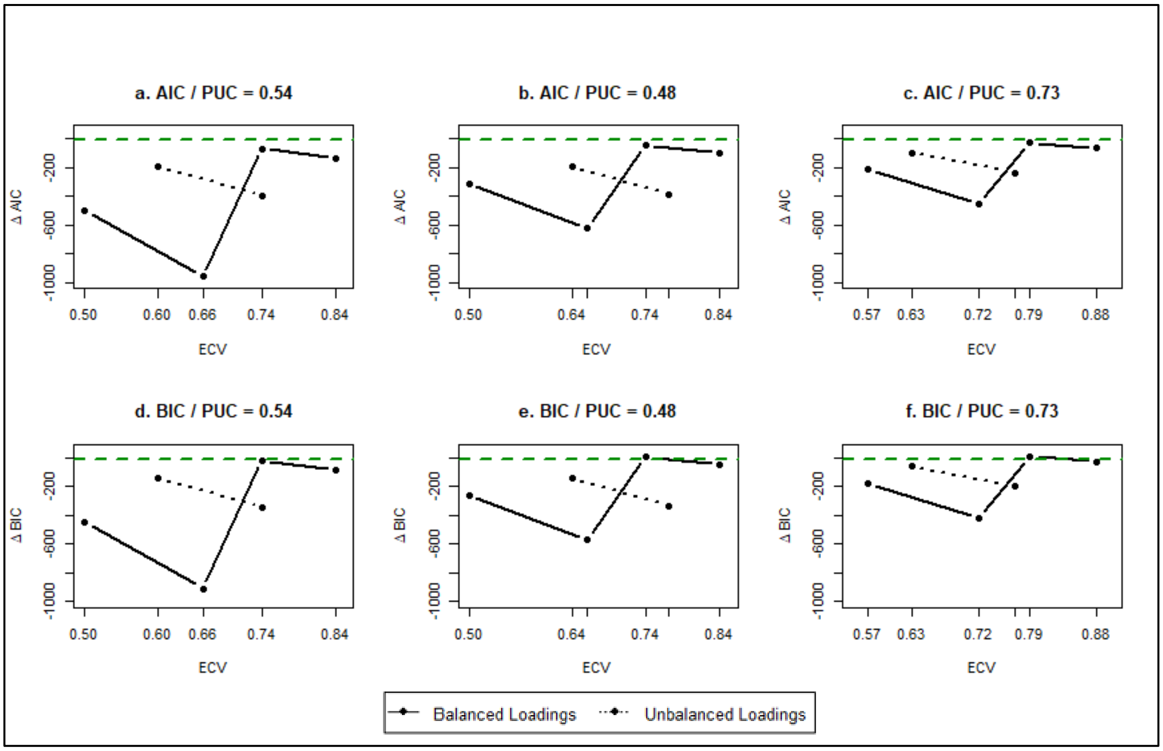

Figure 4.10 *Difference in Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) between the Bifactor Model and the Unidimensional Model by Explained Common Variance (ECV), Percent of Uncontaminated Correlations (PUC), and Loading Type for Normally Distributed Data.* Reference line placed at a difference of less than 10.

CHAPTER 5

DISCUSSION

The use of both positively and negatively worded items is a common design feature used in scales intended to be unidimensional. However, dimensionality investigations of these scales often result in a rejection of the theorized unidimensional model in favor of multidimensional structures. In recent years, the use of bifactor models to model item wording method effects has become increasingly popular (e.g., Alessandri et al., 2015; Gana et al., 2013; Hyland et al., 2014; Hystad & Johnsen, 2020; Marsh et al., 2010; McKay, Morgan, et al., 2015; Salerno et al., 2017). In the context of item wording method effects, a bifactor model contains a general factor, typically the construct of interest, that explains the covariation among all items contained in the scale, and method factors that capture additional covariation among items due to similar wording.

One of the benefits of the bifactor model is that it can be used to assess the strength and utility of a general factor in relation to specific factors. The comparison may be accomplished through the application of several model-derived indices including explained common variance (ECV), omega hierarchical (omegaH), and the percent of uncontaminated correlations (PUC). In the context of item wording method effects, ECV, omegaH, and PUC can be used to assess the strength of the general factor and provide evidence for essential unidimensionality, effectively evaluating the extent to which the method effects related to item wording are contaminating a scale's unidimensional property and structure.

Item wording method effects offer unique data conditions that have not been examined in the context of applying ECV, omegaH, and PUC to establish essentially unidimensional models. These unique conditions relate to the unbalanced nature of the method factors frequently found with item wording method effects. In this context, “unbalanced” refers both the number of items (i.e., positively to negatively worded items) that load on each method factor as well as the unbalanced nature of the factor loadings between the method factors, in which it is sometimes found that loadings on the negatively worded item method factor load more strongly relative to those on the positively worded method factor. Additionally, situations in which non-normal distributions associated with the negatively worded items can arise. These unique data conditions and their impact on utilizing the bifactor model-based indices ECV, omegaH, and PUC as indicators of essential unidimensionality were examined in this study.

Previous studies investigating the effectiveness of ECV, omegaH, and/or PUC in the context of establishing essentially unidimensional models have focused on one of two outcomes from utilizing a misspecified unidimensional model, bias in structural coefficients (Gu et al., 2017; Reise, Scheines, et al., 2013) and bias in factor loadings (Bonifay et al., 2015; Rodriguez et al., 2016a). Although these are different outcomes, as noted by Reise, Scheines, et al. (2013), bias in structural coefficients depends, to some degree, on bias in factor loadings. Issues with the structural model are typically traced back to the measurement model, which may include unmodeled sources of covariation among items or observed measures (Brown, 2015). This study focused on the recovery of the measurement model when method factors are ignored by examining bias in factor loadings from a misspecified unidimensional model, as this is where researchers reach

conclusions about the structure of their data for both subsequent analyses (i.e., fitting structural models) and scoring purposes (i.e., the utility of a composite score).

Technical Issues Related to the Bifactor Model

Before addressing the research questions investigated in this study, technical issues and concerns related to the bifactor model are addressed. Although the functioning of the bifactor model was not the primary focus, the results of this study exposed some issues and concerns related to this complex model. First, unexpected results occurred when item level distributions across the negatively worded items were non-normal. Second, model convergence was problematic for the bifactor model in the current study. These findings are informative for both applied researchers utilizing the bifactor model as well as methodologist examining the functioning of bifactor models.

Non-Normally Distributed Data

This study included two conditions to investigate the impact of normality on the effectiveness of using the bifactor model-based indices as indicators of essential unidimensionality. In one condition, all item-level distributions followed a normal distribution, with skewness and kurtosis of approximately zero. A second condition included item-level distributions that were severely non-normal, with skewness of approximately -3.0 and kurtosis of approximately 7.0. Unexpected results were obtained when data were non-normally distributed. Most concerning were the results obtained for the correctly specified bifactor model. Across most conditions, the general factor loadings estimated from the bifactor model were severely biased and underestimated. In this study, all conditions were estimated using MLMV (maximum likelihood with mean and variance adjusted). As this study included five-point ordered categorical data, it was

expected that MLMV would perform adequately in both normal and non-normal data conditions (Maydeu-Olivares, 2017; Rhemtulla et al., 2012). Maydeu-Olivares' (2017) study on the performance of MLMV included examination of a two-factor model, with extreme non-normality defined as approximate skewness of -2.00 and kurtosis of 3.18, slightly less extreme than what was included in this study. Additionally, the bifactor model represents a more complex measurement model compared to a two-factor model, as in a bifactor model, multidimensionality occurs at the item level (Bader et al., 2021; Rodriguez et al., 2016a).

Few studies have examined the impact of non-normality on the bifactor model. Xinya and Yanyun (2016) examined the impact of non-normality in bifactor models that included one general factor and one specific factor. Their study included item level non-normal distributions with skewness of approximately 2.00 and kurtosis of 7.00, similar to what was included in the current study. One outcome examined in Xinya and Yanyun's study included bias in factor loading estimates using robust maximum likelihood (MLR) and Bayesian estimation with non-informative priors. For a correctly specified bifactor model, Xinya and Yanyun reported both MLR and Bayesian estimation produced general factor loadings that were slightly underestimated, but within acceptable bias levels. In their study, non-normality was present across all items. The current study focused non-normality on the set of items specified as "negatively worded" items. Results from the current study indicated that, in general, more severe bias was associated with general factor loadings for the negatively worded items as estimated from the bifactor model. Isolating the non-normality to one set of items may have resulted in a structure that was too complex for the bifactor model. Others have noted the complexity of the bifactor

model can lead to problematic results (e.g., Maydeu-Olivares & Coffman, 2006). In addition to model convergence issues, also a noted problem in the current study, the bifactor model can lead to so called “anomalous results” which may include ill-defined specific factors and irregular factor loading patterns across both the general and specific factors (Eid et al., 2017).

Given the unexpected patterns in bias resulting from non-normality, a small subset of data, where PUC was 0.54, was re-examined using WLSMV estimation. Results from this analysis indicated proper recovery of general factor loadings from the bifactor model, with patterns similar to those estimated using MLMV. Estimation method appeared to be a confounding variable in this study. Therefore, results regarding how non-normality impacts the use of bifactor model-based indices are inconclusive. Given the potential role of improper estimation under extreme non-normality, accurate comparisons and conclusions cannot be made regarding the relationship between ECV, omegaH, PUC and essential unidimensionality in this context. Additional studies are needed to examine the role of model estimation in this context. Therefore, results in this discussion are evaluated only for the normally distributed data.

Model Convergence

Issues related to model convergence and the bifactor model have been noted in other studies (e.g., Green & Yang, 2018; Maydeu-Olivares & Coffman, 2006). Several distinct features of the bifactor model can impact model convergence, including: the magnitude of the general factor loadings, the number of indicators per specific factor, factor loading variability, and the presence of correlated specific factors (Bader et al., 2021; Green & Yang, 2018). Specially, Green and Yang (2018) have suggested that when

bifactor models follow a "bifactor pure-clustered model" or when all items load on both the general factor and their designated specific factor, and when uniform factor loadings are included across the general and specific factors, researcher should expect to encounter model convergence issues. Green and Yang have suggested that such instances can lead to empirical underidentification at the population level, which is then manifested as model nonconvergence at the sample level. In the current study, when the data were "bifactor pure-clustered models", convergence rates ranged from 57.9% to 68.7%, similar to rates reported by Green and Yang (i.e., model convergence less than 70%).

The convergence issues found in conditions characterized as "pure-clustered bifactor" led to slight instability in the bifactor model in terms of factor loading recovery. This aligns with a recent simulation study that found general factor loading estimates from the bifactor model can approach moderate bias levels when model convergence is poor (Bader et al., 2022). To address concerns with model convergence and the bifactor model, Green and Yang (2018) suggest including items that measure the general factor but not a specific factor. A similar condition was included in this study where PUC was 0.73. This condition was defined with only three of the six positive items loading on the positively worded item method factor and all six negative items loading on the negatively worded item method factor. This condition resulted in non-problematic convergence/proper solution rates. Additionally, Green and Yang (2018) recommend correlating the specific factors to improve model convergence. However, as detailed in Chapter 3, correlated specific factors were not included in this study, as latent correlations can lead to concerns with model interpretability and the application of the

bifactor model-derived indices ECV, omegaH, and PUC (i.e., Canivez, 2016; Bonifay et al., 2015; Gignac & Watkins, 2013; Reise, 2012).

Another condition included in the current study that may be considered problematic was the relatively moderate sample size of 500. However, Green and Yang (2018) reported model convergence rates associated with a "pure-cluster" bifactor model that contained uniform factor loadings across the general and specific factors of less than 70% and a sample size of 1,000. In a recent simulation study investigating sample size and bifactor models, Bader et al. (2022) concluded that a sample size of 500 was sufficient to produce non-problematic convergence rates and parameter estimates. Additionally, in their study investigating item wording effects, Gu et al. (2017) reported convergence/proper solution rates of near 100% for the bifactor model in conditions in which the sample size was 500. The distinguishing characteristics of both Bader et al. and Gu et al. were that the population bifactor models included in these studies either did not include uniform factor loadings (Bader et al., 2022), or included only one specific factor in which select items loaded (i.e., a negatively worded item method factor; Gu et al., 2017). Both conditions which, according to Green and Yang, should yield acceptable convergence rates for the bifactor model. Model convergence issues found in the current study appear to be directly related to the type of population bifactor model specified, and not the sample size of 500 used across all conditions.

Research Question 1: The Relationship between Factor Loading Bias, ECV, omegaH, and PUC in the Context of Item Wording Method Effects.

In evaluating the effectiveness of utilizing the bifactor model-based indices as indicators of essentially unidimensional models, the general consensus across studies

regarding the use of ECV, omegaH, and PUC is that these indices should be as high as possible in order to treat data as essentially unidimensional. However, what constitutes "high" is likely context dependent given the nature of the data examined. Reise, Scheines, et al. (2013) and Bonifay et al. (2015) offer the most comprehensive evaluation of these indices in the context of essential unidimensionality. From their examination of ECV, omegaH, and PUC, Reise, Scheines, et al. offered the following tentative benchmark for researchers, when PUC is greater than 0.80, the value of the strength indices becomes less important; however, when PUC is < 0.80, ECV >0.60, and omegaH >0.70, parameter bias should be minimal, and a unidimensional model can be used. In their study on the impact of factor loading bias from utilizing a misspecified unidimensional model, Bonifay et al. did not offer a specific benchmark, but noted that when PUC was low, or when data follow a more bifactor structure, the strength of the general factor (i.e., ECV) becomes critically important.

The conditions in this study effectively modeled different conceptualizations of item wording method effects, which were reflected in the varying PUC values included, all of which were lower than 0.80. When PUC was 0.54, each method factor included six items, representing a balanced condition in which both the positively worded and negatively worded item method factors were equal in size. When PUC was 0.48, eight items were specified to load on the positively worded item method factor, and four items on the negatively worded item method factor. In this condition, the positively worded item method factor is substantially larger than the negatively worded item method factor, representing a "weaker" negatively worded item method factor. When PUC was 0.73 only three of the six positive items were specified to on the positively worded item

method factor, and all six negative items loaded on the negatively worded item method factor, representing a “strong” negatively worded item method factor, and a “weaker” positively worded item method factor.

In this study, when the method factors were ignored by fitting a misspecified unidimensional measurement model, additional variance was left unexplained in the data. In general, when this occurs, additional covariation among the items that is not modeled will be “absorbed” into other parts of a misspecified model, resulting in biased parameter estimates (Morin et al., 2020). As expected, in this study, it was shown that treating multidimensional (i.e., bifactor) data as unidimensional by fitting the data to a unidimensional measurement model resulted in biased factor loading estimates (i.e., Rodriguez et al., 2016a). This follows the general notion that unaccounted method variance can mask itself with systematic trait variance of the measured general (or target) construct (Podsakoff et al., 2012). However, results from this study indicated that under certain conditions, data can be treated as essentially unidimensional, and that the degree and nature of the biasing effect differed across the various method factor structures modeled in this study. As such, the results of this study indicated that different patterns of ECV and omegaH values as indicators of essentially unidimensional models emerged as a result of the balanced-unbalanced structure in the data, or more accurately, the varying degrees of strength related to the method factors (i.e., PUC values).

In completely balanced conditions, in which both method factors were equal in size and strength (i.e., equal method factor loadings; PUC of 0.54), the bias in general factor loadings from fitting a unidimensional model was less severe across both item wording types (i.e., positively worded and negatively worded items). In this condition,

the method factors contributed equally to additional item-level variance beyond the general factor variance. Consequently, general factor loadings from a misspecified unidimensional model were overestimated to the same extent (i.e., Bonifay et al, 2015; Rodriguez et al., 2016a). When these method factors were ignored by fitting a misspecified unidimensional model, uniform bias was experienced across all 12 general factor loadings, resulting in less severe bias associated with more moderate ECV and omegaH values. When method factors were balanced both in size and magnitude, the benchmark set by Reise, Scheines, et al. (2013) generally held, in that ECV >0.60 and omegaH >0.70 values were associated with negligible bias in general factor loadings estimated from the unidimensional model. Broadly, this suggests that in balanced conditions, slightly lower values of ECV and omegaH can indicate acceptable bias in utilizing a unidimensional model. In other words, the general factor does not have to be as "strong" when method factors are completely balanced both in size and magnitude in order to maintain a unidimensional structure.

When unbalanced method factors were included, both in size and magnitude in loadings, general factor loadings as estimated from the misspecified unidimensional model across the positive and negative items experienced differential patterns in bias. More specifically, general factor loadings associated with the "stronger" method factor were overestimated, and general factor loadings associated with the "weaker" method factor were underestimated. The general factor loadings, as estimated from the unidimensional model, tended to be "pulled" towards the items that shared more covariation or those with stronger local dependence (Liu & Thompson, 2021; Reise et al, 2007). This holds important implications for the meaning of the "general factor" when

the method factors were ignored, and the data treated as unidimensional. The relationship between the bifactor model and the unidimensional model lies in the equivalency of the general factor between both models, as evidenced by the degree of consistency (or distortion) of the factor loadings between the two models (Reise et al., 2007; Reise, 2012). In this study, when the method factors were unbalanced (both in size and magnitude), the general factor loadings tended to be biased towards the “stronger” method factor. In other words, the general factor became more of a representation of the stronger method factor. Rodriguez et al. (2016a) describe this as a weighted composite defined by the size, and strength in the case of the current study, of the specific factors. For example, in this study, when the method factors were defined by a “strong” negatively worded item method factor, and the bias was severe, the general factor captured more variance associated with the negative items and was more conceptualized as a “negative item construct” rather than the intended general construct.

As a result of the differential bias experienced in unbalanced conditions, the general factor needed to be substantially stronger, as assessed by the associated ECV and omegaH values derived from the bifactor model, to compensate for the over/underestimation of general factor loadings across both the positive and negative items in these conditions. When the negatively worded item method factor was strong both in size and strength (i.e., at PUC conditions 0.54 and 0.73 with unbalanced method factor loadings), the general factor loadings on the negative items, as estimated from the unidimensional model, did not reach acceptable bias levels at the highest ECV and omegaH values associated with these conditions (ECV of 0.77; omegaH of 0.83). This indicates that when one method factor is substantially “stronger” both in size and

magnitude, ECV, as derived from the bifactor model, likely needs to surpass 0.80, and omegaH in excess of 0.85, to indicate the data can be treated as essentially unidimensional.

An interesting pattern developed when the method factors were defined by a more sizeable positively worded item method factor (i.e., PUC of 0.48) but "stronger" negatively worded item method factor in terms of the magnitude in loadings (i.e., loadings of 0.3 on the positively worded item method factor and 0.5 on the negatively worded item method factor). In other words, the size and strength of the method factors were counterbalanced. This counterbalancing effect appears to have helped the unidimensional model better recover the general factor loadings compared to other unbalanced conditions. In this condition, there was less severe bias in the factor loadings for both the positive and negative items when estimated from the unidimensional model, and both item types reached acceptable levels of bias when ECV was 0.77 and omegaH was 0.83.

Compared to ECV, there was a more inconsistent relationship between the values of omegaH and subsequent factor loading bias when the data were fit to the unidimensional model. As noted by Reise, Scheines, et al. (2013) this is because omegaH is more effected by the magnitude of the general factor loadings, while ECV is more related to the strength of the general factor in relation to the specific factors, or method factors in the current study. This inconsistent relationship was evident when examining the relative bias in estimated factor loadings when PUC was 0.48 and 0.73 and the method factor loading patterns were balanced. For example, when PUC was 0.48, an omegaH value of 0.70 was associated with bias in estimated factor loadings from the

unidimensional model that was right at the acceptable threshold, while an omegaH value of 0.75 exceeded the acceptable bias threshold (i.e., greater than |10%|). Examination of each condition's associated factor loading pattern and ECV value provides further details. The omegaH value of 0.70 is associated with a factor loading pattern of 0.5, 0.3, 0.3 and an ECV of 0.74. In this condition, the general factor is of moderate magnitude but is stronger is relation to the relatively weak method factors. The omegaH value of 0.75 is associated with a factor loading pattern of 0.7, 0.5, 0.5 and an ECV of 0.66. In this condition, the general factor's loadings are strong, but the method factor loadings are more moderate. This results in a general factor that is not as strong relative to the method factors and therefore a lower ECV value of 0.66. Given the more inconsistent relationship between omegaH and parameter bias when data are fit to a misspecified unidimensional model, it is important for researchers to consider both ECV and omegaH together, and not in isolation, when reaching decisions regarding essential unidimensionality, particularly if the method factors are unbalanced.

A number of findings and recommendations have been proposed as to which values of ECV, omegaH, and/or PUC serve as general guidelines for treating bifactor data as essentially unidimensional. In addition to the benchmark proposed by Reise, Scheines, et al. (2013; PUC < 0.80, ECV >0.60, and omegaH >0.70), Rodriguez et al. (2016a) offered ECV > 0.70 to 0.80, and ECV >0.70 and PUC >0.70 as indicators that bias in factor loadings from a unidimensional model should be minimal, and the data can be treated as essentially unidimensional. Findings from this study suggest there are marked differences in potential thresholds for balanced and unbalanced conditions. In completely balanced conditions, with positively worded and negatively worded item

method factors that are approximately equivalent both in size and strength, more moderate ECV and omegaH values were indicative of essentially unidimensional models, with Reise, Scheines, et al.'s benchmark generally holding true. When the method factors were unbalanced, higher ECV and omegaH values were needed as indicators of essentially unidimensional models, and slightly more robust than what was recommended by Rodriguez et al. A general recommendation based on the results of this study are ECV > 0.80 and omegaH > 0.85 should indicate the use of the unidimensional model is acceptable when the method factors are unbalanced. Furthermore, as detailed in the next section, the actual value of PUC may not be informative as to the degree of expected bias from utilizing a unidimensional model when the structure is unbalanced.

Research Question 2: The Relationship between ECV and PUC in the context of Item Wording Method Effects

Both Reise, Scheines, et al. (2013) and Bonifay et al. (2015) noted the importance of the relationship between ECV and data structure (i.e., PUC) in determining when data can be treated as essentially unidimensional. In essence, as PUC increased, both Reise, Scheines, et al., and Bonifay et al., found that lower ECV values were associated with minimal parameter bias from fitting a unidimensional model to data that follows a bifactor structure. Applying this concept to the current study, when PUC was 0.73, lower ECV values should have indicated acceptable bias levels in factor loadings from utilizing the unidimensional model. However, this was not found. In fact, in this condition, the minimum ECV value associated with negligible bias from utilizing the unidimensional model was 0.79. Additionally, overall, the most severe bias in estimated general factor loadings from the unidimensional model occurred within the PUC of 0.73 condition. As

noted, this condition contained a “strong” negatively worded item method factor. This leads to questions about the utility of PUC as an indicator of essential unidimensionality when data follow an unbalanced bifactor structure.

The influence of unbalanced specific factors is emerging in research investigating the use of bifactor models in the context of establishing essential unidimensionality. A recent study from Liu et al. (2022) found that PUC was non-informative as an indicator of essential unidimensionality when unbalanced specific factors were present. “Unbalanced” in Liu et al.’s study referred only to the size of the specific factors (i.e., the number of items specified to load on each specific factor) and not unbalanced in terms of the magnitude of the loadings. Liu et al. found substantial weaker and non-linear relationships between PUC and subsequent parameter bias (structural coefficients in this study) when data that followed a bifactor structure were subsequently fit to a unidimensional model. They noted this was due to the fact that the PUC value from an unbalanced structure depends not just on the overall number of specific factors and items included, but how the items are distributed within the specific factors. In other words, there is a difference in size between specific factors in unbalanced structures. In unbalanced structures, Liu et al. noted that ECV and omegaH were more useful as these indices are direct indicators of general factor strength and saturation in relation to a scale’s items. Additionally, Liu et al. found that the use of omega subscale (ω_s) was a useful indicator in the context of unbalanced specific factors. Omega subscale is a measure of the total score variance attributable to the group or specific factors (Liu et al., 2022; Rodriguez et al., 2016a, 2016b). Again, this finding suggests that the strength of the unbalanced specific factors (or method factors in the current study) is a key indicator

of subsequent parameter bias. This "strength" is not necessarily indexed by the specific PUC value.

Both Bonifay et al. (2015) and Reise, Scheines, et al. (2013) noted that a data's structure is key to establishing the relationship between ECV and/or omegaH as indicators of essential unidimensionality. "Data structure" was indexed by PUC in their studies, and given the completely balanced conditions, they were able to establish a more linear connection between increased PUC and subsequent parameter bias from the unidimensional model. Subsequently, both studies indicated that PUC moderates the relationship between ECV and parameter bias from utilizing a unidimensional model. This "moderating effect" between the data's structure and the value of ECV as an indicator of essential unidimensionality was found in the current study, however it was not directly tied to a specific PUC value. Rather, the strength of the method factors and how this related to the data's structure was indicative of subsequent bias in utilizing the unidimensional model. More specifically, the imbalance in strength between the method factors. In this study, conditions in which the negatively worded item method factor was substantially stronger, both in size and magnitude, such as when PUC was 0.73 or when PUC was 0.54 and unbalanced method factor loadings were present, more severe bias was associated with ECV values that are considered strong and indicative of a general factor that explains a substantial portion of the variance across all items (i.e., approximately 0.72 to 0.77). Similar to the findings from Liu et al. (2022), in the current study, the values of ECV, as well as omegaH, as derived from the bifactor model were more informative as to the degree of subsequent parameter bias in the unidimensional model than PUC when method factors were unbalanced both in size and magnitude.

However, the values of ECV and omegaH need to be considered in relation to the structure of the method factors.

The non-informative nature of PUC in unbalanced structures also relates to findings from Gu et al. (2017) in their investigation of item wording method effects. In their study, Gu et al. specified only one method factor associated with negatively worded items, creating an unbalanced structure. The data's structure or PUC was not considered as a factor in their study. However, considering the total number of items and items specified to load on the negatively worded item method factor, PUC values included in Gu et al.'s study were 0.76, 0.77, 0.90, and 0.91. However, even within the very high PUC values of 0.90, Gu et al. reported severe bias in structural coefficients at ECV values of 0.75. Although not discussed in their study, the unbalanced nature of the method factors likely created the same effect as was found in this study. The additional covariation among the negatively worded items due to the "strong" method factor associated with these items resulted in more severe bias, even when the general factor was considered relatively strong (i.e., high ECV values). In essence, when unbalanced structures are present, the structure of the data is important, but the actual value of PUC appears not to be informative.

Research Question 3: The Relationship between ECV, PUC, and Model-Fit Indices

The bifactor model forms a nested structure with the unidimensional model, correlated-factors model, and the higher-order model, with the bifactor model being the most complex of these models and the unidimensional the most parsimonious (Reise, 2012; Reise et al., 2010). Gignac (2016) reported that examining practical change in fit indices that reward model parsimony (e.g., TLI, AIC, and BIC) was useful for examining

the degree to which a bifactor model showed better fit compared to a higher-order model. More specifically, Gignac found the degree to which the bifactor model showed improved fit was directly related to the magnitude of violating the proportionality constraint imposed by the higher-order model. In this study, the degree of misspecification between the bifactor model and unidimensional model is a function of ECV (Reise, Scheines, et al., 2013). As such, this study examined the difference in the degree of fit between the bifactor model and the unidimensional model as a function of both ECV and PUC as these are considered statistics of unidimensionality.

This study examined the relationship between ECV, PUC and the model fit indices that include penalties for model complexity (i.e., reward model parsimony), including CFI, TLI, AIC, and BIC. As expected, results from this study indicated exceptional and near perfect fit for the bifactor model with average CFI and TLI values of approximately 1.00, and differences in AIC and BIC that favored the bifactor model in most conditions. In general, the model fit indices examined in this study performed similarly, in that, at very high ECV values, the difference in fit between the models was close, but in almost all conditions, indicated practical improvement in fit for the bifactor model. In other words, as the general factor became more "unidimensional" by accounting for more variance across all items, the degree of difference in fit between the models became smaller.

The finding that the difference in fit was slightly closer for CFI compared to TLI is consistent with Reise, Scheines, et al.'s (2013) finding that CFI and ECV are closely related (i.e., correlating 0.80 in their study), prompting them to label CFI a promising "unidimensional enough" fit index. Although Reise, Scheines, et al. did not directly

examine CFI as a function of ECV, they found CFI was not as strong a predictor in resulting parameter bias (i.e., structural coefficient bias) when bifactor data were fit to a unidimensional model compared to ECV. Similar findings were evident in this study. While both CFI and TLI reached acceptable levels (i.e., ≥ 0.95) across several conditions that were also associated with high ECV values that indicated acceptable bias levels in treating the data as unidimensional, several ECV values that indicated acceptable bias from utilizing the unidimensional model did not reach acceptable CFI and TLI values. Similar to Reise, Scheines, et al., this is an indication that the model fit indices CFI and TLI are perhaps unreliable as indicators of subsequent parameter bias and "essential unidimensionality."

This study included the first examination of AIC and BIC in relation to the bifactor model-based indices ECV and PUC. Slightly more unexpected patterns were seen in the relationship between AIC and BIC with regards to ECV compared to either CFI or TLI. Similar to CFI and TLI, when balanced loadings were present across the method factors, smaller differences were found between the models at the highest ECV values (i.e., as the data became more unidimensional). However, the largest differences in AIC and BIC occurred at more moderate ECV values (0.66 and 0.72) as opposed to the lowest ECV values (0.50 and 0.57). Additionally, when unbalanced factor loadings were included across the method factors, differences in AIC and BIC were smaller at lower ECV values (e.g., 0.60 – 0.64) compared to higher ECV values (e.g., 0.74 – 0.79). In general, this suggests that AIC and BIC are perhaps not as linearly related to ECV as both CFI and TLI. Additionally, the moderate sample size included in this study may have impacted both AIC and BIC. Gignac (2016) reported that AIC and BIC were less likely to

distinguish between the higher-order model and the bifactor model at low to moderate levels of the proportionality constraint violation when the sample size was 500. Ten Berge and Socan (2004) described ECV as a coefficient that represents “closeness to unidimensionality.” Applying the finding from Gignac to the current study, given the more moderate sample size of 500 used in this study, AIC and BIC may have had a harder time recognizing “closeness to unidimensionality” except at very high ECV values. In other words, the more non-linear pattern seen between AIC, BIC, and ECV may be related to the more moderate sample size used in this study.

By definition of PUC (i.e., larger PUC values are indicative of a more unidimensional data structure), it was expected that the overall difference in the model fit indices examined would be inversely related to the PUC value (i.e., largest when PUC was 0.48, smaller when PUC was 0.54, and smallest when PUC was 0.73). However, the difference in fit values between the bifactor model and unidimensional model was largest when PUC was 0.54, narrowed when PUC was 0.48, and was closest when PUC was 0.73. This pattern was consistent across all fit indices examined, CFI, TLI, AIC, and BIC. One reason for this may be related to the elements of the model-implied variance-covariance matrix when unbalanced method factors were present. When PUC was 0.48, the method factors consisted of eight positively worded items and four negatively worded items. The pattern in general factor loading bias as estimated from the unidimensional model indicated an over/underestimation of the factor loadings associated with the positive items and negative items due to the unbalanced nature. This bias pattern is also likely extending to the model implied variance-covariance matrix (Liu & Thompson, 2021). For example, when the bifactor data were fit to the unidimensional model, the

covariances (or correlations in a completely standardized solution) between items as specified in the model-implied variance-covariance matrix ($\hat{\Sigma}$) are the product of the item loadings for each item on the general factor (i.e., $\hat{\sigma}_{x1,x9} = \lambda_{1G}\lambda_{9G}$). The correlations between items from different method factors will appear less biased because of the overestimation of the loadings on the positively worded items and the underestimation of the loadings on the negatively worded items; said another way, the opposite levels of bias "cancel" each other out when estimating the variance-covariance value from the model parameters. In this situation, the model-implied matrix may appear to more closely resemble the true population variance-covariance matrix and the true correlation between the items. Therefore, model fit may appear "closer" in unbalanced conditions because of the over/underestimation of factor loadings on items related to different wording types. In other words, it is possible that model fit may be slightly inflated when unbalanced structures are present.

This finding also suggests, that while the PUC of 0.73 condition is considered the most "unidimensional structure" in this study by definition of PUC, the resulting "narrowing" pattern seen when examining the differences in CFI, TLI, AIC, and BIC in this condition is likely related both to the fact that fewer item correlations are contaminated by both general and method factor variance (i.e., the definition of PUC) and the resulting bias pattern in factor loadings estimated from the unidimensional model, which indicated more severe overestimation for the negative item loadings and underestimation for the positive item loadings. The relationship between the over/under estimation in factor loadings in unbalanced structures may also be related to the findings regarding model-data fit in Gu et al. (2017). The population model used in Gu et al.'s

investigation of item wording method effects consisted of one method factor related only to the negatively worded items, a highly unbalanced structure, that also generated high PUC values. Of note, model-fit in Gu et al. refers to the fit of the structural model. Gu et al. reported a more inconsistent relationship between obtained CFI values and ECV. For example, in one condition in which ECV was 0.89, Gu et al. reported an associated CFI value of 0.983. However, in another condition in which ECV was 0.33, Gu et al. reported an associated CFI value of 0.965. Gu et al. cautioned readers that the use of model fit indices, such as CFI, in selecting the unidimensional model could lead to severe bias. One potential explanation for such robust CFI values is the relationship between unbalanced structures, the over/underestimation of factor loadings when estimated from the unidimensional model, and its impact on the model-implied variance-covariance matrix. Additionally, Liu et al. (2022) reported a slightly stronger relationship between CFI and subsequent structural coefficient bias when data were fit to a unidimensional model in unbalanced structures (r = 0.51) compared to balanced structures (r = 0.44).

Overall, the findings of this study as it relates to model-data fit indices, confirm Reise, Scheines, et al.'s (2013), as well as Gu et al.'s (2017), findings that the use of conventional CFA model-data fit indices are not reliable predictors of subsequent parameter bias from utilizing a unidimensional model. Similar to Reise, Scheines, et al., the findings from this study suggest that the use of the bifactor model-based indices, ECV and omegaH, are more useful and informative as indicators of resulting bias in utilizing a unidimensional model when data follow a bifactor structure, compared to conventional model fit indices. Furthermore, the current study's findings related to the relationship

between model-data fit indices and PUC provides further evidence that PUC may not be a useful index when unbalanced bifactor structures are present.

Implications for Practice

Ignoring Method Factors

When a bifactor model is found to be the “best fitting” model in the context of item wording method effects, researchers may draw several conclusions. First there may be an assumption that the finding of multidimensionality implies the construct itself is multidimensional, with the added assumption that the identified method factors are, in fact, substantive in nature and related to the general construct. Second, this may prompt researchers to assume that the use of subscale scores either in place of, or in addition to, an overall composite score is needed. Third, researchers may assume that an observed sum score is no longer viable and the bifactor model is needed as a “control” for the method effects in order to generate an “uncontaminated” score. All of these assumptions should be thoroughly evaluated by the researcher prior to selecting the bifactor model as the representation of their data in the context of item wording method effects.

Furthermore, the application of the bifactor model-based indices can provide clarity. Results from the current study indicated the use of both ECV and omegaH were informative as to the degree of subsequent factor loading bias when data, simulated to follow a bifactor structure with two method factors, were fit to a unidimensional model (i.e., the intended structure ignoring the method factors). Previous studies indicated that the use of these indices should be considered in relation to the data’s bifactor structure (Bonifay et al., 2015; Reise, Scheines, et al., 2013; Rodriguez et al., 2016a). In these studies, the data structure was indexed by the PUC value. However, as shown in the

current study, the use of PUC as an informative index related to subsequent parameter bias from utilizing a unidimensional model did not hold when unbalanced method factors were present. However, consideration of the data's structure is still critical in the application of both ECV and omegaH in the context of item wording method effects. Researchers need to consider the balanced vs. unbalanced nature of the method factors, as higher ECV and omegaH values are needed to indicate when data can be treated as essentially unidimensional when unbalanced method factors are present. "Unbalanced" in this context refers to both the size (i.e., the number of items that load on both the positively worded method factor and negatively worded method factor) and the magnitude of the loadings between the method factors.

In some instances, the bifactor model may represent an overly complex model in in the context of item wording method effects, particularly if the general factor is strong (i.e., Rodriguez et al, 2016a). As shown in this study, the bifactor model, and the use of the model-derived indices, can provide a clear direction for researchers to assess the unidimensional nature of their scale, or the extent to which the method factors are contaminating the scale. One concern with method factors in the context of item wording method effects are their potential stability across both scales and populations sampled. Bonifay et al. (2015) noted that when small specific factors, or perhaps nuisance dimensions emerge, "…the apparent structure of the secondary dimensions is less likely to be replicable across studies" (p. 515).

Relatedly, one finding across studies examining item wording method effects is that the nature of the effect is likely scale specific, indicating that the meaning of this effect, and its severity, is different depending on the scale used (Kam, 2018). Looking

within scales, this can also be extended to note that item wording method effects may be sample specific. The Rosenberg Self-Esteem Scale (RSES) is one of the most researched scales for investigating item wording method effects. Across studies, researchers have shown that the method effects associated with the positively and negatively worded items can take different forms and magnitudes depending on the population studied. For example, method effects within the RSES have been identified across both the positive and negative items (e.g., Alessandri et al., 2015; Hyland et al., 2014; Marsh et al., 2010; Michaelides et al., 2016; Salerno et al., 2017), primarily associated with the negative items (e.g., DiStefano & Motl, 2006; Gnambs & Schroeders, 2020), or primarily associated with the positive items (e.g., Wang et al., 2001). Additionally, the method effects associated with the RSES have been characterized as response styles associated with certain personality traits in college-aged students (DiStefano & Motl, 2006; Quilty et al., 2006), found to be related to cognitive ability in school-aged children (Gnambs & Schroeders, 2020; Marsh et al., 2010), and related to life satisfaction and depression in older adults (Lindwall et al., 2012). Furthermore, some have suggested that while method effects within the RSES associated with the positively and/or negatively worded items exist, they are not severe enough to impede treating the data as unidimensional (e.g., Donnellan et al., 2016; Hyland et al., 2014; McKay et al., 2014; Reise et al., 2016). Meanwhile, others have questioned the use of a unidimensional model with the RSES (e.g., Alessandri et al., 2015; Gana et al., 2013; Marsh et al., 2010). These differential findings into the nature and magnitude of item wording method effects suggests that method effects related to item wording are perhaps not stable, therefore, applied

researchers need to fully evaluate the nature and degree to which the unidimensional structure has been contaminated by the method effects.

The results of this study indicated that use of the bifactor model-based indices, specifically ECV and omega hierarchical, provides an effective and objective tool for evaluating the degree to which unidimensionality may be violated due to item wording method effects. Across the research investigating the application of the bifactor model-based indices, ECV, omegaH, and PUC, there is somewhat of a reluctance to provide definitive thresholds for exact levels that indicate essential unidimensionality. One reason for this is the influence that a data's structure has on the utility of both ECV and omegaH. Broadly within bifactor models, in the case of unbalanced structures, there exist many patterns that may develop both in size and magnitude of the specific factors that can alter the effectiveness of both ECV and omegaH as indicators of essentially unidimensionality (Liu et al., 2022). The context of item wording method effects offers a slightly more finite structure in that the method factors are defined by the item wording (i.e., no more than two method factors will be specified). The following recommendations and guidelines are provided for researchers utilizing unidimensional scales containing both positive and negative items:

- Researchers should fit both a bifactor model and unidimensional model to the data to examine the nature of method effects and the presence of method factors.
- As model fit will likely indicate superior fit to the bifactor model, researchers should carefully examine the factor loading patterns between the general factor and the method factor(s), as well as between the general factor in the bifactor model and unidimensional model.

- Researchers should utilize the bifactor model-based indices for a comprehensive psychometric evaluation. At minimum, these indices should include ECV and omega hierarchical (omegaH). This will assist researchers in evaluating the strength of the general factor in relation to the identified method factor(s), and to assess the degree to which the data are unidimensional.
- As omegaH is more impacted by the magnitude of the factor loadings, ECV and omegaH should be considered together, and in relation to the data's bifactor structure (i.e., balanced vs. unbalanced method factors), in reaching decisions regarding the unidimensionality of the data.
- If the method factors contain balanced items and approximately balanced loadings across both method factors (i.e., a completely balanced condition), researchers can utilize the benchmark set by Reise, Scheines, et al. (2013), in that ECV >0.60 and OmegaH >0.70 are likely to indicate the data can be treated as essentially unidimensional.
- If unbalanced structures are identified, both in size and magnitude, higher values of ECV and omegaH are desirable. Researchers can use the minimum threshold of ECV >0.80 and OmegaH >0.85 as a tentative benchmark.
- If the data structure, ECV, and omegaH meet the suggested guidelines, researchers can proceed with treating the data as unidimensional by utilizing a unidimensional measurement model in subsequent SEM analyses and/or by computing a composite sum score.
- If non-normality is found, researchers should consider utilizing a categorical estimator such as WLSMV in place of ML-based estimation.

Clearly, if researchers are interested in the functioning of the method factors, they should continue to model the method factors (i.e., the bifactor model) in subsequent analyses. However, if the method factors are found to "contaminate" the unidimensional structure, and a composite score is desired, researchers may conclude that the use of a latent scoring technique, such as estimating factor scores, is needed. In their study on the impact of item wording method effects, Gu et al. (2017) recommended to researchers that if both ECV and omega hierarchical were "large enough", researchers could utilize a summed composite or total score. If not "large enough," they recommended researchers proceed with latent variable modeling. In their study, ECV >0.75 was considered as an acceptable threshold. Gu et al. did not offer a recommendation for omegaH. The current study identified omegaH of at least 0.70 in balanced conditions and at least 0.85 in unbalanced conditions as "large enough" to proceed with utilizing a summed composite score. However, researchers should proceed with caution before utilizing a latent scoring technique. More specifically, they should consider what is being captured by the method factors. For example, Reise et al., (2016), in an investigation of the RSES, noted that method factors, and the selection of a bifactor model, resulted primarily from invalid or random response patterns. In this case, Reise et al. cautioned against the use of factor scores noting that "invalid" responses cannot be corrected or adjusted through latent variable modeling. Ultimately, with the application of a unidimensional scale, researchers are interested in capturing the construct measured by the scale. This construct is either measured by the scale or it is not, and if not, researchers should evaluate why. In other words, if the unidimensional scale is severely contaminated by method factors due to item wording, researchers should evaluate the quality of their data before attempting to

score individuals via latent modeling. However, if a latent or factor score is used in place of an observed sum score, researchers should also consider utilizing the *H* index (i.e., construct replicability) and factor determinacy (*FD*) to fully assess reliability.

The Inclusion of Negatively Worded Items

The primary reasons for including both positively and negatively worded items in scales designed to be unidimensional are that they are thought to control for the effects of certain response styles (e.g., acquiescence; Baumgartner & Steenkamp, 2001; Cloud & Vaughan, 1970), prevent respondents from engaging in response sets (e.g., Barnette, 2000; Podsakoff et al., 2003), and in some instances, improve scale validity by providing more complete coverage of the construct (Weijters & Baumgartner, 2012). Others have questioned the use of including negatively worded items, noting that it is a "counterproductive" practice as additional response biases can occur (van Sonderen et al., 2013). Furthermore, some have advocated for the use of balanced scales (equal positive to negative items) to maximize the utility of including both item types (e.g., Baumgartner & Steenkamp, 2001; Marsh, 1996).

Results of this study indicated both advantages and disadvantages of a balanced scale. On the one hand, in this study, when the scale was balanced and method factors were also balanced, both in size and magnitude, less severe and more uniform bias in estimated factor loadings were found with both item types when the data were treated as unidimensional. However, when the magnitude of the method factor in a balanced scale became more concentrated on one method factor (i.e., a strong negatively worded item method factor), bias was severe, and the ability to treat the data as essentially unidimensional became increasingly more difficult.

As studies investigating item wording method effects associated with the Rosenberg Self-Esteem Scale have shown, it is difficult to predict the nature of potential method effects related to item wording, as it can take different forms and magnitudes depending on the sample investigated. A promising design feature researchers may consider is to include fewer negatively worded items, particularly if a method effect related to the negatively worded items is expected. In this study, when the scale was unbalanced and contained eight positively worded items to four negatively worded items (i.e., PUC of 0.48), and the loadings on the negatively worded item method factor were larger in magnitude, the biasing effect of treating the data as unidimensional was less severe. In this condition, the estimated factor loadings from the unidimensional model were less severe across both item types. Although the general factor needed to be of substantial strength in this condition (i.e., ECV of 0.77, omegaH of 0.83), the degree of bias between the two item wording types was mitigated due to the counterbalancing effect of method factors that differed in size and strength, leading to slightly better equivalency between the general factors estimated from the bifactor model and the unidimensional model. Similarly, Gu et al. (2017) also recommended researchers carefully consider reducing the number of negatively worded items included in scales.

The Use of Bifactor Models

As noted, the results of this study highlighted some of the concerns related to the complexity of the bifactor model, notably issues related to model convergence and when the data was non-normally distributed. These issues resulted in slight instability in the bifactor model in terms of factor loading recovery related to convergence problems and severe underestimation of the general factor loadings as estimated from the bifactor

model (i.e., the true population model) across most non-normally distributed data conditions. However, despite these problematic results, the bifactor model showed exceptional and near perfect fit across conditions (e.g., see Table B.1 in Appendix B). This aligns with other studies indicating good model-data fit for the bifactor model and arguments over the selection of the model based on traditional model fit indices (e.g., Bonifay & Cai, 2017; Greene et al., 2019; Morgan et al., 2015; Murray & Johnson, 2013; Reise et al., 2016). Because of these concerns, it is essential for researchers to fully evaluate the bifactor model through the application of the model-based indices.

This study also exposed the benefits and power of a bifactor model in its usefulness for providing a comprehensive psychometric evaluation of assessments and scales (i.e., Bonifay et al., 2017; Rodriguez et al., 2016a, 2016b). Similar to related studies, this study showed that the bifactor model is a useful and effective model for assisting researchers in evaluating the degree of unidimensionality in their scale (Bonifay et al., 2015; Liu et al., 2022; Reise, Scheines, et al., 2013; Rodriguez et al., 2016a, 2016b). Given the concerns related to the complexity of bifactor models, including convergence issues, it may be beneficial for researchers to consider less complex models for subsequent analyses (i.e., Bonifay et al., 2015). Additionally, the use of composite scores is typically desired when constructs are measured via unidimensional scales. The bifactor model, with its specification of a general construct that explains the covariation among all items, and specific factors (or method factors in this study) that explain additional covariation among sets of items can provide a comprehensive evaluation of the "strength" of the general factor and the utility of a composite score through the application and evaluation of ECV and omega hierarchical (Reise, Bonifay, et al., 2013).

Limitations and Future Research

As with any simulation study, generalizability of the results is directly related to the conditions included. This study included two unbalanced structures related to the size of the method factors (i.e., positive to negative item ratios of 8:4 and 3:6) and two unbalanced structures related to the magnitude of the loadings between the method factors (i.e., loadings of 0.3 on the positively worded method factor and 0.5 on the negatively worded item method factor). Results indicated that the nature of the unbalanced structure is a critical element related to the resulting bias in factor loadings estimated from the unidimensional model. However, different results might have been obtained if, for example, the unbalanced item condition contained a positive to negative item ratio of 9:3 or 7:5. This study, along with the recent contributions of Liu et al. (2022), found that unbalanced specific (method) factor structures can significantly change the minimum ECV and omegaH values associated with identifying essentially unidimensional models. Future research is needed to investigate the discrepancy in the magnitude of the loadings between method factors. For example, are there marked differences found when the method factor loading magnitude differs by 0.5 and 0.6? Or 0.3 and 0.6? Varying loading values in future studies can help to answer this question.

The choice of loading value for the method (specific) factors also relates to the potential role of the omega indices for subscales. As noted by Liu et al. (2022), omega subscale (i.e., the "strength" of the specific factors) was a strong predictor of subsequent parameter bias in unbalanced structures in their study. Omega hierarchical subscale (omegaHS) is also available as a recommended bifactor model-based index. OmegaHS is an estimate of the reliability of a subscale score unique to each specific factor (Rodriguez

et al., 2016a). Gignac and Kretzschmar (2017) proposed the following effect size guidelines related to omegaHS values: < 0.20 are small, 0.20 to 0.30 are typical, and >0.30 are large. How these values and the use of omegaHS in relation to unbalanced structures should be a focus in future research investigating the utility of bifactor model-based indices in the context of establishing essentially unidimensional models.

Results from this study, in addition to the findings from Liu et al. (2022), indicated that PUC may be non-informative as an indicator of essential unidimensionality when unbalanced specific (method) factors are present. The current study included what are considered "low" PUC values of 0.48, 0.54, and 0.73. Although the inclusion of only three PUC values is somewhat limited, each of the values included is representative of PUC values that are typically associated with a bifactor model in the context of item wording method effects. The only model modification that would generate "high" PUC values in this context is to specify only one method factor, thereby creating a highly unbalanced structure. This was the case in Gu et al. (2017) where only a negatively worded item method factor was included in the population bifactor model, which generated PUC values as high as 0.91. More recent studies have referred to this model as the bifactor-(S-1) model (e.g., Eid et al., 2017; Gnambs & Schroeders, 2020). One reason for the use of this bifactor model specification is that when the data were fit to a traditional bifactor model (i.e., one general and two method factors), researchers have identified what Eid et al. (2017) refer to as anomalous results, such as low loadings, non-significant loadings, or irregular factor loading patterns on one of the method factors. However, in the context of modeling item wording method effects, some have questioned the use of this model as it may overestimate the method effect associated with negatively

worded items (Maydeu-Olivares & Coffman, 2006; Marsh et al., 2010). Additionally, in the bifactor-(S-1) model, the unmodeled method factor serves as a reference method to the modeled method factor (Geiser et al., 2008). Therefore, the general factor takes a slightly different interpretation as the common factor associated with the reference method (Geiser et al., 2008). Given both the different interpretation of the general factor, as well as the highly unbalanced bifactor structure created by this model, more research is needed investigating the application of the bifactor model-based indices in this context.

The population model used in this study followed the specifications of a restricted bifactor model in that all items loaded on the general factor and only one method factor (i.e., no cross-loadings). Additionally, all latent variables were orthogonal. While orthogonality is required between general and specific factors (or method factor in this study), it is only recommended between the specific factors in a bifactor model. The primary reason for not considering correlated method factors in this study related to the fact that interpretation of the bifactor model-based indices (ECV, omegaH, and PUC) require these factors to be orthogonal (Canivez, 2016; Bonifay et al., 2015; Gignac & Watkins, 2013). However, several studies utilizing bifactor models in the context of item wording method effects have modeled correlations between the positively and negatively worded item method factors (e.g., Donnellan et al., 2016; Lindwall et al., 2012; Quilty et al., 2006; Rodrigo et al., 2019). Additionally, correlating specific factors is one recommendation for improving model convergence in bifactor pure-clustered models (Green & Yang, 2018). Reise (2012) has cautioned against correlating specific factors noting that this suggests unmodeled general factors are present. However, in the context of item wording method effects, the method factors typically represent nuisance factors

that are unrelated to the general construct. In this sense, it may be reasonable to consider the factors correlated in certain circumstances (Morin et al., 2020). How the inclusion of correlated method factors impacts the estimation of ECV and omegaH should be the focus of future studies. In general, few studies have examined how different data conditions can impact the estimation of ECV and omegaH. Murray et al. (2019) reported that the estimation of ECV and omegaH was not robust when cross-loadings were ignored and estimated via a bifactor CFA model. Furthermore, they also reported that both bifactor exploratory structural equation modeling (ESEM) and exploratory factor analysis (EFA) performed poorly as in estimating both ECV and omegaH. In general, more research is needed to examine how various data factors influence these indices.

The primary outcome investigated in this study included the relative bias in factor loadings estimated from a misspecified unidimensional model. However, other outcomes that have not been investigated in the context of utilizing the bifactor model-based indices to identify essentially unidimensional models include bias in standard errors and coverage. In other words, the accuracy of how well the factor loadings are estimated from the unidimensional model can provide further evidence of the efficacy in treating multidimensional data as unidimensional in this context. Additionally, this study included an examination of select the model-fit indices CFI, TLI, AIC, and BIC in relation to ECV and PUC. Other related studies have also chosen to investigate select model-fit indices, including CFI, RMSEA, and SRMR (Liu et al., 2022; Reise, Scheines, et al.) and CFI and RMSEA (Gu et al., 2017). However, in selecting models in CFA, researchers typically evaluate multiple indices to reach decisions. Future studies should evaluate how all fit indices, including χ^2, function in relation to the unidimensional statistics ECV and PUC.

Finally, more studies are needed to investigate the performance of the bifactor model under a variety of data conditions. For example, this study found severe parameter bias when non-normality was present, and a robust ML estimator was utilized. Future studies are needed to examine the role of non-normality and estimation as it relates to bifactor models. For example, how varying degrees of normality impacts the use of different estimators such as robust ML and WLSMV (i.e., Rhemtulla et al., 2012). In general, the bifactor model remains a highly technical and somewhat misunderstood model (i.e., Reise, 2012).

Summary

The purpose of this study was to investigate the performance of the bifactor model-based indices, explained common variance (ECV), omega hierarchical (omegaH), and the percent of uncontaminated correlations (PUC), under conditions frequently encountered when item wording method effects are identified, including unbalanced method factors (unbalanced in size and magnitude). Results indicated the unidimensional model can be quite robust when method factors are ignored under certain conditions, particularly if the method factors are found to contribute equally to additional item-level variance. The presence of unbalanced method factors (both in size and magnitude) required a substantially "stronger" general factor, as evidenced by higher ECV (>0.80) and omegaH (>0.85) values, to reach "essentially unidimensional" status. Furthermore, PUC was found to be non-informative as an indicator of essential unidimensionality when unbalanced method factors were present (i.e., Liu et al., 2022). Examination of the relationship between ECV and select model-fit indices indicated that while they are somewhat related (i.e., Reise, Scheines, et al., 2013), ECV, as well as omegaH, are more

reliable as indices in selecting essentially unidimensional models in the context of item wording method effects. In general, results from this study offer applied researchers utilizing unidimensional scales that contain both positively and negatively worded items guidance with how to subsequently treat their data if method factors are identified. Ultimately, item wording method effects may not be as problematic as once feared.

REFERENCES

Alessandri, G., Vecchione, M., Eisenberg, N., & Laguna, M. (2015). On the factor structure of the Rosenberg (1965) general self-esteem scale. *Psychological Assessment, 27*(2), 621-635. http://dx.doi.org/10.1037/pas0000073

Arias, V. B., & Arias, B. (2017). The negative wording factor of Core Self-Evaluation Scale (CSES): Methodological artifact, or substantive specific variance? *Personality and Individual Differences, 109,* 28-34. http://dx.doi.org/10.1016/j.paid.2016.12.038

Bader, M., Jobst, L. J., Moshagen, M. (2022). Sample size requirements for bifactor models. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-12. https://doi.org/10.1080/10705511.2021.2019587

Bagozzi, R. P. (1993). Assessing construct validity in personality research: Applications to measures of self-esteem. *Journal of Research in Personality, 27,* 49-87.

Bandalos, D. L., & Gagne, P. (2012). Simulation methods in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 92-108). Guilford Publications.

Bandalos, D. L., & Leite, W. (2013). Use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock, & R. O. Mueller (Eds.), *Structural equation modeling a second course,* (2nd ed., pp. 625-663). Information Age Publishing.

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*(3), 361-370.

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences, 42,* 815-824. DOI:10.1016/j.paid.2006.09.018

Baumgartner, H., & Steenkamp, J. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38,* 143-156.

Bear, G. G., Gaskins, C., Blank, J., & Chen, F. F. (2011). Delaware school climate survey-student: Its factor structure, concurrent validity, and reliability. *Journal of School Psychology, 49,* 157-174. DOI:10.1016/j.jsp.2011.01.001

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. https://www.jstor.org/stable/1435036

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74*(1), 137-143. DOI: 10.1007/S11336-008-9100-1

Billiet, J. B. & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling, 7*(4), 608-628. DOI: 10.1207/S15328007SEM0704_5

Boley, B. B., Jordan, E., & Woosnam, K. M. (2021). Reversed polarity items in tourism scales: Best practice or dimensional pitfall? *Current Issues in Tourism, 24*(4), 466-478. https://doi.org/10.1080/13683500.2020.1774517

Bollen, K. A., & Hoyle, R. H. (2012). Latent variables in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 56-67). Guilford Publications.

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research, 52*(4), 465-484. https://doi.org/10.1080/00273171.2017.1309262

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science, 5*(1), 184-186. DOI:10.1177/2167702616657069

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(4), 504-516. https://doi.org/10.1080/10705511.2014.93859

Brown, T. A. (2003). Confirmatory factor analysis of the Penn State Worry Questionnaire: Multiple factors or method effects? *Behavior Research and Therapy, 41,* 1411-1426. DOI:10.1016/S0005-7967(03)00059-7

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.

Byrne, B. M. (2005). Factor analytic models: Viewing the structure of an assessment instrument from three perspectives. *Journal of Personality Assessment, 85*(1), 17-32. https://doi.org/10.1207/s15327752jpa8501_02

Canivez, G. L. (2016). Bifactor modeling in construct validation of multifactored tests: Implications for understanding multidimensional constructs and test interpretation. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction: Standards and recent advancements* (pp. 247-271). Hogrefe.

Chen, F., Bollen, K., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models. *Sociological Methods & Research*, *29*(4), 468-508.

Chen, H., & Jin, K. (2020). The impact of item feature and response preference in a mixed-format design. *Multivariate Behavioral Research.* https://doi.org/10.1080/00273171.2020.1820308

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41*(2), 189-225.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Cho, S., Cohen, A. S., & Kim, S. (2014). A mixture group bifactor model for binary responses. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(3), 375-395. https://doi.org/10.1080/10705511.2014.915371

Chyung, S. Y., Barkin, J. R., & Shamsy, J. A. (2018). Evidence-based survey design: The use of negatively worded items in surveys. *Performance Improvement, 57*(3), 16-25. DOI: 10.1002/pfi.21749

Cloud, J., & Vaughan, G. M. (1970). Using balanced scales to control acquiescence. *Sociometry, 33*(2), 193-202. http://www.jstor.com/stable/2786329

Corwyn, R. F. (2000). The factor structure of global self-esteem among adolescents. *Journal of Research in Personality*, *34*, 357-379. DOI:10.1006/jrpe.2000.2291

Crocker, L., & Algina, J. (2008). *Introduction to classical & modern test theory.* Cengage Learning.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*(1), 16.

Dalal, D. K. & Carter, N. T. (2015). Negatively worded items negatively impact survey research. In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 112-132). Routledge.

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement, 36*(2), 104-121. DOI: 10.1177/0146621612437403

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing, 13,* 354-378. DOI: 10.1080/15305058.2013.799067

Dierendonck, C., Milmeister, P., Kerger, S., & Poncelet, D. (2020). Examining the measure of student engagement in the classroom using the bifactor model: Increased validity when predicting misconduct at school. *International Journal of Behavioral Development, 44*(3), 279-286. DOI: 10.1177/0165025419876360

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(3), 327-346. https://doi.org/10.1207/S15328007SEM0903_2

DiStefano, C. (2016). Examining fit with structural equation models. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction: Standards and recent advancements* (pp. 166-193). Hogrefe.

DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling: A*

Multidisciplinary Journal, *21*(3), 425-438. https://doi.org/10.1080/10705511.2014.915373

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. https://doi.org/10.1207/s15328007sem1303_6

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, & Evaluation, 14*(20). DOI: https://doi.org/10.7275/da8t-4g52

Donnellan, M. B., Ackerman, R. A., & Brecheen, C. (2016). Extending structural analysis of the Rosenberg Self-Esteem Scale to consider criterion-related validity. Can composite self-esteem scores be good enough? *Journal of Personality Assessment, 98*(2), 169-177. https://doi.org/10.1080/00223891.2015.1058268

Dunn, K. J., & McCray, G. (2020). The place of the bifactor model in confirmatory factor analysis investigations into construct dimensionality in language testing. *Frontiers in Psychology, 11*(1357), DOI: 10.3389/fpsyg.2020.01357

Ebesutani, C., Drescher, C. F., Reise, S. P., Heiden, L., Hight, T. L., Damon, J. D., & Young, J. (2012). The Loneliness Questionnaire-Short Version: An evaluation of reverse-worded and non-reversed worded items via item response theory. *Journal of Personality Assessment, 94*(4), 427-437. DOI: 10.1080/00223891.2012.662188

Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods*, *23*(1), 138-149. doi:10.1037/met0000121

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in *G*-factor models: Explanations and alternatives. *Psychological Methods, 22*(3), 541-562. http://dx.doi.org/10.1037/met0000083

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock, & R. O. Mueller (Eds.), *Structural equation modeling a second course*, (2nd ed., pp. 439-492). Information Age Publishing.

Fiske, D. W. (1982). Convergent-discriminant validation in measurements and research strategies. In D. Brinbirg & L. H. Kidder (Eds.), *Forms of Validity in Research.* Jossey-Bass.

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Sciences, 3*(4), 484-501. https://doi.org/10.1177/2515245920951747

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. https://doi.org/10.1037/1082-989X.9.4.466

Gana, K., Saada, Y., Bailly, N., Joulain, M., Herve, C., & Alaphilippe, D. (2013). Longitudinal factorial invariance of the Rosenberg Self-Esteem Scale: Determining the nature of method effects due to item wording. *Journal of Research in Personality, 47,* 406-416. http://dx.doi.org/10.1016/j.jrp.2013.03.011

Geiser, C., Eid, M., & Nussbeck, F. W. (2008). On the meaning of the latent variables in the CT-C(*M*-1) model: A comment on Maydeu-Olivares and Coffman (2006).

Psychological Methods, 13(1), 49-57. DOI: 10.1037/1082-989X.13.1.49

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57,* 423-436.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007). Full-Information Item Bifactor Analysis of Graded Response Data. *Applied Psychological Measurement*, *31*(1), 4-19. https://doi.org/10.1177/0146621606289485

Gignac, G. E. (2016). The higher-order model imposes a proportionality constraint: That is why the bifactor model tends to fit better. *Intelligence, 55,* 57-68. http://dx.doi.org/10.1016/j.intell.2016.01.006

Gignac, G. E. (2016). On the appropriateness of using items to calculate total scale score reliability via coefficient alpha for multidimensional scales. *European Journal of Psychological Assessment*, *30*(2), 130-139. DOI: 10.1027/1015-5759/a000181

Gignac, G. E., & Watkins, M. W. (2013). Bifactor modeling and the estimation of model-based reliability in the WAIS-IV. *Multivariate Behavioral Research, 48*, 639-662. DOI: 10.1080/00273171.2013.804398

Gignac, G. E., & Kretzschmar, A. (2017). Evaluating dimensional distinctness with correlated-factor models: Limitations and suggestions. *Intelligence, 62,* 138-147. http://dx.doi.org/10.1016/j.intell.2017.04.001

Gnambs, T., & Schroeders, U. (2020). Cognitive abilities explain wording effects in the Rosenberg Self-Esteem Scale. *Assessment, 27*(2), 404-418. https://doi.org/10.1177/1073191117746503

Gnambs, T., & Staufenbiel, T. (2018). The structure of the General Health Questionnaire (GHQ-12): Two meta-analytic factor analyses, *Health Psychology Review*, *12*(2), 179-194. https://doi.org/10.1080/17437199.2018.1426484

Goldberg, D. P., & Williams, P. (1988). *A users' guide to the General Heath Questionnaire*. GL Assessment.

Gordon, R. A., & Peng, F. (2020). Evidence regarding the domains of the CLASS preK in Head Start classrooms. *Early Childhood Research Quarterly, 53,* 23-39. https://doi.org/10.1016/j.ecresq.2020.01.008

Gorsuch, R. (1983). *Factor analysis*. Erlbaum Associates.

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*(1), 155-167. DOI: 10.1007/S11336-008-9099-3

Greene, A. L., Eaton, N. R., Li, K., Forbes, M. K., Krueger, R. F., Markon, K. E., ... & Kotov, R. (2019). Are fit indices used to test psychopathology structure biased? A simulation study. *Journal of Abnormal Psychology*, *128*(7), 740-764. https://doi.org/10.1037/abn0000434

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*, 1241-1254.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6*(4), 430-450. DOI: 10.1037//1082-989X.6.4.430

Gu, H., Wen, Z., & Fan, X. (2015). The impact of wording effect on reliability and validity of the Core Self-Evaluation Scale (CSES): A bi-factor perspective.

Personality and Individual differences, 83, 142-147. http://dx.doi.org/10.1016/j.paid.2015.04.006

Gu, H., Wen, Z., & Fan, X. (2017). Examining and controlling for wording effect in a self-report measure: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(4), 545-555. https://doi.org/10.1080/10705511.2017.1286228

Gustafsson, J. & Aberg-Bengtsson, L. (2010). Unidimensionality and interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 97-121). American Psychological Association. http://dx.doi.org/10.1037/12074-005

Hancock, G. R. & Mueller, R. O. (2001). Rethinking construct reliability with latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 195-216). Scientific Software International, Inc.

Hazlett-Stevens, H., Ullman, J. B., & Craske, M. G. (2004). Factor structure of the Penn State Worry Questionnaire. *Assessment, 11*(4), 361-370. DOI: 10.1177/1073191104269872

Herche, J., & Engelland, B. (1996). Reversed-polarity and scale unidimensionality. *Journal of the Academy of Marketing Science*, *24*(4), 366-374.

Holzinger, K. J., & Swineford, F. (1937). The Bi-factor method. *Psychometrika, 2,* 41-54. https://doi.org/10.1007/BF02287965

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*(3), 329-367. https://doi.org/10.1177/0049124198026003003

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

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. https://doi.org/10.1080/10705519909540118

Huang, F. L. (2014). Using a bifactor model to assess the factor structure of the Phonological Awareness Literacy Screening for grades 1 through 3. *Journal of Psychoeducational Assessment, 32*(7), 638-650. DOI: 10.1177/0734282914525026

Huang, C., & Dong, N. (2012). Factor structure of the Rosenberg Self-Esteem Scale: A meta-analysis of pattern matrices. *European Journal of Psychological Assessment, 28*(2), 132-138. DOI: 10.1027/1015-5759/a000101

Hyland, P., Boduszek, D., Dhingra, K., Shevlin, M., & Egan, A. (2014). A bifactor approach to modelling the Rosenberg Self-Esteem Scale. *Personality and Individual Differences*, *66*, 188-192. http://dx.doi.org/10.1016/j.paid.2014.03.034

Hystad, S. W., & Johnsen, B. H. (2020). The dimensionality of the 12-item General Health Questionnaire (GHQ-12): Comparisons of factor structures and

invariance across samples and time. *Frontiers in Psychology*, *11*, 1-10. https://doi.org/10.3389/fpsyg.2020.01300

Jennrich, R. I. & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika, 76*(4), 537-549. DOI: 10.1007/S11336-011-9218-4

Jennrich, R. I. & Bentler, P. M. (2012). Exploratory bi-factor analysis: The oblique case. *Psychometrika*, *77*(3), 442-454. DOI: 10.1007/S11336-012-9269-1

Johnson, R. L., & Morgan, G. B. (2016). *Survey scales*. The Guilford Press.

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294-316). SAGE.

Judge, T. A., Erez, A., Bono. J. E., & Thoresen, C. J. (2003). The Core Self-Evaluation Scale: Development of a measure. *Personnel Psychology*, *56,* 303-331.

Kam, C. C. S. (2018). Why do we still have an impoverished understanding of the item wording effect? An empirical examination. *Sociological Methods & Research, 47*(3), 574-597. DOI: 10.1177/0049124115626177

Kam, C. C. S., Meyer, J. P., & Sun, S. (2021). Why do people agree with both regular and reversed items? A logical response perspective. *Assessment*, *28*(4), 1110-1124. DOI:10.1177/10731911211001931

Kelley, K., & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for composite measures. *Psychological Methods*, *21*(1), 69-92. http://dx.doi.org/10.1037/a0040086

Kline, R. B. (2011). *Principles and practices of structural equation modeling* (3rd ed.). Guilford Publications.

Kopp, J. P., Zinn, T. E., Finney, S. J., & Jurich, D. P. (2011). The development and evaluation of the Academic Entitlement Questionnaire. *Measurement and Evaluation in Counseling and Development, 44*(2), 105-129. https://doi.org/10.1177/0748175611400292

Lei, P. W. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Quality and Quantity, 43*(3), 495-507. DOI 10.1007/s11135-007-9133-z

Li, C. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavioral Research, 48*, 963-949. DOI 10.3758/s13428-015-0619-7

Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., Raudsepp, L., Liukkonen, J., & Thogersen-Ntoumani, C. (2012). Method effects: The problem with negatively Versus positively keyed items. *Journal of Personality Assessment, 94*(2), 196-204. https://doi.org/10.1080/00223891.2011.645936

Liu, Y., & Thompson, M. S. (2021). General factor mean difference estimation in bifactor models with ordinal data. *Structural Equation Modeling: A Multidisciplinary Journal, 28*(3), 423-439. https://doi.org/10.1080/10705511.2020.1833732

Liu, X., Cai, P., Lai, X. Wen, J., & Yang, Y. (2022). Assessing essential unidimensionality of scales and structural coefficient bias. *Educational and Psychological Measurement,* 1-20. DOI: 10.1177/00131644221075580

Manslof, M. & Reise, S. P. (2016). Exploratory bifactor analysis: The Schmid-Leiman orthogonalization and Jennrich-Bentler analytic rotations. *Multivariate*

Behavioral Research, 51(5), 698-717. DOI:10.1080/00273171.2016.1215898

Markon, K. E. (2019). Bifactor and hierarchical models: Specification, inference, and interpretation. *Annual Review of Clinical Psychology, 15,* 51-69. https://doi.org/10.1146/annurev-clinpsy-050718-095522

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70*(4), 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*, *11*(3), 320-341. https://doi.org/10.1207/s15328007sem1103_2

Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment, 22*(2), 366-381. DOI: 10.1037/a0019225

Maul, A. (2013). Method effects and the meaning of measurement. *Frontiers in Psychology, 4,* 1-13. DOI: 10.3389/fpsyg.2013.00169

Maydeu-Olivares, A. (2017). Maximum likelihood estimation of structural equation models for continuous data: Standard errors and goodness of fit. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(3), 383-394. https://doi.org/10.1080/10705511.2016.1269606

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*(4), 344-362. DOI: 10.1037/1082-989X.11.4.344

McDonald, R. P. (1999). Test theory: A unified treatment. Erlbaum.

McIver, J. P. & Carmines, E. G. (1981). *Unidimensional Scaling*. Sage Publications. https://dx.doi.org/10.4135/9781412986441

McKay, M. T., Boduszek, D., & Harvey, S. A. (2014). The Rosenberg Self-Esteem Scale: A bifactor answer to a two-factor question? *Journal of Personality Assessment*, *96*(6), 654-660. https://doi.org/10.1080/00223891.2014.923436

McKay, M. T., Cole, J. C., & Percy, A. (2015). Further evidence for a bifactor solution for the Considerations of Future Consequences Scale: Measurement and conceptual implications. *Personality and Individual Differences, 83*, 219-222. http://dx.doi.org/10.1016/j.paid.2015.04.022

McKay, M. T., Morgan, G. B., van Exel, N. J., & Worrell, F. C. (2015). Back to "the future": Evidence of a bifactor solution for scores on the Consideration of Future Consequences Scale. *Journal of Personality Assessment*, *97*(4), 395-402. https://doi.org/10.1080/00223891.2014.999338

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412-133. http://dx.doi.org/10.1037/met0000144

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavioral Research Methods, 52*, 2287-2305. https://doi.org/10.3758/s13428-020-01398-0

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development and validation of the Penn State Worry Questionnaire. *Behavior Research and Therapy*, *28*(6), 487-495. https://doi.org/10.1016/0005-7967(90)90135-6

Michaelides, M. P., Koutsogiorgi, C., & Panayiotou, G. (2016). Method effects on an adaptation of the Rosenberg Self-Esteem Scale in Greek and the role of personality traits. *Journal of Personality Assessment, 98*(2), 178-188. https://doi.org/10.1080/00223891.2015.1089248

Morin, A. J. S., Arens, A. K., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 116-139, http://dx.doi.org/10.1080/10705511.2014.961800

Morin, A. J. S., Myers, N. D., & Lee, S. (2020). Modern factor analytic techniques: Bifactor models, exploratory structural equation modeling (ESEM) and bifactor-ESEM. In G. Tenebaum & R. C. Eklund, R. C., *Handbook of sport psychology* (4th ed., pp. 1044-1073). Wiley.

Morgan, G. B., Hodge, K. J., Wells, K. E., & Watkins, M. W. (2015). Are fit indices biased in favor of bi-factor models in cognitive ability research? A comparison of fit in correlated factors, higher-order, and bi-factor models via Monte Carlo simulations. *Journal of Intelligence*, *3*(1), 2-20. DOI:10.3390/jintelligence3010002

Murray, A. L., Booth, T., Eisner, M., Obsuth, I., & Ribeaud, D. (2019). Quantifying the strength of the general factors in psychopathology: A

comparison of CFA with maximum likelihood estimation, BSEM, and ESEM/ EFA bifactor approaches. *Journal of Personality Assessment*, *101*(6), 631-643. https://doi.org/10.1080/00223891.2018.1468338

Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence*, *41,* 407-422. http://dx.doi.org/10.1016/j.intell.2013.06.004

Muthén, B.O. (1998-2004). *Mplus technical appendices*. Muthén & Muthén.

Nandakumar, R., & Ackerman, T. (2004). Test modeling. In D. Kaplan (Ed.), *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 93-106). SAGE Publications

Paulhus, D. L. & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224-239). The Guilford Press

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(2), 287-312. https://doi.org/10.1207/S15328007SEM0802_7

Podsakoff, P. M., MacKenzie, S. B., Lee, J., & Podsakoff, N. P. (2003). Common method bias in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879-903. DOI: 10.1037/0021-9010.88.5.879

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations for how to control it. *Annual*

Review of Psychology, 63, 539-569. DOI: 10.1146/annurev-psych-120710-100452

Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research, 41*(3), 227-259. https://doi.org/10.1207/s15327906mbr4103_1

Quilty, L. C., Oakman, J. M., & Risko, E. (2006). Correlates of the Rosenberg Self-Esteem Scale method effects. *Structural Equation Modeling, 13*(1), 99-117. https://doi.org/10.1207/s15328007sem1301_5

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 111-163. https://doi.org/10.2307/271063

Rammstedt, B. & Farmer, R. F. (2013). The impact of acquiescence on the evaluation of personality structure. *Psychological Assessment*, *25*(4), 1137-1145. DOI: 10.1037/a0033323

Rauch, W. A., Schweizer, K., & Moosbrugger, H. (2007). Method effects due to social desirability as a parsimonious explanation of the deviation from unidimensionality in LOT-R scores. *Personality and Individual Differences, 42,* 1597-1607. DOI:10.1016/j.paid.2006.10.035

Raykov, T. & Marcoulides, G. A. (2006). *A first course in structural equation modeling.* Taylor & Francis Group.

Raykov, T., & Pohl, S. (2012). On studying common factor variance in multiple-component measuring instruments. *Educational and Psychological Measurement, 73*(2), 191-209. DOI: 10.1177/0013164412458673

Raykov, T., & Pohl, S. (2013). Essential unidimensionality examination for multicomponent scales: An interrelationship decomposition approach.

Educational and Psychological Measurement, 73(4), 581-600. DOI: 10.1177/0013164412470451

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667-696. https://doi.org/10.1080/00273171.2012.715555

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95*(2), 129-140, https://doi.org/10.1080/00223891.2012.725437

Reise, S. P., Cook, K. F., & Moore, T. M. (2015). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 13-40).

Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively reweighted least squares to the Rosenberg Self-Esteem Scale. *Multivariate Behavioral Research, 51*(6), 818-838. DOI:10.1080/00273171.2016.1243461

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of* Personality *Assessment, 92*(6), 544-559. https://doi.org/10.1080/00223891.2010.496477

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research Journal, 16*(19), 19-31, DOI: 0.1007/s11136-007-9183-7

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, *73*(1), 5-26. DOI: 10.1177/0013164412449831

Reise, S. P., Waller, N. G., Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12*(3), 287-297. DOI: 10.1037//1040-3590.12.3.287

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145-154. DOI: 10.1007/S11336-008-9102-Z

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354. https://doi.org/10.1037/a0029315

Rodrigo, M. F., Molina, J. G., Losilla, J.-M., Vives, J., & Tomas, J. M. (2019). Method effects associated with negatively and positively worded items on the 12-item General Heath Questionnaire (GHQ-12): Results from a cross-sectional survey with a representative sample of Catalonian workers. *BMJ Open, 9*(e031859), 1-8, DOI:10.1136/bmjopen-2019-031859

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98*(3), 223-237. https://doi.org/10.1080/00223891.2015.1089249

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137-150. http://dx.doi.org/10.1037/met0000045

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

Salerno, L., Ingoglia, S., & Coco, G. L. (2017). Competing factor structures of the Rosenberg Self-Esteem Scale (RSES) and its measurement invariance across clinical and non-clinical samples. *Personality and Individual Differences, 113,* 13-19. http://dx.doi.org/10.1016/j.paid.2017.02.063

Savalei, V., & Reise, S. P. (2019). Don't forget the model in your model-based reliability coefficients: A reply to McNeish (2018). *Collabra: Psychology, 5*(1), 36. DOI: https://doi.org/10.1525/collabra.247

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,* 1063-1078.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23-74.

Schmitt, T. A., Sass, D. A., Chappelle, W., & Thompson, W. (2018). Selecting the "best" factor structure and moving measurement validation forward: An illustration.

Journal of Personality Assessment, 100(4), 345-362. https://doi.org/10.1080/00223891.2018.1449116

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.

Schumacker, R. E., & Lomax, R. G. (2010). *A beginners guide to structural equation modeling* (3rd ed.). Routledge.

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.

Strathman, A., Gleicher, F., Boninger, D. S., & Edwards, C. S. (1994). The consideration of future consequences: Weighing immediate and distant outcomes of behavior. *Journal of Personality and Social Psychology, 66*(4), 742-752.

Stefansson, K. K., Gestsdottir, S., Geldhof, G. J., Skulason, S., & Lerner, R. M. (2016). A bifactor model of school engagement: Assessing general and specific aspects of behavioral, emotional and cognitive engagement among adolescents. *International Journal of Behavioral Development, 40*(5), 471–480. https://doi.org/10.1177/0165025415604056

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589-617. https://doi.org/10.1007/BF02294821

Stucky, B. D., & Edelen, M. O. (2015). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A.

Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 183-206).

Swain, S. D., Weathers, D., & Niedrich, R. W. (2010). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research, 45*(1), 116-131. DOI:10.1509/jmkr.45.1.116

Ten Berge, J. M. F., & Socan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69*(4), 613-625.

Urban, R., Szigeti, R., Kokonyei, G., & Demetrovics, Z. (2014). Global self-esteem and method effects: Competing factor structures, longitudinal invariance and response styles in adolescents. *Behavioral Research Methods, 46*(2), 488-498. DOI:10.3758/s13428-013-0391-5

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. DOI:10.1371/journal.pone.0068967

Van Vaerenberg, Y. & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research, 25*(3), 195-217. DOI:10.1093/ijpor/eds021

Wang, W., Chen, H., & Jin, K. (2015). Item response theory models for wording effects in mixed-format scales. *Educational and Psychological Measurement, 75*(1), 157-178. DOI: 10.1177/0013164414528209

Wang, J., Siegal, H. A., Falck, R. S., & Carlson, R. G. (2001). Factorial structure of Rosenberg’s self-esteem scale among crack-cocaine drug users. *Structural Equation Modeling, 8*(2), 275-286.

https://doi.org/10.1207/S15328007SEM0802_6

Watkins, M. W. (2017). The reliability of multidimensional neuropsychological measures: From alpha to omega. *The Clinical Neuropsychologist, 31*(6), 1113-1126. http://dx.doi.org/10.1080/13854046.2017.1317364

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 209-231). Guilford Publications.

Weems, G. H., Onwuegbuzie, A. J., Schreiber, J. B., & Eggers, S. J. (2003). Characteristics of respondents who respond differently to positively and negatively worded items on rating scales. *Assessment & Evaluation in Higher Education, 28*(6), 587-606. https://doi.org/10.1080/0260293032000130234

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research, 49,* 737-747.

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods, 18*(3), 320-334. DOI: 10.1037/a0032121

Woods, C. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28*(3), 189-194. DOI: 10.1007/s10862-005-9004-7

Xia, X. (2018). *Examining negative wording effect in a self-report measure* [Unpublished doctoral dissertation]. University of Pittsburg.

Xinya, L., & Yanyun, Y. (2016). Confirmatory factor analysis under violations of structural and distributional assumptions: A comparison of robust maximum

likelihood and Bayesian estimation methods. *Journal of Psychological Sciences, 39*(5), 1256-1267. DOI:10.16719/j.cnki.1671-6981.20160536

Xu, M., & Leung, S. (2016). Bifactor structure for the categorical Chinese Rosenberg Self-Esteem Scale. *The Spanish Journal of Psychology, 19*(e67), 1-11. DOI:10.1017/sjp.2016.66

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29*(4), 377-392. DOI: 10.1177/0734282911406668

Ye, S. (2009). Factor structure of the General Health Questionnaire (GHQ-12): The role of wording effects. *Personality and Individual Differences, 46*, 197-201. DOI:10.1016/j.paid.2008.09.027

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*(2), 213-249.

Ziegler, M., & Hagemann, D. (2015). Testing the unidimensionality of items. *European Journal of Psychological Assessment, 31*(4), 231-237, DOI: 10.1027/1015-5759/a000309

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123-133. DOI: 10.1007/s11336-003-0974-7

Zinbarg, R. E., Yovel, I., & Revelle, W. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for $\omega_{H.}$

Applied Psychological Measurement, 30(2), 121-144. DOI:

10.1177/0146621605278814

APPENDIX A

SAMPLE *MPLUS* SIMULATION CODE

```
TITLE: THIS IS CODE FOR DATA GENERATION FOR CONDITION 3. !12GEN
LOADING OF 0.5 6POS and 6NEG LOADING OF 0.3

MONTECARLO:
names = u1-u12;
generate = u1-u12(4);
categorical = u1-u12;
nrep = 1000;
seed = 21822;
repsave=all;
save=66NORM_3_rep*.DAT;
nobs = 500;

ANALYSIS:
estimator = WLSMV; !MLMV implemented in a separate analysis.
processors = 4;

MODEL POPULATION:
f1 BY u1-u12@.5;  !General factor.
f2 BY u1-u6@.3;    !Positive wording factor items 1-6.
f3 BY u7-u12@.3;  !Negative wording factor items 7-12.
f1-f3 @1;
f1 WITH f2@ 0; !All factors are orthogonal.
f1 WITH f3@ 0;
f2 WITH f3@ 0;

!Item thresholds   !These are for the all items normal data condition.
[u1$1-u12$1*-1.50
u1$2-u12$2*-0.5
u1$3-u12$3*0.5
u1$4-u12$4*1.50];

u1-u12*.66  !Uniqueness terms
```

```
TITLE: THIS IS THE BIFACTOR MODEL ANALYSIS FOR CONDITION 3

DATA: FILE IS 66NORM_3_replist.dat;
type=montecarlo;

VARIABLE: NAMES ARE u1-u12;

MODEL:
f1 BY u1-u12*; !General factor.
f2 BY u1-u6*; !Positive wording factor.
f3 BY u7-u12*; !Negative wording factor.
f1-f3@1;
f1 WITH f2@ 0; !All orthogonal.
f1 WITH f3@ 0;
f2 WITH f3@ 0;

ANALYSIS:
estimator=MLMV;

OUTPUT: TECH1;

savedata: results are 66NORM_3_BF_results.DAT;


TITLE: THIS IS THE UNIDIMENSIONAL MODEL ANALYSIS FOR CONDITION 3

DATA: FILE IS 66NORM_3_replist.dat;
type=montecarlo;

VARIABLE: NAMES ARE u1-u12;

MODEL:
f1 BY u1-u12*; !General factor only.
f1@1

ANALYSIS:
estimator=MLMV;

OUTPUT: TECH1;

savedata: results are 66NORM_3_UNI_results.DAT;
```

APPENDIX B

ADDITIONAL RESULTS FOR NON-NORMALLY DISTRIBUTED DATA

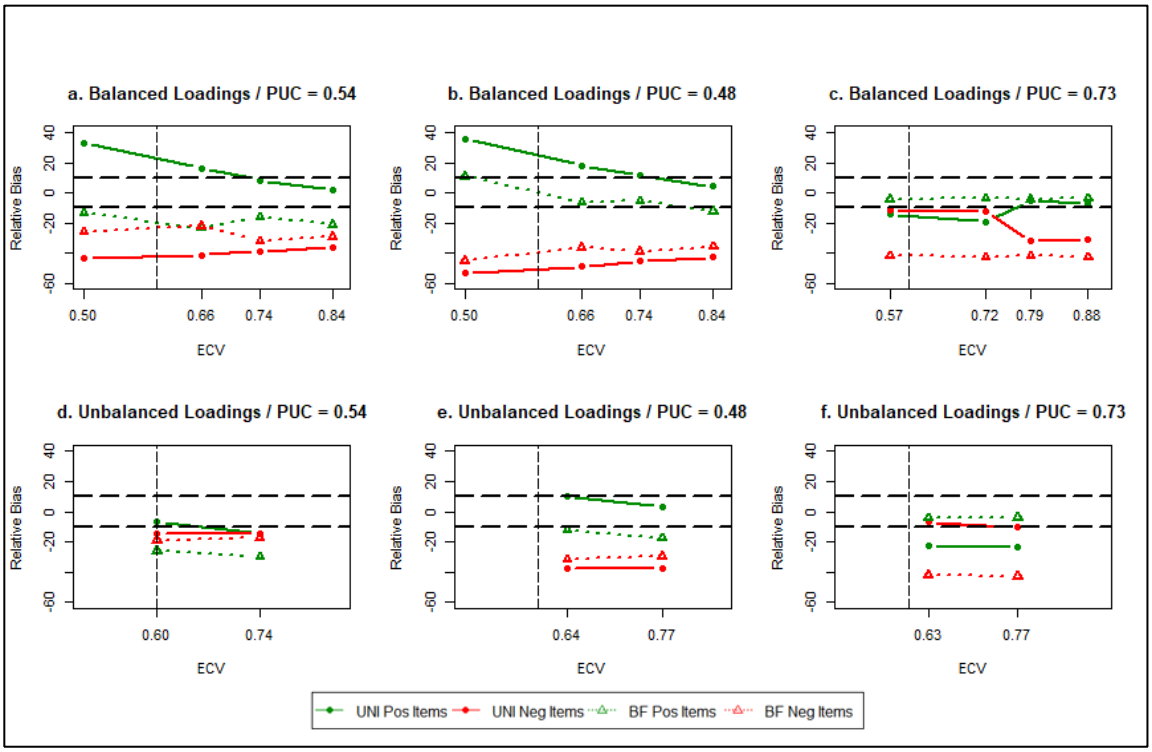

Figure B.1 *Average Percent Relative Bias in Estimated General Factor Loadings by Explained Common Variance (ECV), Percent of Uncontaminated Correlations (PUC), Loading Type, and Model Type for Non-Normally Distributed Data.* Reference lines placed at ±10% to note acceptable bias threshold. Reference line placed at ECV of 0.60 for comparison to Reise, Scheines, et al. (2013)

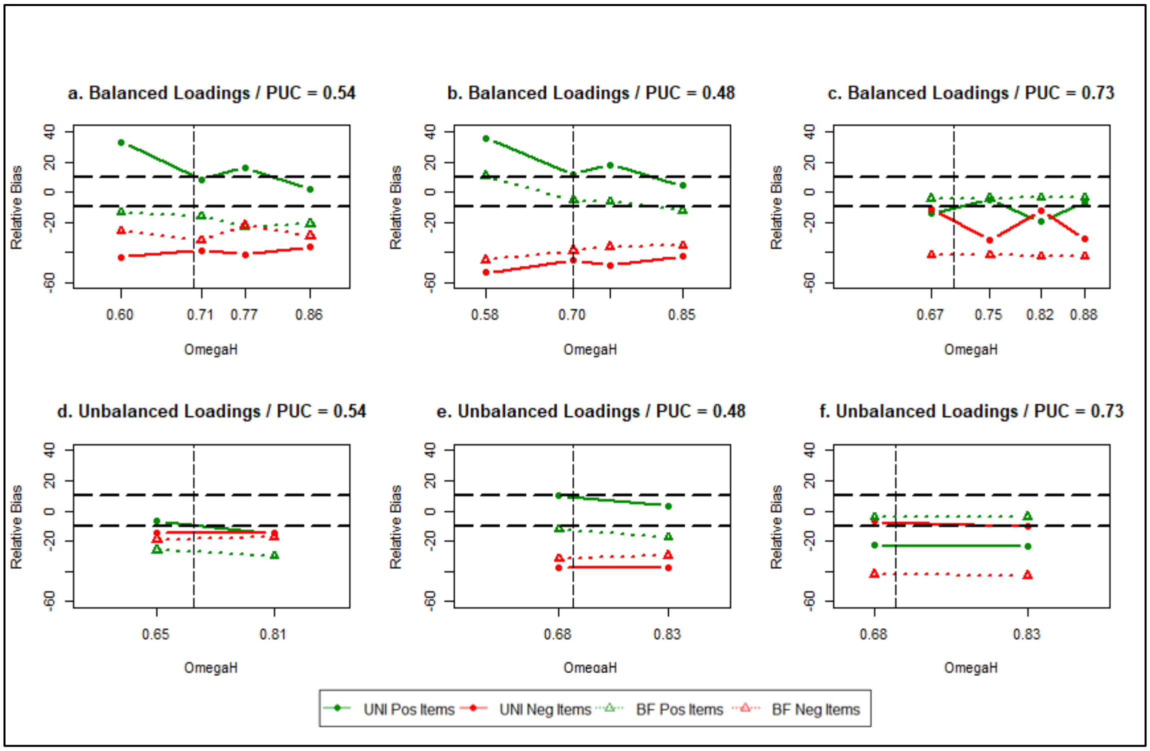

Figure B.2 *Average Percent Relative Bias in Estimated General Factor Loadings by Omega Hierarchical (omegaH), Percent of Uncontaminated Correlations (PUC), Loading Type, and Model Type for Non-Normally Distributed Data.* Reference lines placed at ±10% to note acceptable bias threshold. Reference line placed at omegaH of 0.70 for comparison to Reise, Scheines, et al. (2013)

Table B.1 *Average CFI and TLI Fit Values for the Bifactor Model and Unidimensional Model: Non-Normally Distributed Data*

PUC		Bifactor Model				Unidimensional Model				Difference[b]	
		CFI		TLI		CFI		TLI			
$(N_p:N_n)$	ECV	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	ΔCFI	ΔTLI
0.54 (6:6)	0.84	**1.00**	0.003	**1.01**	0.010	0.87	0.027	0.84	0.033	**0.13**	**0.16**
	0.74	**1.00**	0.007	**1.00**	0.024	0.91	0.034	0.89	0.042	0.09	**0.12**
	0.74[a]	**1.00**	0.003	**1.00**	0.009	0.73	0.031	0.67	0.038	**0.27**	**0.33**
	0.66	**1.00**	0.002	**1.00**	0.006	0.72	0.031	0.66	0.038	**0.28**	**0.34**
	0.60[a]	**1.00**	0.006	**1.00**	0.021	0.76	0.043	0.71	0.053	**0.23**	**0.30**
	0.50	**1.00**	0.004	**1.00**	0.014	0.76	0.041	0.71	0.050	**0.23**	**0.29**
0.48 (8:4)	0.84	**1.00**	0.002	**1.00**	0.007	0.94	0.020	0.92	0.024	0.06	0.08
	0.77[a]	**1.00**	0.002	**1.00**	0.007	0.82	0.028	0.78	0.034	**0.18**	**0.22**
	0.74	**1.00**	0.005	**1.00**	0.017	**0.96**	0.023	**0.95**	0.028	0.04	0.06
	0.66	**1.00**	0.002	**1.00**	0.005	0.86	0.022	0.83	0.027	**0.14**	**0.17**
	0.64[a]	**1.00**	0.005	**1.00**	0.017	0.86	0.039	0.83	0.047	**0.14**	**0.18**
	0.50	**1.00**	0.003	**1.00**	0.010	0.90	0.026	0.87	0.032	**0.10**	**0.13**
0.73 (3:6)	0.88	**1.00**	0.006	**1.00**	0.013	0.88	0.025	0.86	0.030	**0.11**	**0.14**
	0.79	**0.99**	0.013	**1.00**	0.032	0.92	0.036	0.91	0.044	0.07	0.09
	0.77[a]	**1.00**	0.004	**1.00**	0.010	0.78	0.029	0.73	0.035	**0.22**	**0.27**
	0.72	**1.00**	0.005	**1.00**	0.010	0.71	0.030	0.64	0.037	**0.29**	**0.35**
	0.63[a]	**0.99**	0.012	**1.00**	0.026	0.82	0.039	0.78	0.048	**0.17**	**0.22**
	0.57	**0.99**	0.011	**1.00**	0.024	0.73	0.044	0.67	0.054	**0.26**	**0.33**

Note. CFI= Comparative Fit Index, TLI = Tucker-Lewis Index. Gray shaded boxes indicate cases where average fit was within the recommended value of acceptable fit (≥ .95).

[a]Conditions have unbalanced loadings across the positive and negative wording method factors (PWF, NWF); in these conditions, the positive items loaded 0.3 on the PWF and the negative items loaded 0.5 on the NWF.

[b]Difference in fit is the average bifactor model fit index – the average unidimensional model fit index; differences >.10 are bolded.

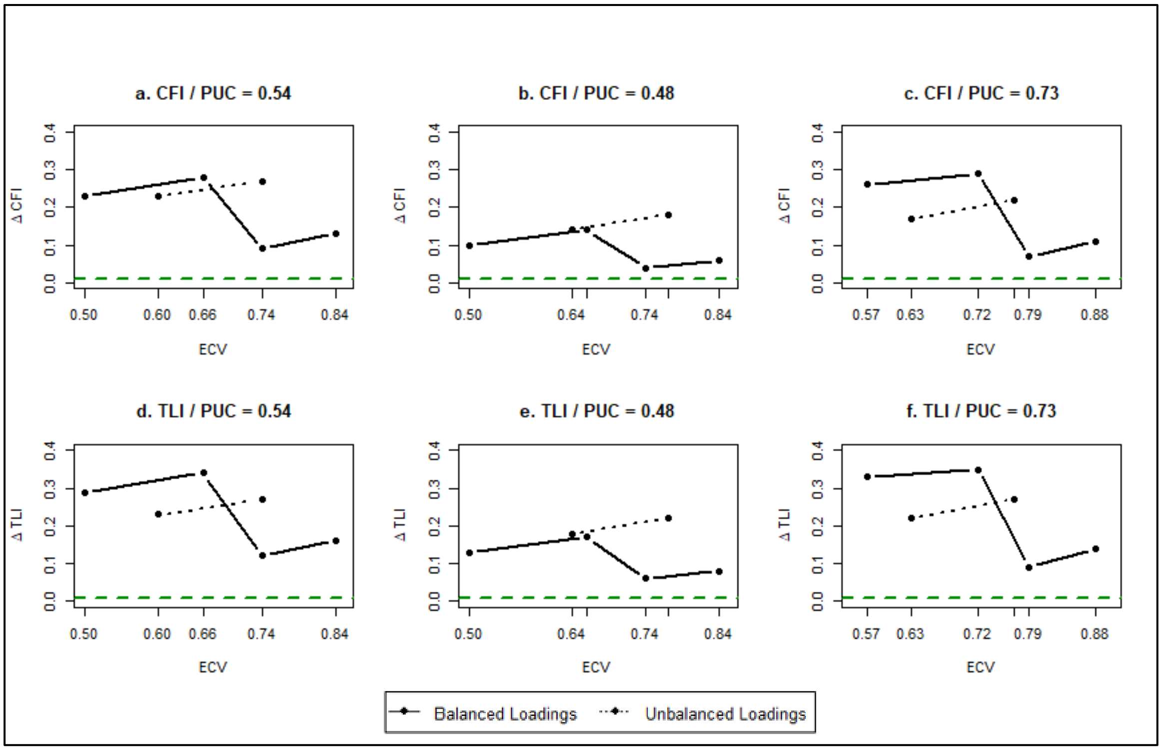

Figure B.3 *Difference in Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) between the Bifactor Model and the Unidimensional Model by Explained Common Variance (ECV), Percent of Uncontaminated Correlations (PUC), and Loading Type for Non-Normally Distributed Data.* Reference line placed at a difference of 0.01.

Table B.2 *Average AIC and BIC Fit Values for the Bifactor Model and Unidimensional Model: Non-Normally Distributed Data*

PUC		Bifactor Model				Unidimensional Model				Difference[b]	
		AIC		BIC		AIC		BIC			
(N_p:N_n)	ECV	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	ΔAIC	ΔBIC
0.54 (6:6)	0.84	**14171.84**	252.48	**14374.14**	252.48	14442.32	244.62	14594.04	244.62	-270.48	-219.90
	0.74	**15552.28**	230.52	**15754.58**	230.52	15617.94	223.02	15769.66	223.02	-65.66	-15.08
	0.74[a]	**13532.13**	274.50	**13734.44**	274.50	14191.63	269.97	14343.36	269.97	-659.50	-608.92
	0.66	**12703.23**	269.54	**12905.53**	269.54	13696.95	278.34	13848.67	278.34	-993.72	-943.14
	0.60[a]	**15296.64**	243.11	**15498.95**	243.11	15508.81	245.24	15660.53	245.24	-212.17	-161.58
	0.50	**14847.90**	239.60	**15050.20**	239.60	15196.13	251.17	15347.85	251.17	-348.23	-297.65
0.48 (8:4)	0.84	**14195.54**	193.21	**14397.84**	193.21	14341.37	191.93	14493.09	191.93	-145.83	-95.25
	0.77[a]	**13846.90**	187.41	**14049.20**	187.41	14313.40	194.19	14465.13	194.19	-466.50	-415.93
	0.74	**15703.20**	174.21	15905.50	174.21	15729.62	174.65	**15881.34**	174.65	-26.42	**24.16**
	0.66	**12660.47**	201.26	**12862.77**	201.26	13211.03	211.44	13362.76	211.44	-550.56	-499.99
	0.64[a]	**15570.05**	190.86	**15772.35**	190.86	15709.51	189.56	15861.23	189.56	-139.46	-88.88
	0.50	**14890.23**	189.37	**15092.53**	189.37	15071.14	199.16	15222.86	199.16	-180.91	-130.33
0.73 (3:6)	0.88	**14425.83**	247.22	**14615.49**	247.22	14620.32	245.10	14772.05	245.10	-194.49	-156.56
	0.79	**15683.02**	219.42	**15872.68**	219.42	15721.88	220.90	15873.60	220.90	-38.86	**-0.92**
	0.77[a]	**13777.81**	261.10	**13967.47**	261.10	14267.71	252.88	14419.43	252.88	-489.90	-451.96
	0.72	**13535.71**	265.83	**13725.37**	265.83	14245.95	260.58	14397.68	260.58	-710.24	-672.31
	0.63[a]	**15439.45**	232.76	**15629.10**	232.76	15571.24	232.99	15722.97	232.99	-131.79	-93.87
	0.57	**15312.22**	232.13	**15501.88**	232.13	15542.38	233.68	15694.10	233.68	-230.16	-192.22

Note. AIC=Akaike Information Criterion, BIC = Bayesian Information Criterion. Gray shaded cells indicate conditions where AIC and/or BIC favor the model.

[a]Conditions have unbalanced loadings across the positive and negative wording method factors (PWF, NWF); in these conditions, the positive items loaded 0.3 on the PWF and the negative items loaded 0.5 on the NWF.

[b]Difference in fit is the average bifactor model fit index – the average unidimensional model fit index

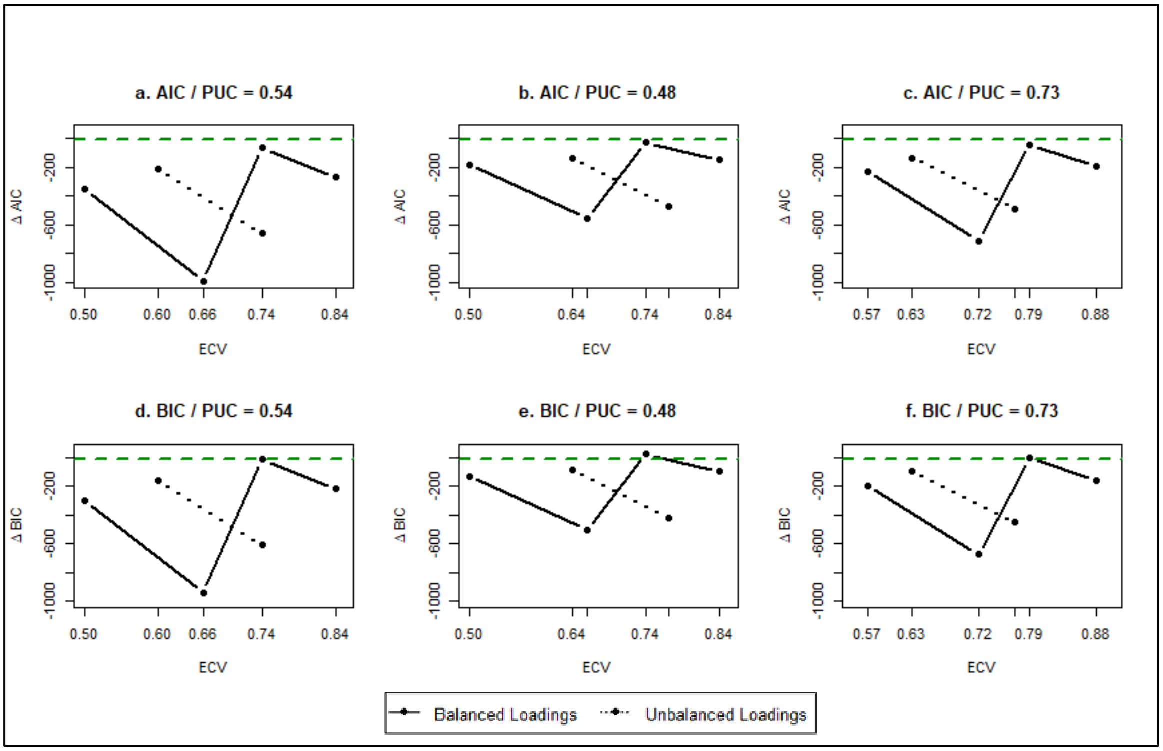

Figure B.4 *Difference in Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) between the Bifactor Model and the Unidimensional Model by Explained Common Variance (ECV), Percent of Uncontaminated Correlations (PUC), and Loading Type for Non-Normally Distributed Data.* Reference line placed at a difference of less than 10.