EXAMINING NEGATIVE WORDING EFFECT IN A SELF-REPORT MEASURE

by

Xiaoyan Xia

B.A., Kean University, 2010

M.A., University of Pittsburgh, 2015

Submitted to the Graduate Faculty of

the School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Xiaoyan Xia

It was defended on

November 29, 2018

and approved by

Clement A. Stone, Professor, Department of Psychology in Education

Lan Yu, Associate Professor, Department of Medicine

Dissertation Advisor: Feifei Ye, Senior Scientist, RAND Corporation, Pittsburgh

Suzanne Lane, Professor, Department of Psychology in Education

EXAMINING NEGATIVE WORDING EFFECT IN A SELF-REPORT MEASURE

Xiaoyan Xia, PhD

University of Pittsburgh, 2018

Researchers often include both positively and negatively worded items in one survey to reduce acquiescence bias. The incorporation of negatively worded items can raise concerns for the internal-consistency coefficients, the validity evidence for criterion relationships and the internal structure of the measure. This study aims to investigate the impact of misspecifying the model when using negatively worded items. Simulated datasets were generated from three models, 1) CFA with two correlated factors, 2) bi-factor CFA with two specific factors for positive and negative wording effects, and 3) bi-factor CFA with one specific factor for negative wording effect, and compared with each other and the unidimensional model. Models were compared with respect to model fit, and their estimation of internal-consistency coefficients, criterion-related validity coefficients, and the internal structure validity.

Approximate and comparative model fit indices were not informative for model comparison because they presented similar fit among the three multidimensional models, although they tended to correctly identify the misfit of the unidimensional model under some conditions. Misspecifying the model for the negative wording effect resulted in biased estimates of internal-consistency coefficients. For the data generation bi-factor model with two specific factors, the under-fitting bi-factor model with the negative wording effect overestimated *the homogeneity coefficient*. When there were positive and negative wording effects, omitting one or both specific factors resulted in underestimated criterion-related validity coefficients and biased factor loadings. However, over-fitting with an additional specific factor did not impact the

estimation of criterion-related validity coefficients or factor loadings of the general factor and the other specific factor.

Results suggest that model fit indices provide limited information for selecting models for negatively worded items. Evaluation of internal consistency reliability, criterion-related validity, and internal structure validity is recommended when selecting an approach for modeling negatively worded items. Researchers still need to rely on substantive and conceptual grounds when examining the nature of negatively worded items.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

The basis for this study initially stemmed from my interest in exploring the nature of negatively worded items. As the use of negatively worded items becomes prevalent in survey instruments, there will be a greater need to justify for modeling the negative wording effect. Misspecifying the model for the negative wording effect presents challenges to internal-consistency coefficients, the validity evidence for criterion relationships and the internal structure of the measure.

This study would have been impossible to complete without receiving support and help in a number of different ways. It is a great pleasure that I write the acknowledgements to thank those people who have been caring, encouraging, and supporting me. First of all, I want to extend my sincerest thanks to all my committee members including Dr. Feifei Ye, Dr. Suzanne Lane, Dr. Clement Stone, and Dr. Lan Yu, each of whom has provided substantive feedback and guidance. Especially, I want to thank Dr. Feifei Ye, my research advisor, for her extraordinary academic guidance and her incredible motivational capabilities. She has been acting as a role model by providing much advice, encouragement, and support throughout my entire doctoral study. She read multiple versions of this dissertation and provided extensive comments during my dissertation research process. I would also like to express my deepest gratitude to Dr. Suzanne Lane, my co-advisor, for her constant support and caring in processing this research. It

was extremely helpful that she provided valuable comments on my write-up which improved the quality of this study.

Secondly, I would like to thank all my friends for their being supportive and keeping me balanced in any possible way. I consider myself very lucky and my warmly thanks go to all of them. Finally, thanks for everything, my beloved parents. Without the love they have always given me, I would not be the person I am.

1.0 INTRODUCTION

The use of negatively worded items (also called negatively keyed items) is prevalent in survey instruments to control for acquiescence bias or response set. However, research has shown that negatively worded items may present challenges to internal-consistency coefficients, the validity evidence for criterion relationships and the internal structure of the measure (Gu, Wen, & Fan, 2017). For example, the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965), which contains a balanced number of positively and negatively worded items, has been studied extensively, but its dimensionality is still under debate (Gnambs, Scharl, & Schroeders, 2018). Researchers have argued about whether this scale represents a unidimensional self-esteem construct with additional covariance among negatively worded items modeled by a method effect, or bi-dimensional self-esteem constructs separated by positively and negatively worded items (Carmines & Zeller, 1979; Marsh, 1996; Tomas & Oliver, 1999).

Method effect related to negatively worded items is prevalent in self-report measures. Researchers have questioned the nature of the method effect: whether it is a measurement artifact and substantively irrelevant, or it represents a response style which can be substantively interpreted in terms of individual characteristics (Gana et al., 2013; Marsh, Scalas, & Nagengast, 2010). The correlated trait-correlated uniqueness (CTCU) (Horan, DiStefano, & Motl, 2003; Tomas & Oliver, 1999) model treats the wording effect as a methodological artifact by correlating the item residuals within positively worded items and/or negatively worded items.

The correlated trait-correlated method (CTCM) (Horan et al., 2003) model treats the wording effect as a distinct latent factor representing a response style that could correlate with other substantive factors such as personality traits (Bollen & Paxton, 1998). The correlated trait-uncorrelated method (CTUM) and the CTCM minus one (CT-C[M-1]) can be considered as variations of the CTCM model. The CTCM model allows for a correlation between wording-effect factors. In contrast, the CTUM model restricts a zero inter-factor correlation between wording effects. The CT-C(M-1) model specifies one fewer wording-effect factors than the CTCM; the CT-C(M-1) incorporates one wording-effect factor only, rather than both. Motl and colleagues (DiStefano & Motl, 2006, 2009; Horan et al., 2003; Motl & DiStefano, 2002) claimed that the method factor has similar psychometric properties as a substantive factor which supports the interpretation of a method effect as a personality trait, while other researchers argued otherwise (Alessandri, Vecchione, Tisak, & Barbaranelli, 2011).

Recent research emerges using bi-factor models to account for wording effects. Bi-factor models have been applied to model multidimensionality of measures when all items share common variances and a set of items share variances over and beyond the common trait (Reise, 2012). When a scale measures one single trait contaminated with wording effects, a bi-factor model is a special case of the CTCM, CTUM, or CT-C(M-1). For example, for the Rosenberg Self-Esteem Scale (RSES), bi-factor models consider specific factor(s) related to positively or negatively worded items or both. In this case, a bi-factor model with two specific factors associated with positive and negative wording is equivalent to a CTCM (i.e., correlated specific factors) or a CTUM (i.e., uncorrelated specific factors) model. A bi-factor model with one specific factor associated with positive or negative wording is identified as a CT-C(M-1) model. The bi-factor model with two specific factors associated with positive and negative wording is

deemed the best model. There is no consensus, however, about whether these two specific factors represent a method or substantive effect (Alessandri, Vecchione, Eisenberg, & Laguna, 2015; Reise, Kim, Mansolf, & Widaman, 2016).

The empirical research on modeling negatively worded items have relied heavily on model fit indices (e.g., CFI, RMSEA) to select the optimal model. For example, Alessandri et al. (2015) compared ten models for the Rosenberg Self-Esteem Scale (RSES) in terms of chi-square, CFI, RMSEA along with 95% confidence interval (CI) for RMSEA, and AIC to identify the bi-factor model with two specific factors as the optimal model. However, to present, research is scarce regarding the performance of model fit indices in selecting the correctly specified model for negatively worded items. There are a few exceptions.

Donnellan, Ackerman, and Brecheen (2016) used TLI, CFI, RMSEA along with a 90% confidence interval (CI) for RMSEA, SRMR, AIC, and BIC to compare and evaluate nine models on the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965) using empirical data. The consistent estimates of the validity evidence for criterion relationships across various-fitting models implied that when the true underlying structure was unknown, model fit indices did not function well in selecting a model. Gu et al. (2017) and Reise, Scheines, Widaman, and Haviland (2013) demonstrated that when the true underlying structure was bi-factor, model fit indices were able to identify misspecified unidimensional model as fitting well under certain conditions. Both Monte Carlo studies (Gu et al., 2017; Reise et al., 2013) focused on the fit comparions between true bi-factor and misspecified unidimensional models only. In addition, Morgan, Hodge, Wells, and Watkins (2015) argued that model fit indices tended to correctly select the true model over misspecified correlated factor models when the true underlying data structure was bi-factor. However, model fit indices favored a bi-factor model under certain conditions when the true

underlying data structure was correlated factor. Simply relying on model fit results is not recommended for judging correct model specification.

Research in bi-factor modelling (Reise et al., 2013; Rodriguez, Reise, & Haviland, 2016) has stressed the use of explained common variance (ECV) and other statistics (e.g., coefficient omega and omega hierarchical) for interpretation of general and specific factors, which can be applied to bi-factor models for method effects. These statistics are valuable for evaluating internal-consistency coefficients, the validity evidence for criterion relationships and the internal structure of the measure. ECV is an indicator of the general factor strength. ECV can vary by changing the number of positively and negatively worded items and their factor loadings. When the ECV is high (e.g., >.75), the scale is judged to be essentially unidimensional. Bias in estimation of validity evidence based on criterion relationships was found to be reversely correlated with ECV (Gu et al., 2017; Reise et al., 2013).

Omega indices are used to disentangle the variance explained by general or specific factors from the total variance. For instance, omega hierarchical treats method effect(s) as measurement error and the square root of omega hierarchical refers to the correlation between the general trait factor and the observed total score. Misspecifying the model when using negatively worded items underestimated the coefficient omega but overestimated the omega hierarchical (Gu et al., 2017). However, Gu et al. (2017) only generated true bi-factor model structures limited to a negative wording effect. Although empirical studies (Gu, Wen, & Fan, 2015; Kam, 2016) have demonstrated the sufficiency of modeling only one wording effect, which was primarily related to negatively worded items, numerous studies have evaluated the bi-factor model with two specific factors associated with positive and negative wording. Nowadays, models such as CFA with two correlated factors, bi-factor CFA with two specific factors, and bi-

factor with one specific factor are still three common options in applied research (e.g., Gana et al., 2013; Gnambs, Scharl, & Schroeders, 2018). Further studies that examine the impact of misspecifying the model for wording effects on the estimation of internal-consistency coefficients, the validity evidence for criterion relationships and the internal structure of the measure are necessary.

The present study is interested in how model fit indices perform when the data generation models for mixed-format scales represent different factorial structures (i.e., two correlated factors, bi-factor model with two specific factors for positive and negative wording effects, and bi-factor model with one specific factor for the negative wording effect). The prior simulation studies compared various fit indices for true bi-factor and misspecified unidimensional models (Gu et al., 2017; Reise et al., 2013), or assessed how model fit indices functioned in selecting a model between bi-factor and correlated factor models when the true underlying structure was known to be one of these two models (Morgan et al., 2015). In contrast, the present study compared various fit indices for four models including the two-factor CFA, the bi-factor with two specific factors for positive and negative wording effects, the bi-factor model with one specific factor for the negative wording effect, and the unidimensional model. The unidimensional model was fitted in each data generating structure as a reference model to investigate the impact of wording effects on the validity evidence for criterion relationships (Donnellan et al., 2016). The present study added to the literature by varying factor loadings, inter-factor correlations, and the degree of prediction of the targeted criterion (i.e., criterion-related validity coefficients). Additionally, the impact on internal-consistency coefficients, the validity evidence for criterion relationships and the internal structure of the measure when misspecifying the models for negatively worded items was examined. Outcome measures

included bias, internal-consistency coefficients, and the validity evidence for criterion relationships in addition to model fit.

1.1 PURPOSE OF THE STUDY AND RESEARCH QUESTIONS

The primary objective was to examine the effect of misspecifying the model when using negatively worded items. Three models were examined, 1) two correlated factor CFA with positively worded items on one factor and negatively worded items on the other factor; 2) bi-factor CFA with two specific factors representing method factors related to positive and negative wording effects; 3) bi-factor CFA with one specific factor representing a method factor related to a negative wording effect. The research questions included

1) How well do model fit indices perform in identifying the correct model for negative wording effects?
2) What are the effects of negative wording on the estimates of internal-consistency coefficients?
3) What are the effects of negative wording on the validity evidence for criterion relationships and the internal structure of the measure?

It was postulated that when the true underlying structure was known to be one of the three models (i.e., correlated factor CFA model, bi-factor CFA with two specific factors, and bi-factor CFA with one specific factor), model fit indices tended to select the data generation model as the optimal-fitting model. It was hypothesized that the wording effect was primarily associated with negatively worded items. If a negative wording effect was misspecified, the internal-consistency coefficients, the validity evidence for criterion relationships and the internal

structure of the measure were biased and misleading inferences would be made. It was not expected that there was a significant positive wording impact on the internal-consistency coefficients, the validity evidence for criterion relationships and the internal structure of the measure.

1.2 SIGNIFICANCE OF STUDY

The current study investigated the performance of model fit indices in identifying the correct specification of negatively worded items, and the impact of misspecifying the model for the negative wording effect on the internal-consistency coefficients, the validity evidence for criterion relationships and the internal structure of the measure. This study compared various fit indices for four models, which added to the literature of model fit comparisons among one-factor, correlated factors, and bi-factor models. It was hypothesized that model fit indices have enough power to select the true model. However, if these hypothesizes were not supported, such as that the three models were not distinguishable in model fit when the true model was a two-factor CFA model, the implication is that model fit comparison is not recommended when researchers examine whether the negatively worded items form a method or substantive factor. This suggests that researchers should exercise extra care when drawing inferences about the corresponding approaches for modeling negatively worded items.

This study has practical significance for researchers using self-reported measures containing negatively worded items. Empirical researchers should first consider the original rationale for including negatively worded items in a measure before directly employing any widely used models. Given the particular constructs of interest and scale items, researchers

should be able to judge whether the positively and/or negatively worded items lead to a methodological artifact. If the items are somewhat confusing or unclear, a method effect may result from such poorly worded items. Moreover, researchers should check to see whether responses are invalid based upon observed responses to positively and negatively worded items. These behaviors provided preliminary justification for modeling wording effects. The present study addresses conditions under what reporting a total score is legitimate for a measure containing negatively worded items. If wording effects were found to be associated with positively and negatively worded items jointly, modeling a negative wording effect only would not be sufficient. If researchers are not sure whether including both wording effects is redundant, researchers are suggested to evaluate the internal-consistency coefficients, the validity evidence for criterion relationships and the internal structure of the measure for both bi-factor models.

2.0 LITERATURE REVIEW

This chapter provides definitions of terms in the area of wording effect in self-report measures, followed by rationales and assumptions for including negatively worded items. This chapter also discusses researchers' concerns posed on the use of the negatively worded items. Moreover, this chapter reviews and evaluates various statistical procedures used in previous research studies to explore wording effects. In addition, performance of selected SEM fit indices is depicted and studies regarding the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965) are summarized as an example.

2.1 MIXED-FORMAT SCALES

A mixed-format scale refers to a self-report inventory containing both positively and negatively worded items. Mixed-format scales are often designed to measure the same latent construct. For example, the Life Orientation Test Revised (Scheier, Carver, & Bridges, 1994) contains both positively worded items (e.g., "I'm always optimistic about my future") and negatively worded items (e.g., "I hardly ever expect things to go my way") to measure optimism/pessimism. Similarly, the Penn State Worry Questionnaire contains positively worded items (e.g., "My worries overwhelm me") and negatively worded items (e.g., "I do not tend to worry about

things") to measure 'anxious experiences' or 'deny the anxious experiences'. Another one of the most widely used scales in psychology is the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965). This scale is a balanced scale with five positively worded items and an equivalent number of negatively worded items. Rosenberg's self-esteem scale was originally conceptualized as measuring one's unitary personal attitudes (either positive or negative) toward the self. In these instances, positively and negatively worded items captured the positive and negative pole of the same underlying construct. Researchers presumed that after negatively worded items were reversely coded, the negatively worded items performed the same as the positively worded items.

2.2 (TYPES OF) NEGATIVELY WORDED ITEMS

A negatively worded item refers to an item that appears in a negative manner opposed to the logic of the construct being measured (Weijters & Baumgartner, 2012). One simple example can be "I am not happy." Developing such items requires creating phrases that denote a negation of the construct through the use of the word "no" or adjectives, adverbs, and even verbs, that offer a negative meaning.

Schriesheim, Eisenbach, and Hill (1991) offered three ways to institute negation: 1) regular or direct negation (i.e., reverse oriented), 2) polar opposites (i.e., reverse wording), and 3) negation of the polar opposite. In particular, the inclusion of negative particles ("not" or "no") or affixal negations ("un" or "less") can create regular or direct negation negatively worded items. Using words with an opposite meaning produce the polar opposite negatively worded items. For example, if a regular item is 'I am happy,' then a corresponding 1) regular or direct negation negatively worded item could be 'I am not happy', a corresponding 2) polar opposite negatively

worded item could be ‘I am sad’ and a corresponding 3) negation of the polar opposite negatively worded item could be ‘I am not sad.’ Psychological measures popularly use 1) regular or direct negation and 2) polar opposites (Zhang & Savalei, 2016). The majority of the negatively worded items were created using the first method: regular or direct negation (Swain, Weathers, & Niedrich, 2008). Since agreeing to these items implies low levels of the target construct, observed scores to such items should be reversed-scored.

2.2.1 Negatively worded versus negatively keyed

By definition, when an item is reversely scored, such an item is negatively keyed. A negatively keyed item can be (grammatically) negatively worded or (grammatically) positively worded. In contrast, a negatively worded item can be negatively keyed (i.e., reversed-scored prior to summing to create a total score) or positively keyed (i.e., summed to produce a scale score without reverse scoring). A significant number of items were both negatively worded and negatively keyed (Coleman, 2013, presented a detailed analysis of the different combinations of wording and keying). However, many researchers did not distinguish the term of negatively worded from the other term of negatively keyed. For instance, Weijters and Baumgartner (2012) defined items as negatively worded when items were written in the opposite pole of the construct being measured and when the observed responses were reversed before computing attribute standing. Essentially, Weijters and Baumgartner’s (2012) definition of negatively worded items somewhat pointed to the definition of negatively keyed items. This dissertation used Weijters and Baumgartner (2012)’s definition of negatively worded items.

2.2.2 Rationales of including negatively worded items

The inclusion of negatively worded items has become so commonplace that the majority of published works incorporated such items in the studied scales without specifying the reason of such inclusion. The two most frequently stated reasons for including negatively worded items are 1) to reflect the past scales that contain negatively worded items, that is, others already included negatively worded items and 2) to minimize response styles (Dalal & Carter, 2015). For instance, Sauley and Bedeian (2000) stated the reason for adopting both positively and negatively worded items was to lessen the acquiescent bias. Consistently, later work, including one study by Sanders (2009), recommended the incorporation of negatively worded items.

In survey research, respondent acquiescence refers to respondents uncritically agreeing with items, regardless of the item content (Messick, 1991; Paulhus, 1991; Ray, 1983). The cognitive process underlying acquiescence is in line with Gilbert's (1991) dual-stage model of belief (Knowles & Condon, 1999). According to Gilbert (1991), respondents first understand a statement by instinctively accepting the content; the next stop includes the gathering of essential information. In the dual-stage model, therefore, acquiescence eliminates this second level; pertinent and perhaps contradictory material is neither gathered nor constructed (Knowles & Condon, 1999; Krosnick, 1999). Acquiescence intrinsically leads to correct responses for true items but incorrect responses for false items.

Ideally, acquiescence to positively worded items compensates for acquiescence to negatively worded items (Billiet & McClendon, 2000), which leads to an unbiased summed scale score (Marsh, 1996). Stemming from such an ideal expectation, researchers suggest using a balanced number of positively and negatively worded items in a self-report measure (e.g. Paulhus, 1991). Since acquiescent respondents tend to say 'yes' to all items, their summed scores

on responses inflate scale means when items are phased in one direction. Involving both positively and negatively worded items address such inflation of scale means because responses to positively worded items are biased in one direction, and responses to negatively worded items are biased in the opposite direction.

Balanced scales neither eliminate acquiescent responding nor remove bias from individual items, however, this approach is intended to ensure that on a given scale, acquiescent respondents receive a summated score near the scale mean (Cloud & Vaughan, 1970). "Without this balance, it is difficult to establish how much of the distinction between different factors is due to differences in the underlying constructs being measured as opposed to method effects" (Marsh, 1996; p. 817).

2.2.3 Assumptions for including negatively worded items

An overarching assumption underlying reverse-scoring of negatively worded items is the interchangeability between positively and negatively worded items. According to Dalal and Carter (2015), four assumptions are involved in the inclusion of negatively worded items.

First, the use of negatively worded items is assumed to either minimize response tendencies or help detect respondents engaging in response tendencies.[1] Inspection of responding patterns to positively and negatively worded items can be used to identify individuals who are engaging in a particular response tendency (Swain et al., 2008). Second, the use of negatively worded items is assumed to not impair internal-consistency coefficients. Researchers expect no added measurement error or additional concern with the utilization of mixed-format scales.

[1] See the first point from 'Potential problems associated with negatively worded items' for dissimilar functions of incorporating negatively worded items on response tendencies.

Third, researchers postulate that mixed-format scales are valid. When involving negatively worded items, this assumption regarding inferences about the validity evidence for criterion relationships must be investigated. Fourth, negatively worded items are assumed to measure a given construct in an equivalent way as positively worded items (Marsh, 1996). Items written with different wordings are expected to gauge the same construct.

Unfortunately, empirical studies have not included pairs of reverse-worded items to ensure the measure of a same target construct, such as using both "I am happy" and "I am not happy" to measure respondents' happiness. Rather, in an attempt to increase the breadth of the construct while keeping the number of items small, researchers may be tempted to include negatively worded items that are slight variations of the positively worded items. Therefore, responses to subsets of positively worded or negatively worded items do not necessarily measure matched components of the target construct.

2.2.4 Potential problems associated with negatively worded items

Many concerns have been posed on the use of the negatively worded items. First, some researchers argued that the use of negatively worded items does not lessen the acquiescence bias. For instance, Sauro and Lewis (2011) noted a similar amount of extreme reactions between positively and negatively worded items. Sonderen, Sanderman, and Coyne (2013) also claimed that such bias was not reduced by reversing half of the items. Consistently, Weijters, Geuens, and Schillewaert (2009) indicated that when negatively worded items were located very closely to each other, respondents perceived positively and negatively worded items similarly at one cognitive level. When the negatively worded item appeared at every sixth item, then negatively worded items functioned to lessen the acquiescence bias.

Second, negatively worded items may confuse respondents due to increased difficulty in interpreting such items. Participants spend more time reading the questions and response options of negatively worded items (Kamoen, Holleman, Mak, Sanders, & van den Bergh, 2011). Respondents may feel challenged to map their agreement level to the item with a negation. Mapping replies to response options in negatively worded items can be a harder, longer process (Chessa & Holleman, 2007). Longer processing might mirror the processing complexity (Bassili & Scott, 1996). Not understanding the negatively worded items may lead to an increase of non-responses (Colosi, 2005) and a decrease of mean scores on negatively worded items (Weems, Onwuegbuzie, Schreiber, & Eggers, 2003). When negated negatively worded items are included, respondents might not notice a negative particle in the item. Such carelessness might cause them to incorrectly read 'I am not happy' as 'I am happy.' If at least 10% of participants respond carelessly, then a method effect emerges in a principal component analysis (Schmitt & Stults, 1985) and a one-factor solution is unacceptable in terms of model fit (Woods, 2006).

Third, negatively worded items may lead to aberrant psychometric properties of the mixed-format scales. If researchers are unware of the impact of negatively worded items, this systematic bias will be treated as item residuals; as a result, measurment error will increase (Bollen, 1989; Mulaik, 1971). Researchers who employed mixed-format scales often discovered that negatively worded items have a slightly lower internal-consistency coefficient and weaker item-to-total correlations when compared to positively worded items (Barnette, 1999; Benson & Hocevar, 1985; Cronbach, 1942; Roszkowski & Soven, 2010; Schriesheim & Eisenbach, 1995). Across a series of studies, the internal-consistency coefficients are lowest in the mixed-format scales while the internal-consistency coefficients are highest in the scales with positively-worded items only; this result further supports the argument that the difficulty in understanding

negatively worded items causes the increased error (Schriesheim & Eisenbach, 1995; Schriesheim et al., 1991; Schriesheim & Hill, 1981). Using both empirical and simulated data in the IRT framework, Wang, Chen, and Jin (2015) demonstrated that when the true data structure was unidimensional without a wording effect, fitting the bi-factor with a wording effect to the unidimensional data presented little harm. However, when the true data structure was with a wording effect, ignoring the wording effect resulted with a positive bias in internal-consistency coefficients. The internal-consistency coefficient is very sensitive to the existence of negatively worded items: even a small proportion of negatively worded items (e.g., 2 out of 20 items) can diminish the internal consistency (Roszkowski & Soven, 2010). Thus, factor analyses often favor a two-factor solution over the one-factor solution of a measure (Reise, Morizot, & Hays, 2007; Woods, 2006).

Rather than increasing measurement error, some researchers have questioned if the systematic bias introduced by negatively worded items may actually introduce a common method variance, thereby inflating correlations across different scales (e.g., Magazine et al., 1996). If this is the case, convergent validity evidence tainted with common method bias would be artificially high. Independent of whether the systematic bias increases measurement error (thereby deflating internal-consistency coefficients and criterion-related validity coefficients) or increases common method variance (thereby inflating convergent validity coefficients), using mixed-format scales can have serious implications for the validity evidence based on relations to other variables.

Fourth, negatively worded items might affect the dimensionality of the target construct. Consider the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965) as one example. Some researchers have frequently reported that a single factor accounts for significant variance in the

RSES, supporting a one-factor solution (Bagley, Bolitho, & Bertrand, 1997; Pullmann & Allik, 2000; Shevlin, Bunting, & Lewis, 1995; Tomas & Oliver, 1999). In this case, determining the respondent's score on the RSES involves summing the respondent's responses to the statements and using the overall score to determine the respondent's self-esteem. The negatively worded items were reversely scored while the positively worded items were taken as they were. This one-factor finding is in accordance with the scale's conceptualization as a unidimensional scale measuring self-esteem.

However, alternative factor structures, such as a dual dimensional model has been proposed by some researchers (e.g., Ang, Neubronner, Oh, & Leong, 2006; Boduszek, Hyland, Dhingra, & Mallett, 2013; Boduszek, Shevlin, Mallett, Hyland, & O'Kane, 2012). In the dual dimensional model, researchers argued two distinct but correlated constructs separated by positively worded and negatively worded items are a reflection of the fundamental dimensions of self-esteem (e.g., Boduszek, Hyland, Dhingra, & Mallett, 2013; Boduszek, Shevlin, Mallett, Hyland, & O'Kane, 2012). The researchers may, in this instance, incorrectly conclude that the measure taps into two distinct yet correlated psychological variables when in reality, these two variables are only a function of the wording of the items and not an accurate representation of the respondent's score on the measure. Still, no consensus has been reached upon the dimensionality of the RSES.

In summary, in mixed-format scales, negatively worded items tend to be inter-correlated, regardless of whether positively and negatively worded items measure the same dimension of the target construct. Negatively worded items could adversely impact the internal-consistency coefficients, the validity evidence for criterion relationships and the internal structure of the measure (Benson & Hocevar, 1985; Chessa & Holleman, 2007; Clark, 1976; Cronbach, 1946;

Goldsmith & Desborde, 1991; Holleman, 1999; Kamoen et al., 2011; Ory, 1982; Riley-Tillman, Chafouleas, Christ, Briesch, & LeBel, 2009; Schriesheim & Hill, 1981; Weems et al., 2003). Negatively worded items impact the way in which respondents think and use the latent construct to organize their beliefs, thereby impacting the validity evidence for criterion relationships and the internal structure of the measure (Weijters & Baumgartner, 2012). When the wording effect is not modeled, biased estimates of internal-consistency coefficients and criterion-related validity coefficients occur (Gu et al., 2015).

2.2.5 Findings on the nature of item wording effects

Researchers have also questioned the nature of this wording effect factor—whether this factor is a spurious method factor or a substantive factor representing response style and individual characteristics (e.g., Lance et al., 2009). If wording effects are meaningfully interpretable and reflect personality traits, they might have specific substantive correlates. In particular, Rauch, Schweizer, and Moosbrugger (2007) claimed that respondents have various social desirability response styles related to positively and negatively worded items. Such response styles could cause respondents to react differently to a positively worded item and its counterpart negatively worded item, thereby artificially generating a factor due to item wording (DiStefano & Motl, 2006). In their empirical study (2007), Rauch et al. found that the correlation between the artificial factor extracted from positively worded items and social desirability response style is significant (r =.35). This indicates that social desirability response style contributes to respondents' tendency of positive self-reporting. In contrast, a negative wording effect may represent a consistent behavioral trait such as apprehension about others (e.g., DiStefano & Motl,

2006). A recent study conducted by Alessandri et al. (2011) also supported that the wording effect does not substantively represent any latent construct.

To address whether correlational findings in terms of wording effects in a particular survey instrument can be generalized to findings from other survey instruments, researchers examined the correlations among the wording factors extracted from various instruments. For example, DiStefano and Motl (2006) found that one negative wording factor from the Rosenberg Self-Esteem Scale (RSES) weakly correlated with the negative wording factor from an anxiety instrument (r =.37). Consistently, Pohl and Steyer (2010) found the correlation between a negative wording factor from a calmness measure and a negative wording factor from an alertness measure is weak (r =.35). Item wording is scale-specific as correlations are low (Kam, 2016).

Moreover, according to Kam (2016), it is not necessary to model wording factors from both positively and negatively worded items. Kam (2016) suggested that researchers should be aware that redundancy could occur if they model both positive and negative wording effects as separate factors; modeling a single factor eliminates the need for another factor (Gu et al., 2015). Further, wording effect is primarily associated with negatively worded items (Lindwall et al., 2012; Quilty, Oakman, & Risko, 2006), possibly a result of the interpretational difficulty from negatively worded items (Sonderen et al., 2013; Swain et al., 2008). In addition, the positively-worded totals have a closer similarity to the scores for items of direct negation than for those of negative items with polar opposite wording, suggesting that different types of negatively worded items do not evoke the same reactions from participants (Solís Salazar, 2015).

To enrich the current understanding of the manner in which people analyze and react to survey items requires an understanding of the nature of item wording (Borsboom, Mellenbergh,

& van Heerden, 2004). It is important to recognize that not all cultures share the same issues related to the use of negatively worded items (Wong, Rindfleisch, & Burroughs, 2003). Examination of the same survey in different languages also indicates that different cultures react differently to the same negative items. Even using the same scale, if translated into different languages, respondents react to the same negatively worded items dissimilarly. Whether the factor formed from negatively worded items is an artifact or a substantive factor depends upon the nature of the construct and the quality and type of the negatively worded items. Correctly modelling wording effects would minimize the bias in estimation of internal-consistency coefficients, the validity evidence for criterion relationships and the internal structure of the measure, though no conclusion has been reached on the exact nature of the item wording effect (Kam, 2016). Researchers unaware of the existence of the effect of negatively worded items would either ignore or inappropriately model such effect, which leads to inaccurate conclusions.

2.2.6 When negatively worded items are appropriate and necessary?

Although the use of negatively worded items is related to a number of issues, which seems to suggest researchers should exclude negatively worded items in surveys, such exclusion is a step backward from previous literature because scales with positively worded items produce acquiescence bias. Researchers have stressed the importance of using negatively worded items to diminish the acquiescence bias (Barnette, 2000; Baumgartner & Jan-Benedict, 2001; Cronbach, 1946; Nunnally, 1978). Such bias emerges when respondents select the statement that does not convey how they feel (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003), or when respondents react to a pattern out of lethargy, apathy, or an automatic adaptation. Acquiescence bias taints the

covariance structure of the data (Savalei & Falk, 2014), therefore, including negatively worded items is beneficial.

There are instances in which negatively worded items are not only appropriate but also necessary. One instance is when creating bi-dimensional scales (Cacioppo, Gardner, & Berntson, 1997). For example, the Positive and Negative Affect Scale (Watson & Clark, 1988) was designed to assess multiple constructs separated by positively and negatively worded items. Researchers cannot sum up a single score by combining the positively and negatively worded items for bi-dimensional scales. Instead, these scales require a separate score for the positively and negatively worded items. Researchers should not only use negatively worded items for bi-dimensional scales, but also they must pay attention to scoring such scales.

The other instance relates to Thurstone scales (Thurstone, 1928) which represent an ideal-point response process. Items from the Thurstonian approach are developed to cover all aspects of a self-report continuum, including positive, moderate, and negative regions. In order to scale extreme-positive, moderate, and extreme-negative attribute standings, it is necessary to have items tapping these levels of the trait (Drasgow, Chernyshenko, & Stark, 2010; Roberts, Donoghue, & Laughlin, 2000). Therefore, negatively worded items play an essential role in tapping the negative attribute standing on ideal-point scales. In each of these situations, researchers should pay significant attention to ensure that the negatively worded items are clearly developed and are measuring the construct of interest.

Researchers need to consider the consequences that result from validating a scale or analyzing a conceptual model when they decide if it is better to involve only positively worded items and risk exposure to potential acquiescence bias, or whether it is better to involve both positively and negatively worded items, a choice that might lead to erratic responses and a

decrease in the internal consistency of the scale. A scale with positively worded items can result in a shared variance bias, leading to exaggerated associations and a more favorable evaluation of a theoretical-based model. The use of both positively and negatively worded items could invalidate a suggested scale or model, which actually is valid and reliable. Dimensionality issues are a reflection of the nature of the scale and have a number of implications for the scoring, evaluation, and interpretation of the scale. Researchers should be cognizant of the fact that using a mixed-format scale can introduce one factor upon which only the negatively worded items load. If researchers decide to include negatively worded items, they must try to verify that the respondents have the ability to discern the negatively worded items in the pilot test of the measures (Hughes, 2009).

2.3 STATISTICAL PROCEDURES USED TO DEAL WITH WORDING EFFECTS

Due to the lack of consensus on the nature of wording effect, it is challenging to determine appropriate statistical approaches to identify and control such effect. Early researchers either directly ignored those item wording effects or conducted simple correlation analyses. Advancements in statistical and methodological strategies in recent decades have opened up new possibilities to address method effects. Alternative analyses to investigate wording effect, for example, include Multitrait-Multimethod (MTMM) strategies and various bi-factor modeling.

2.3.1 One-factor solution

In a unifactor approach, all items on one scale, including positively worded items and reverse-coded negatively worded items, should assess a sole latent construct. This approach does not take wording effect into account. See Model 1 below as one example: five positively and five negatively worded items from the Rosenberg Self-Esteem Scale (RSES). Model 2-10 were illustrated using the same scale.

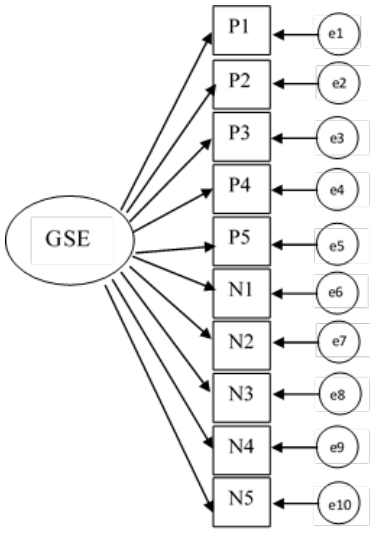

Figure 1. Model 1 One trait factor, no correlated residuals

2.3.2 Two-factor solution

A two-factor approach models the positively worded items as one factor and the negatively worded items as the other distinct factor; these two factors are assumed to measure different latent constructs but they are expected to be correlated (see Model 2). Model 3 is a reduced model which represents two orthogonal traits. Such an approach may violate the intent of a single latent construct and thus raise concerns with interpretability of the scale.

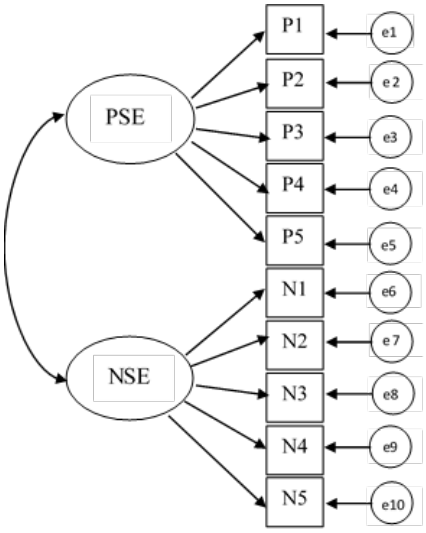

Figure 2. Model 2 Two correlated traits: Correlated positive and negative self-esteem factors

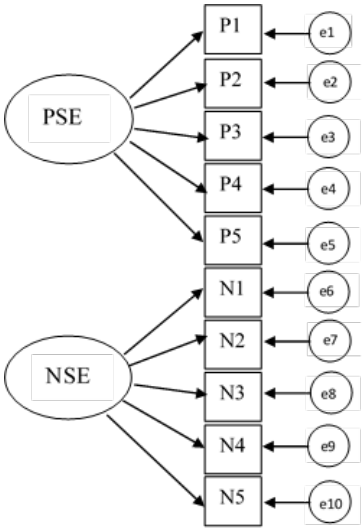

Figure 3. Model 3 Two orthogonal traits: Uncorrelated positive and negative self-esteem factors

2.3.3 Correlation with external criteria

Among the initial work examining the nature of item wording effects in a self-report survey instrument, researchers simply investigated the patterns of correlations 1) between the summed scores of the survey's positively worded items with the external variables and 2) between the

summed scores of the survey's negatively worded items with the external variables. With the use of this strategy, different patterns of correlations were expected because of item wording effects.

For instance, Marshall, Wortman, Kusulas, Hervig, and Vickers (1992) conducted the correlation between positively worded items with external variables as well as the correlation between negatively worded items with external variables. Positively worded items loaded on the factor named optimism and negatively worded items loaded on the other factor named pessimism. Marshall et al. (1992) discovered a stronger association between optimism and extroversion, and a stronger association between pessimism and neuroticism; therefore, Marshall et al. (1992) concluded that positively and negatively worded items actually measure separate constructs due to the different patterns of correlations.

In contrast, to evaluate the structure of the Rosenberg Self-Esteem Scale (RSES), Carmines and Zeller (1979) correlated each dimension of self-esteem (i.e., positive and negative self-esteem factors) with 16 external variables (criteria) falling into three areas: 1) socioeconomic background, 2) psychological predispositions, and 3) social and political attitudes. Due to negligible differences between correlations across all 16 variables (with largest difference being .05) and such nonsignificant difference (p >.25), Carmines and Zeller (1979) concluded that the dual dimensionality is a function of a single dimension of self-esteem contaminated by a method artifact.

2.3.4 CTCU

Numerous confirmatory factor models for separating the underlying construct and method variance have complemented the MTMM design (Marsh, 1989; Marsh & Bailey, 1991). The MTMM matrix is defined as a structured matrix of zero-order correlations between several traits

examined by several methods to assess validity evidence regarding relationships with conceptually related constructs (Campbell & Fiske, 1959). The three most frequently applied models are the correlated trait-correlated uniqueness (CTCU) model (Kenny, 1976; Marsh, 1989; Marsh & Bailey, 1991), the correlated trait-correlated method (CTCM) model (Widaman, 1985), and the CTCM minus one (CT-C[M-1]) model (Eid, Lischetzke, Nussbeck, & Trierweiler, 2003).

The CTCU (Kenny, 1976; Marsh, 1989) model suggests the presence of a single latent factor representing the construct of interest and correlated residual variances among the positively worded items and/or the negatively worded items (Vasconcelos-Raposo, Fernandes, Teixeira, & Bertelli, 2012). The CTCU model infers method effects from a series of correlated residuals among items using the same method; the CTCU model does not allow correlation between different method effects. See Model 4, 5, and 6 below.

Such models treat the wording effect as a methodological artifact only. With all covariances related to wording effects modeled, it assumes method effects to be non-unidimensional and rarely produces ill-defined solutions. The process of correlating residual terms has been heavily criticized by various authors (e.g., Brown, 2006); residual variances should not be correlated for the purpose of model fit because such correlation produces an additional unspecified latent construct, which would result with interpretation and replication problems.

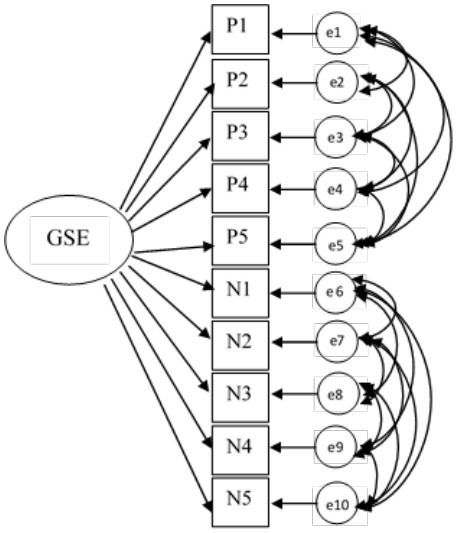

Figure 4. Model 4 One trait factor with correlated residuals among both positively and negatively worded items

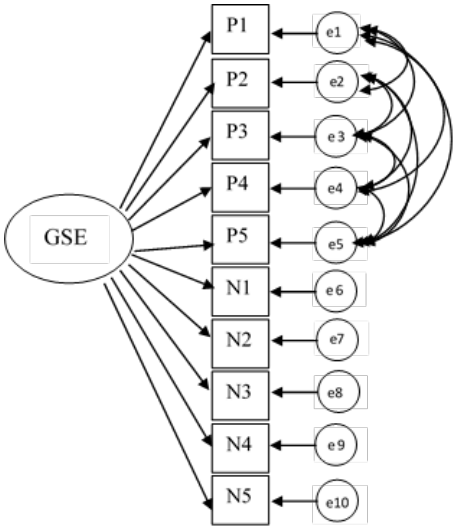

Figure 5. Model 5 One trait factor with correlated residuals among positively worded items

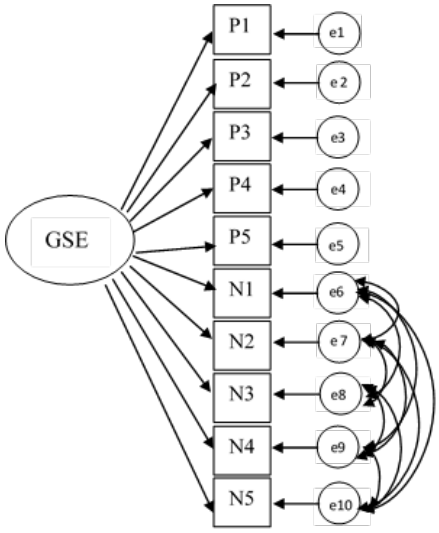

Figure 6. Model 6 One trait factor with correlated residuals among negatively worded items

2.3.5 CTCM

See Model 7 below for the CFA with correlated trait-correlated method (CFA-CTCM) (also called a general CFA model or block-diagonal model) (Tomas & Oliver, 1999). When methods are orthogonal, the reduced model is named as the correlated trait-uncorrelated method (CTUM) model; see Model 8. The CTCM (Widaman, 1985) model includes specific latent method effect factors underlying scale items of the same wording (i.e., positively or negatively worded items) along with a latent substantive factor. Such model decomposes observed variance into trait, method, and residual effects. The CTCM permits correlations between different method effects. However, the CTCM model suffers more from identification and estimation problems (Marsh, 1989; Marsh & Grayson, 1995).

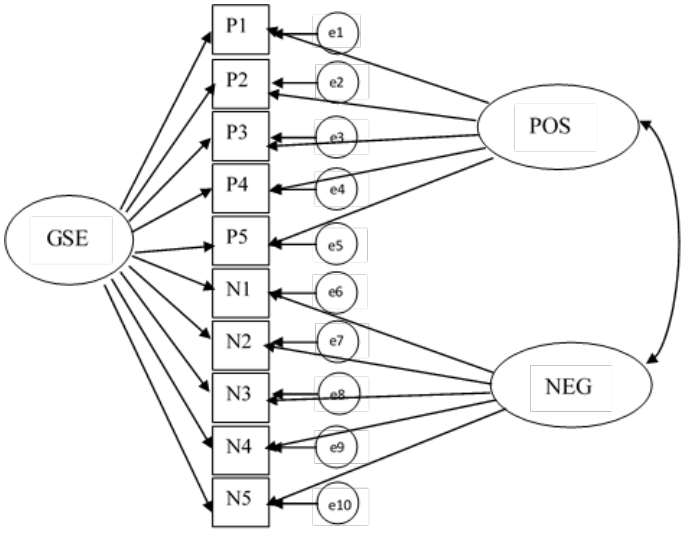

Figure 7. Model 7 One trait factor plus correlated positive and negative latent method factors

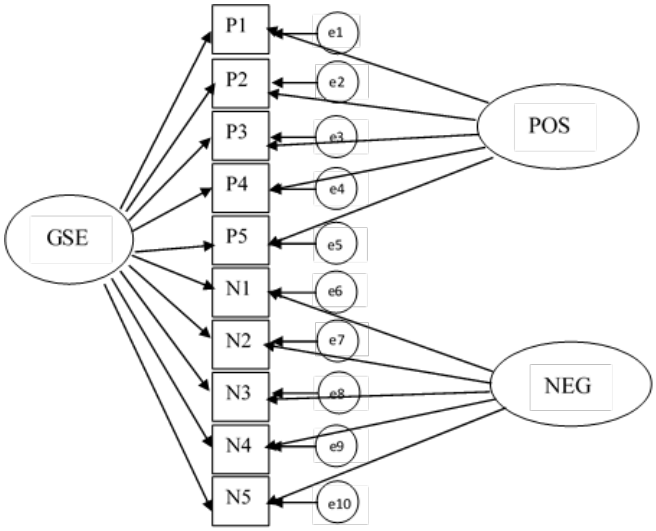

Figure 8. Model 8 One trait factor plus positive and negative latent method factors (uncorrelated method factors)

2.3.6 CTCU versus CTCM

The CTCU and CTCM models have led to debates concerning their wording effect. Morin, Arens, and Marsh (2016) and Schweizer (2012) claimed that it is inappropriate to employ CTCU models to statistically control for wording effect. CTCU models partial out the wording effect, prohibiting the addition of new information to the model and therefore, it is impossible to

investigate the nature of the wording effect (Lance, Noble, & Scullen, 2002). If using correlated residuals to represent the method effect, the internal-consistency coefficients would be significantly underestimated because the random error confounds the wording effect. CTCU models would biasedly estimate trait factor loadings when the method factor loadings are medium or high (Conway, Lievens, Scullen, & Lance, 2004; Urbán, Szigeti, Kökönyei, & Demetrovics, 2014). Lance et al. (2002) suggested the use of CTCM model over CTCU; the CTCU model should be employed only when the CTCM model fails. The CTCM model is favored by some researchers (e.g., DiStefano & Motl, 2006; Horan et al., 2003) because it assesses the method effect as a unique factor that allows for the empirical examination of the substantive relevant relations with external variables.

However, some researchers (e.g., Tomas, Hontangas, & Oliver, 2000) demonstrated the lack of methodological evidence for selecting one model over the other. CTCM and CTCU models have different underlying rationales and should not be used interchangeably (Tomas & Oliver, 1999). These two models are in fact operationally equivalent when wording effects are orthogonal (Bagozzi, 1993) and when limited to three items load on a method effect (Quilty et al., 2006). When the number of items is more than three, these two models can be examined and compared. The CTCU model can handle a method effect regardless of its dimensionality, whereas the CTCM model limits to unidimensional method effects (Tomas & Oliver, 1999). A better fitting CTCU model may indicate multidimensionality of the method effect.

In sum, when multidimensional method effects are present, the CTCU model is the appropriate choice. When method factors are correlated, the CTCM model needs to be adequately applied. When method factors are orthogonal and unidimensional, both CTCU and CTCM perform well. Unfortunately, whether method effects are multidimensional and/or

whether method factors are correlated is unknown before applying analytical tools. Therefore, an a priori preference between the CTCU and CTCM is not justified, unless in a replication study (Byrne, 1993; Marsh & Grayson, 1995; Tomas & Oliver, 1999).

2.3.7 CT-C(M-1)

In essence, the CT-C(M-1) (Eid, 2000; Eid et al., 2003) specifies one fewer method factors than the actual number of methods the MTMM strategies use; see Model 9 and 10. The CT-C(M-1) model, by selecting one of the methods to function as a comparison method standard, allows for an examination of convergent evidence regarding method by contrasting one method against the other. Like alternative MTMM strategies, the CT-C(M-1) also has weaknesses (e.g., Eid et al., 2003; Lance et al., 2002). For example, further examination is needed on which method effect (positive or negative) should be used as the reference method (the one not modeled).

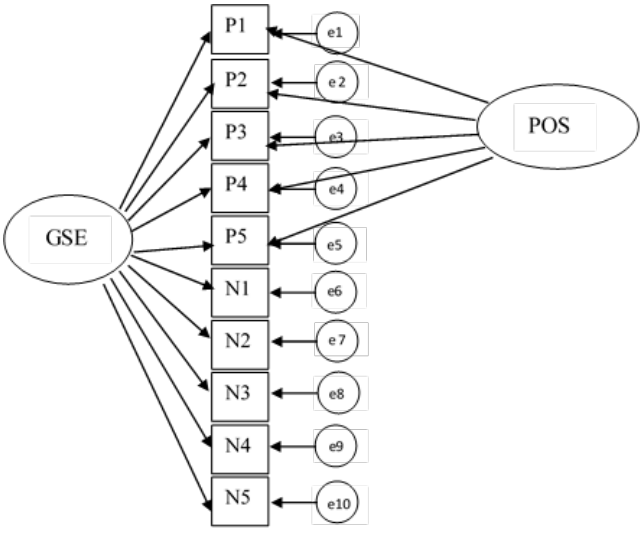

Figure 9. Model 9 One trait factor plus a positive latent method factor

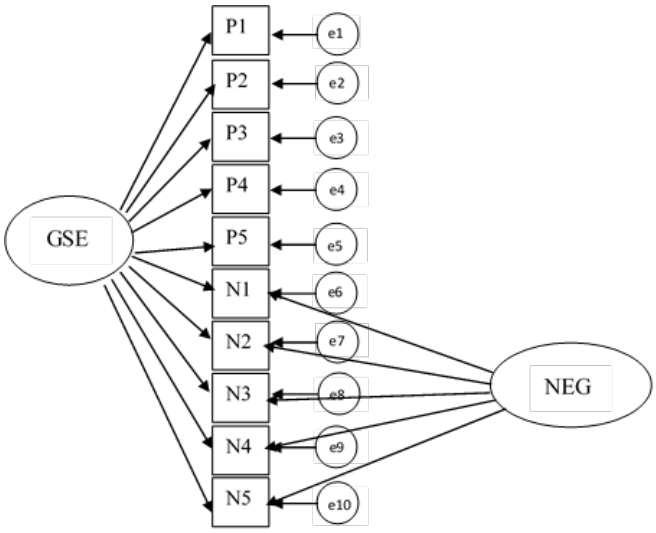

Figure 10. Model 10 One trait factor plus a negative latent method factor

2.3.8 Bi-factor model

Many psychometric experts (e.g., Myers, Martin, Ntoumanis, Celimli, & Bartholomew, 2014) have highlighted the effectiveness of using a bi-factor model to assess the structure of multidimensional scales (Holzinger & Swineford, 1937). A bi-factor model is specified to include one general factor directly influenced by all items on a measure and one or more specific factors that are directly influenced by subset(s) of items; the paths of these two influences occur simultaneously (Reise, 2012). From this perspective, a bi-factor model permits researchers to simultaneously explore the validity evidence for criterion relationships of both the general factor(s) and the specific factors.

The canonical bi-factor model (also termed as “restricted bi-factor model”) sets all correlations between the general and specific factors as zero (Chen, West, & Sousa, 2006); see Model 8, 9, and 10 as special cases. The oblique bi-factor model (Jennrich & Bentler, 2012) relaxes the assumption of orthogonality; see Model 7 as a special case.

Recent research has emerged to address the wording effect using a form of the bi-factor model which is the same as CTUM and CT-C(M-1) models. The bi-factor model considers the common variance shared by all items (i.e., for the target trait), and one or two specific factors for method variance in terms of the systematic variance from the positively and negatively worded items (e.g., Reise, Morizot, & Hays, 2007; Vecchione, Alessandri, Caprara, & Tisak, 2014). The bi-factor model permits the investigation of whether items on a measure are sufficiently unidimensional to allow for the interpretation of its scores (Reise, 2012). With a suitably modeled wording effect, researchers can determine the impact of a wording effect on the psychometric traits (e.g., internal-consistency coefficients, the validity evidence for criterion relationships and the internal structure of the measure) of the measure (see, Reise, 2012 for details concerning bi-factor model applications and associated methodology issues). Across a variety of measures, researchers demonstrated that the bi-factor model provides a good fit to data.

Research in bi-factor modelling (Reise et al., 2013; Rodriguez et al., 2016) stresses the use of explained common variance (ECV) and other statistics (e.g., coefficient omega and omega hierarchical) for interpretation of general and specific factors, which can be applied to bi-factor models for method effects. These statistics are valuable for evaluating the estimation of internal-consistency coefficients, the validity evidence for criterion relationships and the internal structure of the measure.

ECV is the ratio of variance attributable to a general factor and variance attributable to general and specific factors. The ECV is computed as:

Equation 1

$$\text{ECV} = \frac{\left(\sum \lambda_{gi}^2\right)}{\left(\sum \lambda_{gi}^2\right)+\left(\sum \lambda_{sj}^2\right)},$$

where λ_{gi} is the factor loading onto the general factor and λ_{sj} is the factor loading onto the specific factor. The ECV estimates the relative strength of the general factor to the specific factor(s). The ECV can be varied by changing number of positively and negatively worded items and their factor loadings. A higher value of ECV means a stronger general factor relative to the specific factor, then less wording effects. Researchers (Gu et al., 2017) concluded that when ECV is high (e.g., >.75), the use of unidimensional model is sufficient. When ECV less than .75, it is important to control for wording effects. The ECV has a negative correlation with the bias in estimates of the validity evidence for criterion relationships (Gu et al., 2017; Reise et al., 2013).

Omega indices are used to disentangle the variance explained by general or specific factors. Coefficient omega (also named as internal consistency reliability or composite reliability) is computed as:

Equation 2

$$\text{omega} = \frac{\left(\sum \lambda_{gi}\right)^2+\left(\sum \lambda_{sj}\right)^2}{var(total)},$$

where var(total) is the total variance. Omega is an estimator of variance attributed to both the general and specific factors.

Omega hierarchical (also named as homogeneity coefficient) is computed as:

Equation 3

$$omega\text{H} = \frac{\left(\sum \lambda_{gi}\right)^2}{var(total)}.$$

Omega hierarchical is an estimator of variance attributed to the general factor. This index reflects the degree of unidimensionality. Omega hierarchical treats the method effect as measurement error and the square root of omega hierarchical refers to the correlation between the general trait factor and the observed total score. Misspecifying the model for the negative wording effect underestimates the coefficient omega but overestimates the omega hierarchical (Gu et al., 2017).

2.4 SEM FIT INDICES

All the aforementioned models addressing method effects are applications of structural equation modeling (SEM), and as a result, model fit indices are commonly used to compare these models to identify the optimal model for factor structure of the construct of interest. Widely used model fit indices include χ^2 goodness-of-fit test statistic, Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), root mean square error of approximation (RMSEA), and standardized root mean-square residual (SRMR). In addition, Akaike's information criterion (AIC; Akaike, 1974), Bayesian information criterion (BIC; Schwarz, 1978) and sample size-adjusted BIC (SABIC) are popular indices used for comparing non-nested models.

Researchers can use a χ^2 goodness-of-fit test statistic to test the null hypothesis that the specified model leads to an approximate representation of the observed data when the model was specified correctly and the distributional assumptions for the data were satisfied. A non-significant test statistic indicates a fitting model. The χ^2 statistic is sensitive to sample size and

probably overestimates the model misfit (Bollen, 1989). Therefore, researchers suggest to use a variety of indices from different families of measures to supplement the utilization of the χ^2 statistic. The Tucker-Lewis Index (TLI) and Comparative Fit Index (CFI) are incremental fit indices; a larger value is an indicator of a better fit. The TLI, also named the non-normed fit index (NNFI), compares the lack of fit of the proposed model to the lack of fit of the null model. TLI is not significant dependent on sample size. The CFI notates the relative reduction in lack of fit, as estimated by the noncentral chi-square of a proposed model versus a null model. The TLI and CFI differ primarily in that the TLI compensates for the effect of model complexity; its penalty for complexity is the ratio of chi-square and degree of freedom (Marsh, 1996). The root mean square error of approximation (RMSEA) and standardized root mean-square residual (SRMR) are absolute fit indices. The RMSEA and SRMR measure absolute fit of the data to the model; a smaller value is an indicator of a better fit. As Hu and Bentler (1999) suggested, CFI equals to or greater than .95, RMSEA equals to or less than .06, and SRMR equals to or less than .08 indicate a good fit between hypothesized model and the data. See Table 1 for the formulas and descriptions associated with different indices.

Table 1. Formulas and descriptions for some selected incremental and absolute fit indices

	Formula	Description
Incremental Fit Indices	CFI = 1- max[($T_{proposed}$ - $df_{proposed}$), 0]/max[($T_{proposed}$ - $df_{proposed}$), (T_{null} - df_{null}), 0]	Normed. Noncentrality-based.
	TLI (or NNFI) = [(T_{null}/df_{null}) –	Non-normed. Compensates for the

Table 1 continued

	$(T_{proposed}/df_{proposed})]/\ [(T_{null}/df_{null}) - 1]$	effect of model complexity.
Absolute Fit Indices	RMSEA = $\sqrt{\hat{F}_0/df_{proposed}}$, where $\hat{F}_0$ = max[($T_{proposed}$ - $df_{proposed}$)/(N-1), 0]	Has a known distribution. Compensates for the effect of model complexity. Noncentrality-based.
	SRMR = $\sqrt{\left\{2\sum_{i=1}^{p}\sum_{j=1}^{i}\left[\left(s_{ij} - \hat{\sigma}_{ij}\right)/s_{ii}s_{jj}\right]^2\right\}/p(p+1)}$	Standardized root mean squared residual

Note. $T_{proposed}$ = T statistic for the proposed model. $df_{proposed}$ = degrees of freedom for the proposed model. T_{null} = T statistic for the null model. df_{null} = degrees of freedom for the null model. p = number of observed variables. s_{ij} = observed covariances. $\hat{\sigma}_{ij}$ = reproduced covariances. s_{ii} and s_{jj} are the observed standard deviations. CFI = Comparative Fit Index. TLI = Tucker-Lewis Index. NNFI = non-normed fit index. RMSEA = root mean square error of approximation. SRMR = standardized root mean-square residual.

Akaike's information criterion (AIC; Akaike, 1974), Bayesian information criterion (BIC; Schwarz, 1978), and sample size-adjusted BIC (SABIC) are formulated as the sum of negative log-likelihood and a penalty term that increases with the number of parameters in a given model. The negative log-likelihood represents the goodness of fit of a proposed model with a smaller value indicating a better fit. The penalty term shows the complexity of a model and the smaller it is, the more parsimonious the model is. SABIC, like the BIC, includes the penalty for adding parameters based on sample size, but less penalty than the BIC. Thus, a model

with a minimal value of AIC, BIC, or SABIC among all the competing models indicates an optimal balance between model fit and model complexity and is the prefered model. If the models are with different values of model-fit criteria, the model with the smaller value is favored. Otherwise, the more parsimonious model is favored. Ideally, different fit indices will point to the same conclusion. If fit indices lead to different conclusions, the conservative choice is to reject the model.

Oftentimes, empirical researchers determine the final structure mainly based upon fit indices (e.g., CFI, RMSEA, SRMR). The majority of studies found that the bi-factor model was better than alternative models in model fit. However, Donnellan et al. (2016) claimed that model fit indices did not perform well in model selection when the underlying true structure was unknown. They used TLI, CFI, RMSEA along with a 90% confidence interval (CI) for RMSEA, SRMR, AIC, and BIC to compare and evaluate nine models on the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965). Based on consistent general factor loadings across various models, Donnellan et al. (2016) suggested that the validity evidence for criterion relationships of the general self-esteem factor seemed not to be affected when wording effects were not controlled for. They further concluded that the study of the factor structure of the RSES does not have significant impact on the practical implications of the RSES. This statement was also supported by Michaelides, Koutsogiorgi, and Panayiotou (2016b).

In addition to empirical analyses, one simulation study (Morgan, Hodge, Wells, & Watkins, 2015) compared the fit (using CFI, TLI, SRMR, RMSEA, AIC, BIC, and aBIC) of correlated factors and bi-factor models. Morgan et al. (2015) specified four design factors including 1) three true models, 2) three fitted models, 3) two sample sizes, and 4) two factor identifications. They argued that when the true underlying model was a bi-factor model, model

fit indices tended to correctly select the true model over misspecified correlated factor models. However, when the true underlying model was a correlated factor model, model fit indices biasedly favored a bi-factor model under certain conditions.

Moreover, Reise et al. (2013) found that model fit indices (CFI, RMSEA and SRMR) did not effectively detect the model-misfit between unidimensional and "strictly" bi-factor models under different 1) relative strength of general and specific factor loadings, 2) number of specific factors, and 3) number of items. In contrast, factor strength indices, including the explained common variance (ECV) and omega hierarchical (*omega*H) had substantive impact on the bias of the validity evidence for criterion relationships. The ECV was found to negatively correlate with the bias of the validity evidence for criterion relationships.

Gu et al. (2017) concluded that the fit indices (CFI and RMSEA) performed unsatisfactorily in selecting models between a true bi-factor model (with one specific factor for a negative wording effect) and a misspecified unidimensional model. They manipulated four factors in their simulation: 1) the number of positively and negatively worded items, 2) loadings in accordance with the trait and the wording effect factors, 3) sample size, and 4) the relation of the measure to a relevant criterion. Results also suggested the use of ECV, coefficient omega, and coefficient omega hierarchical for selecting the analysis model between bi-factor and unidimensional models. As the ECV increases, the statistical power for detecting the validity evidence for criterion relationships increases. The contamination of spurious wording effect underestimated the coefficient omega and the relation of the measure to the criterion, but overestimated the coefficient omega hierarchical.

2.5 ARE WORDING EFFECTS IN THE RSES SUBSTANTIVE OR ARTIFACTUAL?

This section reviews research on the factor structure of the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965) as this scale was used to illustrate the design for the Monte Carlo study conducted in the present study. This scale was used due to its popular use and the ongoing debate upon the factor structure. See Appendix A for a review of the items. The dominance of Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965) in self-esteem research is reflected in its translation into 28 different languages across 53 countries and in its ability to perform well in different settings (Schmitt & Allik, 2005).

The search for primary studies reporting on the factor structure of the RSES included major scientific databases (ERIC, PsycINFO) and Google Scholar. Additional studies were derived from the references of all identified articles using a rolling snowball method. In July 2018, after reviewing the titles and the abstracts, a total of 84 articles were retained. Eligible studies supported either an oblique/orthogonal two-factor solution (i.e., correlated or uncorrelated positive self-esteem and negative self-esteem) or a global self-esteem factor (with method effects). Studies used for obtaining descriptive statistics in Table 4 met the following two additional criteria: 1) model fit indices were used for model comparison/selection and 2) factor loadings of their final models were reported.

Exploratory and confirmatory factor analyses, using different versions of the RSES (6, 7 and 10 items), have reported that a single factor sufficiently accounts for significant variance in the RSES, supporting a unidimensional structure of self-esteem (Bagley et al., 1997; Gray-Little, Williams, & Hancock, 1997; Pullmann & Allik, 2000; Robins, Hendin, & Trzesniewski, 2001; Shevlin et al., 1995; Tomas & Oliver, 1999). However, researchers have also argued different facets of self-esteem underlying the RSES (Dobson, Goudy, Keith, & Powers, 1979; Hensley &

Roberts, 1976; Owens, 1993, 1994). For example, Hensley and Roberts (1976) scored the 10 items across a five-point response framework and employed a varimax rotation. They found a two-factor solution with all item loadings between .53 and .71 on the positive self-esteem factor and between .52 and .71 on the negative self-esteem factor.

2.5.1 Positive and negative self-esteem

A number of factor analytic studies produced findings that support a dual dimensionality of the RSES (Ang et al., 2006; Boduszek et al., 2013; Boduszek et al., 2012; Greenberger, Chen, Dmitrieva, & Farruggia, 2003; Hensley & Roberts, 1976; Owens, 1993; Supple, Su, Plunkett, Peterson, & Bush, 2013). Researchers suggested to interpret the positive component as positive self-worth; the reflection of the degree to which one believes in one's own capacities or worth. They interpreted the negative component as self-deprecation; the reflection of the degree to which one underestimates self-capacities or self-worth (Owens, 1994). In this regard, the positive component distinguishes from the negative component, though these two correlate with each other.

For example, Owens (1993) conducted EFA and CFA to examine the dimensionality of the RSES. He used a scale containing six positively worded items and four negatively worded items. Owens (1993) exploratory findings demonstrated a two factorial structure and he further supported a bi-dimensional model over a unidimensional model via assessing model fit and parameter estimates. The results showed that the unidimensional model had a poor fit to the data, while the dual dimensional model exhibited an adequate fit.

Another validation study of the RSES conducted by Ang et al. (2006) argued that if RSES measures two dimensions separated by positively and negatively worded items, these two

distinct factors should correlate with external variables differentially and substantively. They expected that the factor extracted from positively worded items significantly predicted mastery goal orientation and academic self-efficacy while the other factor extracted from negatively worded items significantly predicted disruptive behavior. In their study, a nine-item (five positively worded and four negatively worded) Rosenberg Self-Esteem Scale (RSES) was used with the item of "I wish I could have more respect for myself." excluded due to 20.9% nonresponse. Model comparison resulted in supporting a dual dimensional model as a better fitting model. The correlation between positive and negative self-esteem factors was .33, which indicates moderate amount of shared variance between the two factors, further supporting that two-factor model appeared to be adequate.

Moreover, positive self-esteem significantly predicted both students' mastery and self-efficacy but not disruptive behavior, negative self-esteem significantly predicted students' disruptive behavior but not students' mastery or self-efficacy. A bi-dimensional structure of the RSES was also favored by studies involving samples of prisoners/ex-prisoners (Boduszek et al., 2013; Boduszek et al., 2012). Boduszek and his colleagues favored a bi-dimensional model over a one-factor model via model comparison and external criterion verification.

2.5.2 One substantive self-esteem

If the RSES is indeed a bi-dimensional scale then that would mean that each dimension would have to be scored separately and each dimension would require psychometric evaluation. However, there is no clear answer to the nature of the RSES and the two-factor solution poses challenges to the initial conceptualization of the RSES. According to a meta-analysis based upon 23 factor analytic studies of the RSES, two factors were generated to explain 93.7% of the

variance. The low discriminant validity evidence between positive and negative self-esteem factors, however, was indicative of the appropriateness of the single-factor solution (Huang & Dong, 2011).

Many researchers have realized that the unidimensional model may be overly simplistic. Models with method effects, either in CTCU or CTCM or both models, outperformed the competing models without method effects (Corwyn, 2000; Horan et al., 2003; Marsh, 1996; Marsh et al., 2010; Tomas & Oliver, 1999). This implies that the RSES is contaminated with method effects. Moreover, some researchers claim that method effects are mainly attributable to negatively worded items (Corwyn, 2000; Donnellan et al., 2016; Horan et al., 2003; Marsh, 1996; Marsh et al., 2010; Tomas & Oliver, 1999; Wang, Kong, Huang, & Liu, 2016) while several researchers (Gana et al., 2013; Lindwall et al., 2012; Salerno, Ingoglia, & Lo Coco, 2017; Wang, Siegal, Falck, & Carlson, 2001) claim that method effects are mainly associated with positively worded items. Other studies have demonstrated that models including method effects for both positively and negatively worded items reach an optimal fit (Marsh et al., 2010; Quilty et al., 2006; Wu, Zuo, Wen, & Yan, 2017).

The wording effect was interpreted as response style or enduring individual characteristics in many studies (Gana et al., 2013; Horan et al., 2003; Lindwall et al., 2012; Marsh, 1996; Marsh et al., 2010; Michaelides, Koutsogiorgi, & Panayiotou, 2016a; Quilty et al., 2006; Urbán et al., 2014). In contrast, Alessandri and his colleagues (Alessandri, Vecchione, Donnellan, & Tisak, 2013; Alessandri et al., 2015) argued for the substantive interpretation of the two specific factors in bi-factor modeling. They considered the negatively worded items of the RSES as ‘self-derogation’ which is a reflection of intense negative affect toward the self and they considered the positively worded items of the RSES as ‘self-competence’ which mirrors

individual's self-appraisal of his or her competences. Their interpretation of the two specific factors aligns to the interpretation in the two-factor solution.

A recent meta-analysis of the RSES conducted by Gnambs et al. (2018) supported a bi-factor model with two specific factors related to positive and negative wording. An initial EFA resulted in a two-factor solution: five positively worded items had salient loadings (between .51 and .75) on the positive self-esteem factor and five negatively worded items had salient loadings (between .45 and .80) on the negative self-esteem factor. These two extracted factors were correlated at .68, indicating the covariances of the RSES items were attributable to a common factor. Further, multiple model fit indices including chi-square, CFI, TLI, RMSEA, SRMR, AIC, and BIC were assessed for model comparison. Three models provided acceptable but marginally inferior model fit compared to the bi-factor model with two specific factors. These three models were 1) the oblique two-factor (i.e., correlated positive and negative self-esteem) model, 2) the bi-factor model with a positive specific factor, and 3) the bi-factor model with a negative specific factor.

In a bi-factor model with two specific factors, the positive specific factor exhibited only a single substantial loading larger than .40 and two loadings exhibited negative values which were close to zero. In the oblique two-factor model, positively worded items loaded on the positive self-esteem factor ranging from .56 to .76, negatively worded items loaded on the negative self-esteem factor ranging from .54 to .74, and the inter-correlation between factors was .79. In the bi-factor model with a negative specific factor, all item loadings on the general factor were greater than .40 (ranging from .43 to .76), negatively-worded items' loadings on the negative specific factor ranged from .29 to .55; only one negatively-worded item loaded on the negative specific factor (.55) marginally higher than on the general factor (.54).

Researchers (2018) concluded that the RSES essentially represents a unidimensional scale because most of the explained common variance in the RSES (up to 85%) was captured by the general factor. Gnambs et al. (2018)'s findings were consistent to other researchers' conclusion that the structure of one general self-esteem factor with two specific factors was the best-fitting solution among alternative models (Alessandri et al., 2015; Lindwall et al., 2012; Marsh et al., 2010; Michaelides, Koutsogiorgi, et al., 2016a; Michaelides, Zenger, et al., 2016; Quilty et al., 2006).

3.0 METHODS

This chapter presents a Monte Carlo simulation study to evaluate the performance of model fit indices in identifying the correct specification of negatively worded items, the impact of misspecifying the model for the negative wording effect on the estimates of internal-consistency coefficients, the validity evidence for criterion relationships and the internal structure of the measure. Three research questions were answered.

Research Question 1: how well do model fit indices perform in identifying the correct model for negative wording effects?

Research Question 2: what are the effects of negative wording on the estimates of internal-consistency coefficients?

Research Question 3: what are the effects of negative wording on the validity evidence for criterion relationships and the internal structure of the measure?

Three data generation models are introduced first, followed by the simulation design and data validation.

3.1 DATA GENERATION

Prior studies have demonstrated that measures with potential wording effects introduced by positive versus negative wording in a self-report measure can be modeled jointly or separately in

the bi-factor model. In particular, it can be hypothesized that both the general factor and one or two specific factors together account for the items' covariation. In self-report measures, the general factor refers to the trait of interest that explains the common variance shared by all the items. The specific factor refers to the positive-wording factor or the negative-wording factor, or both. The positive-wording factor accounts for the method variance introduced by positively worded items. The negative-wording factor accounts for the method variance introduced by negatively worded items. When both positive-wording and negative-wording factors are included in bi-factor modeling, the two specific factors are assumed to be orthogonal with each other and uncorrelated with the general factor.

Though empirical studies have claimed the redundancy of incorporating both positive and negative wording effects in a bi-factor modeling, many studies still favored the bi-factor modeled with two wording effects. In addition, many studies concluded that the wording effect was primarily associated with negatively worded items. Hence, this dissertation included both variations of the bi-factor model for comparison.

For illustration purpose, a self-report measure such as the Rosenberg Self-Esteem Scale (RSES) was assumed to compose five positively worded items (x_1-x_5) and five negatively worded items (x_6-x_{10}). In a two-factor CFA model, items of (x_1-x_5) measure a positive trait factor (*P*) and items of (x_6-x_{10}) measure a distinct negative trait factor (*N*). These two factors (*P* and *N*) can be either correlated or uncorrelated. In a bi-factor model with two specific factors, items of (x_1-x_{10}) measure a general trait factor (*G*) and two specific method factors associated with the positive item wording (S_P) and the negative item wording (S_N), respectively. In a bi-

factor model with one specific factor, all items measure a general trait factor (G) and the specific factor is associated with the negative item wording (S_N). These three models were used in the current study. See the equations for each model below.

Correlated two-factor model:

Equation 4

$$x_i = \lambda_{pi}P + e_i, \text{ for } i = 1, \ldots, 5$$

$$x_i = \lambda_{ni}N + e_i, \text{ for } i = 6, \ldots, 10$$

Bi-factor with two specific factors:

Equation 5

$$x_i = \lambda_{gi}G + \lambda_{si}S_P + e_i, \text{ for } i = 1, \ldots, 5$$

$$x_i = \lambda_{gi}G + \lambda_{si}S_N + e_i, \text{ for } i = 6, \ldots, 10$$

Bi-factor with one specific factor:

Equation 6

$$x_i = \lambda_{gi}G + e_i, \text{ for } i = 1, \ldots, 5$$

$$x_i = \lambda_{gi}G + \lambda_{si}S_N + e_i, \text{ for } i = 6, \ldots, 10$$

where λ_{pi} is the factor loading of x_i on the positive trait factor (P), λ_{ni} is the factor loading on the negative trait factor (N), λ_{gi} is the factor loading of x_i on the general factor (G), λ_{si} is the factor loading on the positive specific factor (S_P) or the negative specific factor (S_N), and e_i is

the residual of x_i. These latent factors (two first-order factors in two-factor CFA, general and specific factors in bi-factor models) were assumed to be standard normal. Variances of the unique errors terms were computed based on the factor loadings so that the variance of each manifest variable will be unity.

For assessing the validity evidence for criterion relationships, a criterion variable was specified to regress on the general factor in bi-factor models and both the positive and negative factors in the two-factor model. The criterion variable is a normal variable with mean zero and residual variance to be calculated so that R^2 is .25. The simulated true two-factor model is shown in Figure 11, the simulated true bi-factor model with two specific factors for positive and negative wording effects is shown in Figure 12, and the simulated true bi-factor model with one specific factor for negative wording effect is shown in Figure 13.

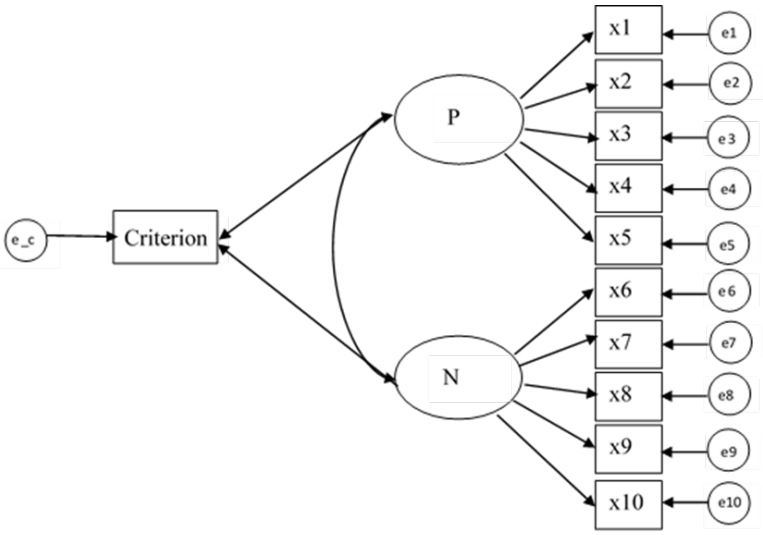

Figure 11. Simulated true two-factor model

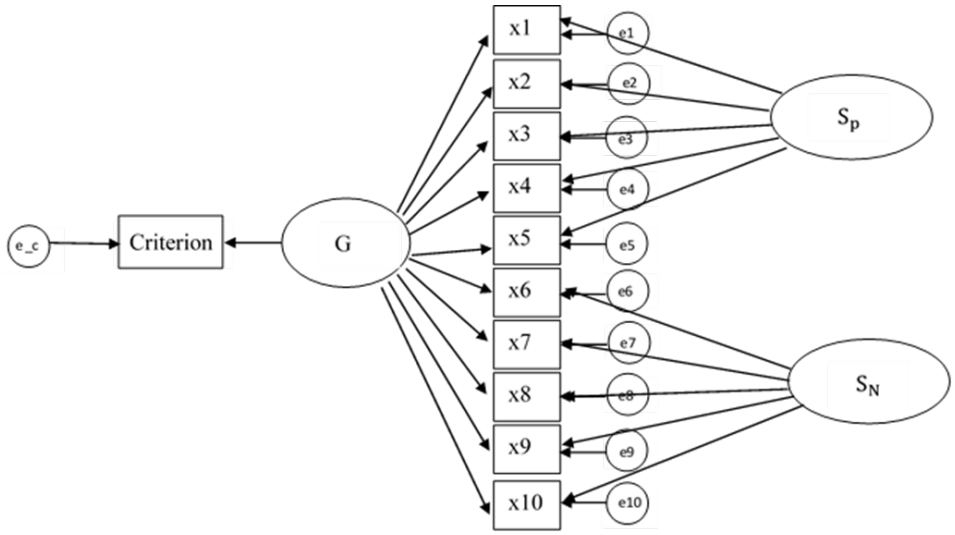

Figure 12. Simulated true bi-factor model with two specific factors

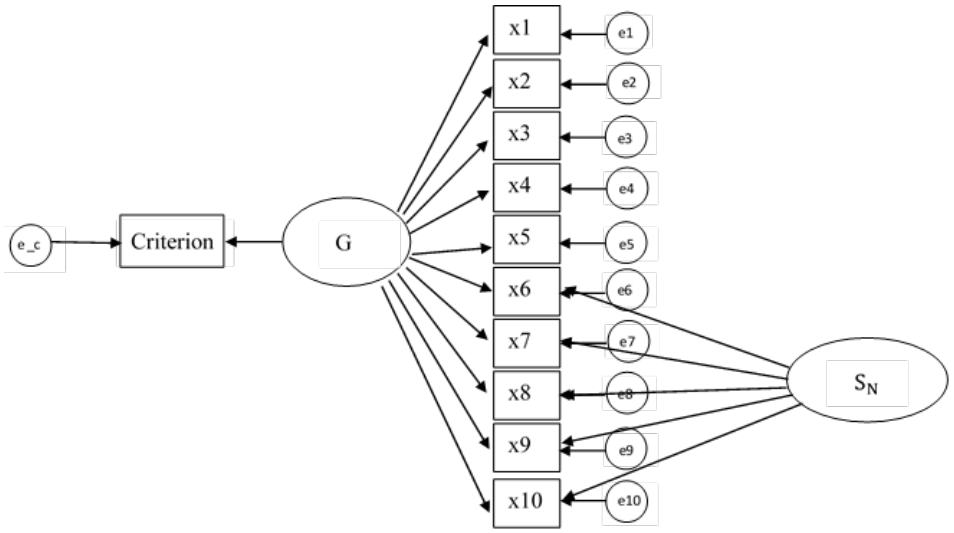

Figure 13. Simulated true bi-factor model with one specific factor for negative wording effect

Prior simulation studies (e.g., Gu et al., 2017) suggested that there is no difference in bias of internal-consistency coefficients and the validity evidence for criterion relationships for simulation conditions with different sample sizes. In this study, the sample size was constrained to be 1,000 to assure sufficient ability in the estimation of the model parameters.

3.2 SIMULATION DESIGN

The current study simulated three data generation models: 1) a two-factor CFA, 2) a bi-factor with two specific factors for positive and negative wording effects, and 3) a bi-factor with one specific factor for negative wording effect. Four design factors were manipulated for the data generation two-factor CFA model. These four design factors are 1) two levels for number of positively and negatively worded items (i.e., 5, 5 and 7, 3), 2) two levels for item loadings on positive and negative factors (i.e., .6, .6 and .6, .3), 3) three levels for criterion-related validity coefficient of positive and negative factors (i.e., 0, 0; .5, .5; and .5, .1), and 4) two levels for correlation between factors (i.e., .4 and .7). Three design factors were manipulated for the data generation bi-factor model with two specific factors for positive and negative wording effects. These three design factors are 1) two levels for number of positively and negatively worded items (i.e., 5, 5 and 7, 3), 2) five levels for item loadings on the general factor, the positive specific factor, the negative specific factor (i.e., .6, .6, .6; .6, .6, .3; .6, .3, .3; .3, .6, .6; and .3, .6, .3), and 3) two levels for criterion-related validity coefficient of the general factor (i.e., 0 and .5). Three design factors were manipulated for the data generation bi-factor model with one specific factor for negative wording effect. These three design factors are 1) two levels for number of positively and negatively worded items (i.e., 5, 5 and 7, 3), 2) three levels for item loadings on the general factor and the negative specific factor (i.e., .6, .6; .6, .3; and .3, .6), and 3) two levels for criterion-related validity coefficient of the general factor (i.e., 0 and .5).

Altogether, there were 24 (2 x 2 x 3 x 2 for the two-factor CFA) + 20 (2 x 5 x 2 for the bi-factor with two specific factors for positive and negative wording effects) + 12 (2 x 3 x 2 for the bi-factor with one specific factor for negative wording effect) = 56 unique cell conditions. For each cell, a thousand sample data sets were generated based on a set of specified population

parameters. Four models (including two-factor CFA, bi-factor model with two specific factors, bi-factor model with one specific factor for negative wording effect, and one-factor model) were fitted to each simulated sample data set. The one-factor model was fitted to serve as a useful point of comparison to evaluate whether the unidimensional model is sufficient under certain conditions for the purpose of obtaining validity evidence regarding relationships with criteria given data multidimensionality. SAS 9.4 and M*plus* 8.0 were used to generate and analyze the data. Maximum likelihood estimation was used. The rationales of design conditions and levels were described below. See Table 5 for varied design factors in the Monte Carlo study.

First, two combinations of positively and negatively worded items were manipulated. According to various self-report measures shown in Table 2, the total number of items in a self-report measure ranged from 8 to 16; the proportion of positively to negatively worded items was 1:1, 4:3, 3:2, and 11:5. Table 3 illustrates the number of positively and negatively worded items in prior simulation studies. For example, Gu et al. (2017) used 4 different combinations of positively and negatively worded items in their simulation study: 1) 6, 6, 2) 8, 4, 3) 9, 9, and 4) 12, 6. They only had a total of 12 or 18 items on the self-report measures and the proportion of positively to negatively worded items was either 1:1 or 2:1. This dissertation adopted two levels for number of positively and negatively worded items, balanced (i.e., 5 positively and 5 negatively worded items) and unbalanced (i.e., 7 positively and 3 negatively worded items), while constraining the total number of items to be 10.

Table 2. Number of positively and negatively worded items in selected self-report measures

Scale	N_t	N_p	N_n
The Life Orientation Test Revised (Scheier et al., 1994)	8	4	4
The Penn State Worry Questionnaire (Meyer et al.,	16	11	5

Table 2 continued

1990)			
The Rosenberg Self-Esteem (Rosenberg, 1965)	10	5	5
	7	4	3
	10	6	4
General Health Questionnaire-12 (Aguado et al., 2012)	12	6	6

Note. N_t= total number of items. N_p = number of positively worded items. N_n= number of negatively worded items.

Table 3. Number of positively and negatively worded items in simulation studies

Source	N_t	N_p	N_n
Gu et al. (2017)	12	6	6
		8	4
	18	9	9
		12	6
Wang et al. (2015)	11	6	5

Note. N_t= total number of items. N_p = number of positively worded items. N_n= number of negatively worded items.

Second, in the two-factor CFA, two combinations of item loadings on the positive trait factor and the negative trait factor (λ_p, λ_n) were specified as (.6, .6) and (.6, .3). In the bi-factor model with two specific factors for positive and negative wording effects, five combinations of item loadings on the general factor, the positive specific factor, and the negative specific factor (λ_g, λ_{sp}, λ_{sn}) were specified as (.6, .6, .6), (.6, .6, .3), (.6, .3, .3), (.3, .6, .6), and (.3, .6, .3). In the

bi-factor model with one specific factor for negative wording effect, three combinations of item loadings on the general factor and the negative specific factor (λ_g, λ_{sn}) were specified as (.6, .6), (.6, .3), and (.3, .6).

Any item with loading of less than .3 is not worth considering (Reise et al., 2013). The loading of .6 was specified to mimic the computation of the mean factor loadings ($\sim$ .60) from empirical studies on the Rosenberg Self-Esteem Scale (RSES) and an indicator with a loading greater than .6 is considered as a strong indicator. Descriptive statistics for factor loadings obtained from applied studies (favoring any of these three models: two factor CFA, bi-factor model with two specific factors, or bi-factor model with one specific factor for negative wording effect) are reported in Table 4. In the two-factor CFA model, mean item loadings on positive and negative trait factors are close to each other. In the two bi-factor models, the loadings of the positively worded items on the general factor are higher than those of the negatively worded items. In the bi-factor model with two specific factors, the average specific factor loadings related to negatively worded items is higher than the average specific factor loadings related to positively worded items. These results align to prior literature that negatively worded items may contaminate the construct of interest and wording effect is primarily associated with negatively worded items (Corwyn, 2000; Donnellan et al., 2016; Horan et al., 2003; Marsh, 1996; Marsh et al., 2010; Tomas & Oliver, 1999; Wang et al., 2016).

Table 4. Descriptive statistics for factor loadings in applied studies

Model	Factor Loadings	Mean (SD)	Min, Max
Two-factor CFA	Positive Trait Factor Loadings	.63 (.13)	.31, .81
	Negative Trait Factor Loadings	.64 (.15)	.16, .90

Table 4 continued

Bi-factor Model with Two Specific Factors for Positive and Negative Wording Effects	General Factor Loadings related to Positively Worded Items	.57 (.19)	.02, .89
	General Factor Loadings related to Negatively Worded Items	.49 (.18)	.02, .86
	Specific Factor Loadings related to Positively Worded Items	.44 (.22)	.06, .94
	Specific Factor Loadings related to Negatively Worded Items	.47 (.22)	.02, .91
Bi-factor Model with One Specific Factor for Negative Wording Effect	General Factor Loadings related to Positively Worded Items	.63 (.06)	.52, .78
	General Factor Loadings related to Negatively Worded Items	.51 (.09)	.33, .69
	Specific Factor Loadings related to Negatively Worded Items	.47 (.21)	.13, .74

In bi-factor models, when the general factor loading is lower than the specific factor loading, the interpretation of the general factor is questionable. However, such cases are still happening empirically. For instance, Corwyn (2000) selected the bi-factor model with two specific factors to represent the underlying structure of the Rosenberg Self-Esteem Scale based on model-fit indices. In his selected bi-factor model, for the sample of adults in a follow-up survey 30 months later: the general factor loadings related to positively worded items ranged from .14 to .39 while the specific factor loadings related to positively worded items ranged from .39 to .73; the general factor loadings related to negatively worded items ranged from .02 to .42

while the specific factor loadings related to negatively worded items ranged from .51 to .84 (p.369, Corwyn, 2000). All the ten items loaded on the general factor lower than on the specific factor. For the sample of National Longitudinal Survey of Youth (NLSY) in 1987: the general factor loadings related to positively worded items ranged from .19 to .71 while the specific factor loadings related to positively worded items ranged from .51 to .82; the general factor loadings related to negatively worded items ranged from .00 to .48 while the specific factor loadings related to negatively worded items ranged from .65 to .85. Two positively worded items loaded on the general factor a bit higher than on the specific factor and all negatively worded items loaded on the general factor lower than on the specific factor (p.371, Corwyn, 2000). For the sample of adolescents: the general factor loadings related to positively worded items ranged from .02 to .14 while the specific factor loadings related to positively worded items ranged from .64 to .87; the general factor loadings related to negatively worded items ranged from .05 to .64 while the specific factor loadings related to negatively worded items ranged from .34 to .91 (p.373, Corwyn, 2000). Only one out of ten items loaded on the general factor (.64) marginally higher than on the specific factor (.55). Gu et al. (2017) also specified item loadings on the general factor and the negative specific factor as .3 and .6, respectively, in their simulation study. In summary, items with general loadings lower than specific loadings are common in empirical studies, and the current study incorporates such scenario in the simulation conditions for bi-factor models.

Third, different levels of validity coefficient (also called structural path coefficient; Reise et al., 2013) for the effect of the target latent variable (i.e., positive and negative trait factors in the two-factor CFA, general factor in the two bi-factor models) were specified for assessing bias in the validity evidence for criterion relationships, power and type I error rates for the validity

evidence regarding relationships with criteria. The simulated criterion-related validity coefficient, in the two-factor CFA, was set to be 1) zero on the positive trait factor and zero on the negative trait factor, 2) .5 on the positive trait factor and .5 on the negative trait factor, and 3) .5 on the positive trait factor and .1 on the negative trait factor. The simulated criterion-related validity coefficient, in the bi-factor models, were 1) 0 and 2) .5, respectively. This dissertation used the same criterion-related validity coefficient of .5 as Reise et al. (2013) specified in their study. The specification of .1 on the negative trait factor was selected to represent a negligible effect of the negative trait factor on the criterion manifest variable.

Fourth, the inter-factor correlations in the two-factor CFA were specified to be 1) .4 (medium) and 2) .7 (high).

Table 5. Varied design factors in the Monte Carlo study

Data Generation Model	Item Loading	Criterion-related Validity Coefficient	Correlation between Factors
Two-factor CFA	Positive, Negative .6, .6 .6, .3	Positive, Negative 0, 0 .5, .5 .5, .1	Medium: .4 High: .7
Bi-factor with Positive and Negative Wording Effects	General, SpecificP, SpecificN .6, .6, .6 .6, .6, .3 .6, .3, .3 .3, .6, .6	0 .5	

Table 5 continued

	.3, .6, .3		
Bi-factor with a Negative Wording Effect	General, SpecificN .6, .6 .6, .3 .3, .6	0 .5	

3.3 EVALUATION CRITERIA

Four analyses were fitted to each generated data, including a two-factor CFA, a bi-factor model with two specific factors for positive and negative wording effects, a bi-factor model with one specific factor for negative wording effect, and a one-factor CFA. The unidimensional model was fitted in each data generating structure as a reference model to investigate the impact of wording effects on the validity evidence for criterion relationships because Donnellan et al. (2016) indicated that misspecifying the model for the negative wording effect seemed not to impact the validity evidence for criterion relationships in their empirical study. The number of nonconvergent or improper solutions were recorded. Only sample data sets with proper solutions were used in evaluating goodness of model fit, estimation of internal-consistency coefficients, the validity evidence for criterion relationships and the internal structure of the measure. The following criteria were used.

3.3.1 Model fit indices

Model fit indices of chi-square, CFI, TLI, RMSEA, SRMR, AIC, BIC, and SABIC were used to compare the true and misspecified models. In addition to the non-significant chi-square, the criteria recommended by Hu and Bentler (1999) were used: CFI and TLI equal to or greater than .95, RMSEA equals to or less than .06, SRMR equals to or less than .08, and smaller AIC, BIC, and SABIC.

3.3.2 Pooled mean of factor loadings

Following Flora and Curran (2004), the pooled mean of the factor loading at each level was examined instead of examining the factor loading of each individual item:

Equation 7

$$\text{Pooled Mean} = n^{-1}\sum_{i=1}^{n}\overline{\hat{\lambda}_i},$$

where n is the number of indicators and $\overline{\hat{\lambda}_i}$ is the mean across replications of each factor loading for each factor. First, the mean of factor loadings across replications of each cell was calculated. Then the pooled mean of the factor loading of all items was calculated. For example, the pooled mean of the factor loading for the general factor was calculated across 10 items, while for the specific factor of negatively worded items, it was calculated across 5 or 3 items depending on the number of negatively worded items. Moreover, the means were calculated for general factor loading of positively worded items and negatively worded items separately, as well as the specific factor loadings.

The pooled standard deviation of the factor loading was calculated as:

Equation 8

$$\text{Pooled SD} = \sqrt{n^{-1}\sum_{i=1}^{n} VAR(\hat{\lambda}_i)},$$

where $VAR(\hat{\lambda}_i)$ is the sample variance of each factor loading across replications.

3.3.3 Bias in strength indices

The relative bias in strength indices (including ECV, composite reliability, and homogeneity coefficient) was computed only for two data generation bi-factor models and two data analysis bi-factor models. The relative bias of ECV was calculated by subtracting the true ECV from the average of ECV estimates in each condition and then dividing by the true ECV. For example, in the bi-factor model with two specific factors (five positively and five negatively worded items), when factor loadings were .6 on the general factor and .3 on both specific factors, the true ECV was .80. In the bi-factor model with one specific factor for negative wording effect (five negatively worded items), when factor loadings were .6 on the general factor and .3 on the specific factor, the true ECV was .89. See Equation 1 for the formula of ECV.

The relative bias of composite reliability was calculated by subtracting the true composite reliability from the average of composite reliability estimates in each condition and then dividing by the true composite reliability. For example, in the bi-factor model with two specific factors (five positively and five negatively worded items), when factor loadings were .6 on the general factor and .3 on both specific factors, the true composite reliability for the total score was .88. In the bi-factor model with one specific factor for negative wording effect (five negatively worded items), when factor loadings were .6 on the general factor and .3 on the specific factor, the true

composite reliability for the total score was .87. See Equation 2 for the formula of composite reliability.

The relative bias of homogeneity coefficient was calculated by subtracting the true homogeneity coefficient from the average of homogeneity coefficient estimates in each condition and then dividing by the true homogeneity coefficient. For example, in the bi-factor model with two specific factors (five positively and five negatively worded items), when factor loadings were .6 on the general factor and .3 on both specific factors, the true homogeneity coefficient was .78. In the bi-factor model with one specific factor for negative wording effect (five negatively worded items), when factor loadings were .6 on the general factor and .3 on the specific factor, the true homogeneity coefficient was .81. See Equation 3 for the formula of homogeneity coefficient. See Table 6 for the true ECV, composite reliability, and homogeneity coefficient for the bi-factor model with two specific factors and see Table 7 for the true ECV, composite reliability, and homogeneity coefficient for the bi-factor model with one specific factor. Relative bias less than 5% is the trivial bias, between 5% and 10% is the moderate bias, and greater than 10% is the substantial bias (Yang-Wallentin et al., 2010).

Table 6. True ECV, composite reliability, and homogeneity coefficient for the bi-factor model with two specific factors in various conditions

λ_g, λ_{sp}, λ_{sn}	N_p, N_n	True ECV	True ω	True ω_H
.6, .6, .6	5, 5	.50	.95	.63
	7, 3	.50	.95	.60
.6, .6, .3	5, 5	.62	.92	.70
	7, 3	.56	.94	.62
.6, .3, .3	5, 5	.80	.88	.78

Table 6 continued

	7, 3	.80	.88	.77
.3, .6, .6	5, 5	.20	.83	.28
	7, 3	.20	.84	.25
.3, .6, .3	5, 5	.29	.75	.33
	7, 3	.24	.81	.27

Note. λ_g = item loadings on the general factor. λ_{sp} = item loadings on the positive specific factor. λ_{sn} = item loadings on the negative specific factor. N_p = number of positively worded items. N_n = number of negatively worded items. ω = composite reliability coefficient. ω_H = homogeneity coefficient.

Table 7. True ECV, composite reliability, and homogeneity coefficient for the bi-factor model with one specific factor in various conditions

λ_g, λ_{sn}	N_p, N_n	True ECV	True ω	True ω_H
.6, .6	5, 5	.67	.91	.73
	7, 3	.77	.88	.81
.6, .3	5, 5	.89	.87	.81
	7, 3	.93	.86	.84
.3, .6	5, 5	.33	.71	.36
	7, 3	.45	.60	.44

Note. λ_g = item loadings on the general factor. λ_{sn} = item loadings on the negative specific factor. N_p = number of positively worded items. N_n = number of negatively worded items. ω = composite reliability coefficient. ω_H = homogeneity coefficient.

3.3.4 Bias of criterion-related validity coefficient

The relative bias of the validity evidence for criterion relationships was computed by subtracting the true criterion-related validity coefficient from the average of criterion-related validity estimates in each condition and then dividing by the true criterion-related validity coefficient. Relative bias less than 5% is the trivial bias, between 5% and 10% is the moderate bias, and greater than 10% is the substantial bias (Yang-Wallentin et al., 2010).

3.3.5 Power and type I error rates

Power in statistically detecting the validity evidence for criterion relationships was examined when the true criterion-related validity coefficient was nonzero for the true model and three misspecified models. Type I error rates is the percentage of the number of models with non-zero criterion-related validity coefficient over the total number of replications in each condition when the true criterion-related validity coefficient was zero.

3.4 VALIDATION OF DATA GENERATION

In the data validation part, data were generated using three data generation models, and analyzed with the corresponding true model only. Sample size was set to be 1000, with the number of replications set to be 500.

For the data generated for the two-factor CFA, factor loadings were set to be .6 on both factors. The criterion-related validity coefficient was set to be .5 on the positive trait factor and .1

on the negative trait factor. The average chi-square was 44.76 with *df* = 44, RMSEA was .006 (SD = .007), and SRMR was .02 (SD = .003). The average unstandardized factor loadings of the general factor ranged from .597 to .602 and the average was .600, same as the true value of .6. The average criterion-related validity coefficients were .498 and .103, close to the true value of .5 and .10.

For the data generated for the bi-factor model with specific factors for positively and negatively worded items, factor loadings were .6 on the general factor and .3 on both specific factors. The criterion-related validity coefficient was .5. The average chi-square was 34.79 with *df* = 35, RMSEA was .006 (SD = .008), and SRMR was .013 (SD = .002). The average unstandardized factor loadings of the general factor ranged from .597 to .603 and the average was .600, which was quite close to the true value of .6. The average unstandardized factor loadings of the specific factor ranged from .287 to .303 and the average was .297, which was quite close to the true value of .3. The average criterion-related validity coefficient was .499, close to the true value of .5.

For the data generated for the bi-factor model with the specific factor for negatively worded items, factor loadings were .6 on the general factor and .3 on the specific factor. The criterion-related validity coefficient was .5. The average chi-square was 40.60 with *df* = 40, RMSEA was .006 (SD = .007), and SRMR was .015 (SD = .002). The average unstandardized factor loadings of the general factor ranged from .598 to .600 and the average was .600, which was quite close to the true value of .6. The average unstandardized factor loadings of the specific factor ranged from .294 to .302 and the average was .299, which was quite close to the true value of .3. The average criterion-related validity coefficient was .500, close to the true value of .5.

4.0 RESULTS

This chapter presents the non-convergence percentage, the evaluation of model fit, parameter estimates, the estimation of internal-consistency coefficients, followed by the validity evidence for criterion relationships and the internal structure of the measure. First, the sample data sets that did not converge were removed. Second, the true and misspecified models in terms of model goodness of fit were compared. Model fit indices include chi-square, CFI, TLI, RMSEA, SRMR, AIC, BIC, and SABIC. Finally, pooled mean of factor loading, relative bias of composite reliability, relative bias of homogeneity coefficient, power in statistically detecting the validity evidence for criterion relationships (when the true criterion-related validity coefficient was nonzero), and Type I error rates (when the true criterion-related validity coefficient was zero) under the true and misspecified models were examined.

4.1 CONVERGENCE

All analyses using the unidimensional model for all conditions resulted in full convergence. All analyses in the two-factor CFA and bi-factor with negative wording effect for all conditions resulted in the percentages of convergence close to 100%. However, the convergence rate for the bi-factor model with two specific factors depended on the data generation model and the criterion-related validity coefficient. When the data generation model is a bi-factor model with

positive and negative wording effects, the analysis bi-factor model with positive and negative wording effects resulted in the percentages of convergence close to 100% only when the criterion-related validity coefficient was .5. When the criterion-related validity coefficient was 0, the percentages of convergence were around 80%. It seems that criterion-related validity of the general factor was related to convergence of the bi-factor model with two specific factors. For the other two data generation models, slight difference in the percentages of non-convergence was found across levels of the criterion-related validity coefficient. Specifically, when the generation model is a bi-factor with negative wording effect or a two-factor CFA, the percentages of non-convergence for bi-factor with positive and negative wording effects were around 20% at each level.

4.2 EVALUATION OF MODEL FIT

Model fit indices of chi-square, CFI, TLI, RMSEA, SRMR, AIC, BIC, and SABIC, were used to compare the true and misspecified models. In addition to the non-significant chi-square, the criteria recommended by Hu and Bentler (1999) were used: CFI and TLI equal to or greater than .95, RMSEA equals to or less than .06, SRMR equals to or less than .08. Percentage of each of these indices meeting the criteria for indicating good fit is discussed in terms of identifying the true model versus three misspecified models. For the information criteria, including AIC, BIC, and SABIC, the percentage of each index identifying the true model (i.e., smallest index across four analysis models) was computed. Appendix B presents these percentages by data generation model and simulation conditions. Appendix C presents percentage of non-significant chi-square, percentage of CFI and TLI equal to or greater than .95, percentage of RMSEA equals to or less

than .06, and percentage of SRMR equals to or less than .08 for the unidimensional model only. Results are summarized as follows.

4.2.1 Two-factor CFA

When the true underlying model was a two-factor CFA, percentages of non-significant chi-square for the unidimensional model were greater than 80% across conditions when 1) the factor loadings on the positive trait and negative trait factors were .6 and .3, respectively, 2) the number of positively and negatively worded items was 7 and 3, respectively, 3) the criterion-related validity coefficient of positive and negative factor was .5 and .1, respectively, and 4) the correlation between factors was .7. In addition, percentages of non-significant chi-square for the true model and the two bi-factor models were greater than 90% across all conditions. Therefore, chi-square did not function well in correctly identifying the true model; chi-square tended to favor the bi-factor model with positive and negative wording effects more frequently.

Almost 100% of CFI, TLI, RMSEA, and SRMR meeting the criteria for good fit across the analysis two-factor CFA model and the two analysis bi-factor models, indicating none of these indices worked correctly identifying the true model. When the unidimensional model was fitted, percentages of CFI and TLI were close to 100% across conditions when 1) the factor loadings on the positive trait and negative trait factors were .6 and .3, respectively, and 2) the inter-factor correlation was .7 in a balanced scale, or 2) in an unbalanced scale. Percentages of RMSEA in the analysis unidimensional model were close to 100% across conditions when 1) the factor loadings on the positive trait and negative trait factors were .6 and .3, respectively, or 1) the factor loadings on both positive trait and negative trait factors were .6 and 2) the inter-factor correlation was .7 in an unbalanced scale. Moreover, most conditions in the analysis

unidimensional model had 100% of SRMR indicating good fit except under conditions when 1) the factor loadings on both positive trait and negative trait factors were .6 and 2) the inter-factor correlation was .4 in a balanced scale.

In addition, the information criteria AIC, BIC, and SABIC performed well in identifying the true model when the number of positively and negatively worded items was balanced (i.e., 5, 5), but poorly in identifying the true model when the number of positively and negatively worded items was unbalanced (i.e., 7, 3). The percentage of AIC correctly selecting the data generation model was at least 80% under conditions wherein the number of positively and negatively worded items was balanced while approximately 10% under conditions wherein the number of positively and negatively worded items was unbalanced. Likewise, the percentage of BIC correctly selecting the data generation model was at least 76% under conditions wherein the number of positively and negatively worded items was balanced while approaching zero under conditions wherein the number of positively and negatively worded items was unbalanced. The percentage of SABIC was at least 95% under conditions wherein the number of positively and negatively worded items was balanced while around 2% under conditions wherein the number of positively and negatively worded items was unbalanced.

If one of these information criteria has to be chosen for identifying the true model for a balanced scale, SABIC would be selected since all its percentages were above 95%, followed by BIC. The percentages of BIC were 100% except for three conditions wherein the inter-factor correlation was .7 and the factor loadings on the positive trait and negative trait factors were .6 and .3, respectively. These three conditions had the percentage around 80%. If one of these information criteria has to be chosen for identifying the true model for an unbalanced scale, AIC would be selected as its percentage was highest, followed by SABIC, then BIC.

4.2.2 Bi-factor with positive and negative wording effects

When the true underlying model was a bi-factor model with positive and negative wording effects, 100% chi-square statistics were significant in the analysis unidimensional model, indicating the unidimensional model was identified as a poor fit. At least 93% non-significant chi-square statistics identified the true model and the two-factor CFA, indicating good fit of the models. The percentage of non-significant chi-square was slightly higher in the true model than that in the two-factor CFA, except when 1) the item loadings on the general factor, the positive specific factor, and the negative specific factor were .3, .6, and .3, respectively, and 2) the criterion-related validity coefficient was .5 in a balanced scale. When the bi-factor model with negative wording effect was fitted, the percentage of non-significant chi-square was low when the criterion-related validity coefficient was .5, indicating a poor fit. Therefore, chi-square statistics identified the true model most frequently, followed by the two-factor CFA. Clearly, based upon 100% significant chi-square statistics, the analysis unidimensional model was identified as a model with unacceptable fit across all conditions.

Other approximate indices did not work in correctly selecting the true model versus misspecified models. In particular, CFI, TLI, RMSEA, and SRMR identified two-factor CFA and both bi-factor models as those with good fit because percentages of each index indicating good fit were 100% for all conditions. When the unidimensional model was fitted, percentages of satisfactory CFI were close to 100% and percentages of TLI were greater than 80% in an unbalanced scale across conditions when the factor loadings on the general factor, on the positive specific factor, and on the negative specific factor were .6, .3, and .3, respectively. Percentages of RMSEA in the analysis unidimensional model were close to 100% in an unbalanced scale across conditions when the factor loadings on the general factor, the positive specific factor, and

the negative specific factor were 1) .6, .3, and .3, respectively, or 2) .3, .6, and .3, respectively. Percentages of SRMR in the analysis unidimensional model were 100% in a balanced scale across conditions when the factor loadings on the general factor, on the positive specific factor, and on the negative specific factor were 1) .6, .3, and .3, respectively, or 2) .3, .6, and .3, respectively. In addition, percentages of SRMR in the analysis unidimensional model were 100% in an unbalanced scale across conditions when the factor loadings on the negative specific factor were .3.

Moreover, the percentage of all information criteria correctly selecting the true model was close to 0. Therefore, neither approximate index nor information criteria correctly selected the true model; each approximate index identified the bi-factor with positive and negative wording effects with good fit, but not the only one model with good fit.

4.2.3 Bi-factor with negative wording effect

When the bi-factor with negative wording effect was the true underlying model, chi-square tended to favor bi-factor with two specific factors more frequently. Almost 100% of chi-square statistics in the analysis unidimensional model was significant, indicating that chi-square correctly identified the unidimensional model as a model with unacceptable fit. Similar to analysis for the data generation bi-factor model with positive and negative wording effects, other approximate indices did not work in identifying the true model and misspecified models. Specifically, CFI, TLI, RMSEA, and SRMR identified two-factor CFA and both bi-factor models as those with good fit because percentages of each index were 100% for all conditions. When the unidimensional model was fitted, percentages of CFI and TLI were very high when the factor loadings on the general factor and the negative specific factor were .6 and .3, respectively.

Percentages of RMSEA in the analysis unidimensional model were close to 100% in a balanced scale across conditions when the factor loadings on the general factor and the negative specific factor were .3 and .6, respectively. Percentages of RMSEA were also close to 100% across conditions when the factor loadings on the general factor and the negative specific factor were .6 and .3, respectively. Percentages of SRMR in the analysis unidimensional model were 100% across all conditions.

In addition, results of the information criteria showed that AIC, BIC, and SABIC functioned poorly in identifying the data generation bi-factor model with negative wording effect; AIC might work in identifying correctly the data generation model as its percentage ranged from 10% to 25%, such percentage was slightly higher than that of BIC and SABIC while the percentages of all information criteria selecting correctly the data generation model was lower than 25%.

4.3 POOLED MEAN OF FACTOR LOADING

The pattern of the pooled means for each analysis model was examined to explore any discrepancy in terms of factor loadings. The pooled means of standardized factor loading were calculated for general factor loading of positively worded items and negatively worded items separately, as well as the specific factor loadings. The pooled standard deviation of the factor loading was also calculated. Because of similar pooled means across levels of criterion-related validity and a much larger pooled standard deviation resulted from conditions wherein the criterion-related validity coefficient was zero, only results in conditions when criterion-related validity coefficient was non-zero were presented within each data generation model. When the

data generation model was the two-factor CFA, only results in conditions when criterion-related validity coefficient of positive and negative factors was both .5 were presented.

4.3.1 Two-factor CFA

Table 8 presents pooled mean in all four analysis models when the true underlying model was a two-factor CFA. For all conditions, pooled means from the analysis two-factor CFA model matched those true factor loadings and the average pooled standard deviations for positive trait and negative trait factors were .03 and .04, respectively. In the analysis unidimensional model, the pooled means of positively worded items were close to the true value of .6, with an average of .58 (pooled SD = .03) and a range from .49 to .60. The pooled means of negatively worded items were lower than their corresponding true value under different conditions, with an average of .45 (pooled SD = .04) and a range from .30 to .55 when the true value was .6 and an average of .18 (pooled SD = .04) and a range from .12 to .23 when the true value was .3.

In the analysis bi-factor model with positive and negative wording effects, the pooled means of factor loading for the general factor loading of positively worded items ranged from .41 to .53 with an average of .49 (pooled SD = .12), the pooled means for specific factor loading of positively worded items ranged from .24 to .38 with an average of .29 (pooled SD = .40); and all positively worded items loaded higher on the general factor than on the specific factor. When the true factor loading of negatively worded items was .6, the average pooled means for the general factor loading of negative items was .43 (pooled SD = .11) and ranged from .33 to .51; when the true factor loading of negative items was .3, the average pooled means for general factor loading of negative items was .21 (pooled SD = .07) and ranged from .16 to .25. The pooled means for specific factor loading of negative items ranged from .18 to .48 with an average of .31 (pooled

SD = .40). Negatively worded items loaded higher on the specific factor than on the general factor under conditions when 1) item loadings on the positive trait factor and negative trait factor were both .6 and the inter-factor correlation was .4 in an unbalanced scale, 2) item loadings on positive trait factor and negative trait factor was .6 and .3, respectively, and the inter-factor correlation was .4 in a balanced scale, 3) item loadings on positive trait factor and negative trait factor was .6 and .3, respectively, and the inter-factor correlation was .4 in an unbalanced scale, and 4) item loadings on positive trait factor and negative trait factor was .6 and .3, respectively, and the inter-factor correlation was .7 in an unbalanced scale.

In the analysis bi-factor model with negative wording effect, the pooled mean of the general factor loadings of positively worded items were .60 (pooled SD = .03) across all conditions. When the true factor loading of negatively worded items was .6, the average pooled means for general factor loading of negatively worded items was .33 (pooled SD = .03) and ranged from .24 to .42. When the true factor loading of negative items was .3, the average pooled means for the general factor loading of negative items was .16 (pooled SD = .03) and ranged from .12 to .21. Pooled means of specific factor loading of negatively worded items ranged from .20 to .55 with an average of .37 (pooled SD = .12). Only one negative item loaded slightly lower on the specific factor than on the general factor under the condition when 1) the criterion-related validity coefficient of positive and negative factor was both .5, 2) item loadings on positive and negative factors was .6 and .3, respectively, and 3) the inter-factor correlation was .7 in a balanced scale.

Table 8. Pooled mean of factor loadings for the data generation two-factor CFA model when criterion-related validity for both positive and negative trait factors was .5

Simulation Conditions			Analysis Model										
			1F		**2F**		Bi2				Bi1		
λ_p, λ_n	N_p, N_n	r	P	N	**P**	**N**	G_P	G_N	S_P	S_N	G_P	G_N	S_N
.6, .6	5, 5	.4	.49	.49	**.60**	**.60**	.41	.40	.38	.40	.60	.24	.55
		.7	.55	.55	**.60**	**.60**	.51	.51	.29	.30	.60	.42	.43
	7, 3	.4	.59	.30	**.60**	**.60**	.48	.33	.28	.48	.60	.24	.55
		.7	.59	.47	**.60**	**.60**	.52	.50	.26	.34	.60	.42	.43
.6, .3	5, 5	.4	.60	.14	**.60**	**.30**	.46	.17	.32	.24	.60	.12	.27
		.7	.60	.23	**.60**	**.30**	.51	.25	.28	.19	.60	.21	.20
	7, 3	.4	.60	.13	**.60**	**.30**	.49	.16	.28	.28	.60	.12	.28
		.7	.60	.22	**.60**	**.30**	.53	.24	.24	.25	.60	.21	.28

Note. λ_p = item loadings on the positive trait factor. λ_n = item loadings on the negative trait factor. N_p= number of positively worded items. N_n = number of negatively worded items. r = inter-factor correlation. 1F = unidimensional model. 2F = two-factor CFA. Bi2 = bi-factor with positive and negative wording effects. Bi1 = bi-factor with a negative wording effect. P = pooled mean of factor loadings related to positively worded items. N = pooled mean of factor loadings related to negatively worded items. G_P = general factor loadings related to positively worded items. G_N = general factor loadings related to negatively worded items. S_P = specific factor loadings related to positively worded items. S_N = specific factor loadings related to negatively worded items. Values in bold indicate the pooled means when the analysis model matched the true model.

4.3.2 Bi-factor model with positive and negative wording effects

Table 9 presents the pooled means in all four analysis models when the true underlying model was the bi-factor model with positive and negative wording effects. For all conditions, pooled means from the true analysis model matched those true factor loadings and pooled standard deviations around .1. In the analysis unidimensional model, when the true general factor loading was .6, the average pooled mean of positive items was .75 (pooled SD = .03) and ranged from .63 to .85 and the average pooled mean of negative items was .55 (pooled SD = .04) and ranged from .44 to .69. When the true general factor loading was .3, the average pooled mean of positive items was .61 (pooled SD = .07) and ranged from .45 to .67 and the average pooled mean of negative items was .23 (pooled SD = .08) and ranged from .15 to .46.

In the analysis two-factor CFA, when the true general factor loading was .6, the average pooled mean of positive items was .79 (pooled SD = .01) and ranged from .67 to .85 and the average pooled mean of negative items was .73 (pooled SD = .02) and ranged from .67 to .85. When the true general factor loading was .3, the pooled means of positive items were all .67 (pooled SD = .02) and the average pooled mean of negative items was .55 (pooled SD = .03) and ranged from .42 to .67.

In the analysis bi-factor model with negative wording effect, when the true general factor loading of positive items was .6, the average pooled mean of the general factor loading of positive items was .79 (pooled SD = .01) and ranged from .67 to .85. When the true general factor loading of positive items was .3, the pooled means of general factor loading of positive item were all .67 (pooled SD = .02). When the true general factor loading of negative item was .6, the average pooled mean of the general factor loading of negative items was .47 (pooled SD = .03) and ranged from .43 to .54. When the true general factor loading of negative items was .3,

the pooled means of the general factor loading of negative item were all .14 (pooled SD = .03). When the true specific factor loading of negative item was .6, the average pooled mean of the specific factor loading of negative items was .69 (pooled SD = .02) and ranged from .66 to .73. When true specific factor loading of negative item was .3, the average pooled mean of specific factor loading of negative items was .44 (pooled SD = .04) and ranged from .39 to .52. Negatively worded items loaded higher on the specific factor than on the general factor when factor loadings on the positive specific factor was specified as .6 in the true model.

Table 9. Pooled mean of factor loadings for the data generation bi-factor model with positive and negative wording effects

Simulation Conditions		Analysis Model										
$\lambda_g, \lambda_{sp}, \lambda_{sn}$	N_p, N_n	1F		2F		**Bi2**				Bi1		
		P	N	P	N	**G-P**	**G-N**	**S-P**	**S-N**	G-P	G-N	S-N
.6, .6, .6	5, 5	.70	.69	.85	.85	**.60**	**.60**	**.60**	**.60**	.85	.43	.73
	7, 3	.85	.46	.85	.85	**.60**	**.60**	**.60**	**.60**	.85	.43	.73
.6, .6, .3	5, 5	.84	.48	.85	.67	**.60**	**.60**	**.60**	**.29**	.85	.43	.52
	7, 3	.85	.44	.85	.67	**.60**	**.60**	**.60**	**.31**	.85	.43	.52
.6, .3, .3	5, 5	.63	.63	.67	.67	**.60**	**.60**	**.30**	**.30**	.67	.54	.39
	7, 3	.66	.58	.67	.67	**.60**	**.60**	**.29**	**.31**	.67	.54	.39
.3, .6, .6	5, 5	.45	.46	.67	.67	**.30**	**.30**	**.60**	**.60**	.67	.14	.66
	7, 3	.67	.16	.67	.67	**.30**	**.30**	**.60**	**.60**	.67	.14	.66
.3, .6, .3	5, 5	.67	.17	.67	.42	**.30**	**.30**	**.60**	**.30**	.67	.14	.40
	7, 3	.67	.15	.67	.43	**.30**	**.30**	**.60**	**.25**	.67	.14	.40

Note. λ_g= item loadings on the general factor. λ_{sp}= item loadings on the positive specific factor. λ_{sn} = item loadings on the negative specific factor. N_p = number of positively worded items. N_n = number of negatively worded items. 1F = unidimensional model. 2F = two-factor CFA. Bi2 = bi-factor with positive and negative wording effects. Bi1 = bi-factor with a negative wording effect. P = pooled mean of factor loadings related to positively worded items. N = pooled mean of factor loadings related to negatively worded items. G_P = general factor loadings related to positively worded items. G_N = general factor loadings related to negatively worded items. S_P = specific factor loadings related to positively worded items. S_N = specific factor loadings related to negatively worded items. Values in bold indicate the pooled means when the analysis model matched the true model.

4.3.3 Bi-factor model with negative wording effect

Table 10 presents pooled means in all four analysis models when the true underlying model was the bi-factor model with negative wording effect. For all conditions, pooled means from the analysis bi-factor model with negative wording effect matched those true factor loadings and pooled standard deviations less than .1. In the analysis unidimensional model, when the true general factor loading was .6, the average pooled mean of positive items was .54 (pooled SD = .03) and ranged from .46 to .59 and the average pooled mean of negative items was .73 (pooled SD = .02) and ranged from .63 to .84. When the true general factor loading was .3, the average pooled mean of positive items was .17 (pooled SD = .04) and ranged from .15 to .18 and the average pooled mean of negative items was .66 (pooled SD = .03) and ranged from .65 to .67.

In the analysis two-factor CFA model, the pooled means of positive items under each condition matched the true general factor loading of positive items. When the true general factor loading was .6, for positively worded items, the pooled means were all .6 (pooled SD = .02) while for the negatively worded items, the average pooled mean was .76 (pooled SD = .02) and ranged from .67 to .85. When the true general factor loading was .3, for positively worded items,

the pooled means were all .3 (pooled SD = .04) while for negatively worded items and pooled means were all .67 (pooled SD = .02).

In the analysis bi-factor model with positive and negative wording effects, both the pooled means of the general factor loading of positive and negative items matched the true general factor loading and the pooled means of specific factor loading of negative items matched the true specific factor loading of negative item. The average pooled means of the specific factor loading of positive items was .13 (pooled SD = .56) and ranged from .10 to .17.

Table 10. Pooled mean of factor loadings for the data generation bi-factor model with negative wording effect

Simulation Conditions		Analysis Model										
		1F		2F		Bi2				**Bi1**		
λ_g, λ_{sn}	N_p, N_n	P	N	P	N	G_P	G_N	S_P	S_N	**G_P**	**G_N**	**S_N**
.6, .6	5, 5	.46	.84	.60	.85	.59	.61	.14	.59	**.60**	**.60**	**.60**
	7, 3	.52	.80	.60	.85	.59	.60	.11	.59	**.60**	**.60**	**.60**
.6, .3	5, 5	.58	.66	.60	.67	.59	.60	.15	.29	**.60**	**.60**	**.30**
	7, 3	.59	.63	.60	.67	.60	.60	.10	.29	**.60**	**.60**	**.30**
.3, .6	5, 5	.15	.67	.30	.67	.30	.30	.17	.60	**.30**	**.30**	**.60**
	7, 3	.18	.65	.30	.67	.29	.30	.11	.58	**.30**	**.30**	**.60**

Note. λ_g = item loadings on the general factor. λ_{sn} = item loadings on the negative specific factor. N_p = number of positively worded items. N_n = number of negatively worded items. 1F = unidimensional model. 2F = two-factor CFA. Bi2 = bi-factor with positive and negative wording effects. Bi1 = bi-factor with a negative wording effect. P = pooled mean of factor loadings

related to positively worded items. N = pooled mean of factor loadings related to negatively worded items. G_P = general factor loadings related to positively worded items. G_N = general factor loadings related to negatively worded items. S_P = specific factor loadings related to positively worded items. S_N = specific factor loadings related to negatively worded items. Values in bold indicate the pooled means when the analysis model matched the true model.

4.4 BIAS IN STRENGTH INDICES

Table 11 and 12 present relative bias of ECV, composite reliability, and homogeneity coefficient for the two data generation bi-factor models and the two data analysis bi-factor models when criterion-related validity coefficient was non-zero. As depicted in Table 11, for the data generation bi-factor model with positive and negative wording effects in various conditions, the relative biases of ECV, ω, and ω_H were all less than 5%, indicating the estimation of ECV, homogeneity coefficient, and composite reliability in the bi-factor model with positive and negative wording effects was accurate with negligible bias. For the bi-factor model with a negative wording effect, 90% of conditions resulted in relative bias of ECV and a homogeneity coefficient greater than 10%, indicating substantial bias. The relative bias of the composite reliability for the bi-factor model with a negative wording effect was zero, indicating that the estimation of composite reliability in this bi-factor model was accurate without noticeable bias.

As shown in Table 12, for the data generation bi-factor model with a negative wording effect in various conditions, the relative biases of ECV, ω, and ω_H were all less than 5%, indicating the estimation of ECV, homogeneity coefficient, and composite reliability in the bi-factor model with a negative wording effects was accurate with negligible bias. For the bi-factor model with positive and negative wording effects, all the relative biases in ECV were negative

and their absolute values were larger than 10%, indicating that the model underestimated the ECV. For about 50% of the conditions in Table 12, the relative bias of composite reliability was moderate or substantial. Relative biases in homogeneity coefficient were all negative but within 5%, which were considered unnoticeable.

Table 11. Relative bias of ECV, ω, and ω_H for the data generation bi-factor model with positive and negative wording effects

Data Generation Model					Analysis Model					
Simulation Conditions		True Statistics			Bi2			Bi1		
$\lambda_g, \lambda_{sp}, \lambda_{sn}$	N_p, N_n	True ECV	True ω	True ω_H	Bias_E	Bias_ω	Bias_ω_H	Bias_E	Bias_ω	Bias_ω_H
.6, .6, .6	5, 5	.50	.95	.63	.01	.00	.00	.26	.00	.13
	7, 3	.50	.95	.60	.00	.00	.00	.55	.00	.45
.6, .6, .3	5, 5	.62	.92	.70	-.01	.00	.00	.25	.00	.13
	7, 3	.56	.94	.62	.00	.00	.00	.55	.00	.45
.6, .3, .3	5, 5	.80	.88	.78	-.01	.00	.00	.04	.00	.02
	7, 3	.80	.88	.77	-.02	.00	.00	.12	.00	.11
.3, .6, .6	5, 5	.20	.83	.28	.04	.00	.01	1.61	.00	.82
	7, 3	.20	.84	.25	.04	.00	.02	2.56	.00	1.89
.3, .6, .3	5, 5	.29	.75	.33	-.01	.01	.00	1.60	.00	.81
	7, 3	.24	.81	.27	.00	.01	.01	2.55	.00	1.89

Note. λ_g = item loadings on the general factor. λ_{sp} = item loadings on the positive specific factor. λ_{sn} = item loadings on the negative specific factor. N_p= number of positively worded items. N_n = number of negatively worded items. ω = composite reliability coefficient. ω_H = homogeneity coefficient. Bias_E = relative bias in ECV. Bias_ ω = relative bias in composite reliability coefficient. Bias_ω_H = relative bias in homogeneity coefficient. Bi2 = bi-factor model with positive and negative wording effects. Bi1 = bi-factor with a negative wording effect.

Table 12. Relative bias of ECV, $\boldsymbol{\omega}$, and $\boldsymbol{\omega_H}$ for the data generation bi-factor model with negative wording effect

Data Generation Model					Analysis Model					
Simulation Conditions		True Statistics			Bi2			Bi1		
λ_g, λ_{sn}	N_p, N_n	True ECV	True ω	True ω_H	Bias _E	Bias _ ω	Bias _ω_H	Bias _E	Bias _ ω	Bias _ω_H
.6, .6	5, 5	.67	.91	.73	-.16	.04	.00	.00	.00	.00
	7, 3	.77	.88	.81	-.15	.03	-.01	.00	.00	.00
.6, .3	5, 5	.89	.87	.81	-.20	.05	.00	.00	.00	.00
	7, 3	.93	.86	.84	-.15	.04	-.01	.00	.00	.00
.3, .6	5, 5	.33	.71	.36	-.29	.15	-.01	.01	.00	.00
	7, 3	.45	.60	.44	-.29	.20	-.02	.01	.00	.00

Note. λ_g = item loadings on the general factor. λ_{sn} = item loadings on the negative specific factor. N_p= number of positively worded items. N_n = number of negatively worded items. ω = composite reliability coefficient. ω_H = homogeneity coefficient. Bias_E = relative bias in ECV. Bias_ ω = relative bias in composite reliability coefficient. Bias_ ω_H = relative bias in homogeneity coefficient. Bi2 = bi-factor model with positive and negative wording effects. Bi1 = bi-factor with a negative wording effect.

4.5 BIAS OF CRITERION-RELATED VALIDITY COEFFICIENT

Table 13 presents the mean criterion-related validity estimates under the correct two-factor CFA. Those estimated validity coefficient in the analysis two-factor CFA model matched the true criterion-related validity at each level. For the unidimensional model and bi-factor model with a negative wording effect, mean validity estimations were greater than .7 when both the true positive and negative criterion-validity coefficients were .5 while mean validity estimations were between .5 and .6 when the true positive and negative criterion-validity coefficient was .5 and .1, respectively. Most conditions of the bi-factor model with two specific factors resulted in comparable validity coefficients at each condition when compared to that from the bi-factor model with negative wording effect.

Table 13. Mean criterion-related validity estimates for the data generation two-factor CFA model

Simulation Conditions				Analysis Model				
Criterion-related Validity Coefficient (Positive, Negative)	λ_p, λ_n	N_p, N_n	r	1F	2F_pos	2F_neg	Bi2	Bi1
.5, .5	.6, .6	5, 5	.4	.86	.50	.50	.68	.70
			.7	.93	.50	.50	.87	.85
		7, 3	.4	.75	.50	.50	.71	.70
			.7	.90	.50	.50	.88	.85
	.6, .3	5, 5	.4	.73	.50	.51	.64	.70
			.7	.88	.47	.53	.81	.85
		7, 3	.4	.71	.49	.51	.63	.70

Table 13 continued

			.7	.86	.43	.57	.81	.85
.5, .1	.6, .6	5, 5	.4	.52	.50	.10	.38	.54
			.7	.56	.50	.10	.48	.57
		7, 3	.4	.55	.50	.11	.47	.54
			.7	.58	.50	.10	.47	.57
	.6, .3	5, 5	.4	.54	.50	.10	.43	.54
			.7	.58	.50	.10	.49	.57
		7, 3	.4	.54	.50	.10	.45	.54
			.7	.57	.46	.14	.51	.57

Note. λ_p = item loadings on the positive trait factor. λ_n = item loadings on the negative trait factors. N_p = number of positively worded items. N_n= number of negatively worded items. r = inter-factor correlation. 1F = unidimensional model. 2F_pos = positive trait factor from two-factor CFA. 2F_neg = negative trait factor from two-factor CFA. Bi2 = bi-factor with positive and negative wording effects. Bi1 = bi-factor with a negative wording effect.

Tables 14 and 15 present the relative biases of the validity evidence for criterion relationships for the conditions of the true criterion-related validity coefficient of .5. The relative biases of the validity evidence for criterion relationships under the bi-factor model with positive and negative wording effects were close to 0 in the two data generation bi-factor models. As shown in Table 14, all relative biases in misspecified models were negative and at least 80% of the conditions resulted in relative biases greater than 10%, indicating that the estimation on the validity evidence for criterion relationships in the misspecified models were underestimated and those biases were substantial. There was no difference between the balanced and unbalanced conditions within the same fitted model. When the true validity coefficient was 0, all

misspecified models performed the same as the correct model, and the absolute bias was zero for all conditions.

Table 15 presents relative bias of criterion-related validity estimates under the data generation bi-factor model with negative wording effect. For the correct model, all biases were zero. For the misspecified bi-factor model with positive and negative wording effects, all biases were close to zero. For the unidimensional models, all biases were negative and most biases were substantial. For two-factor CFA, biases for the positive trait factor were close to zero while all biases for the negative trait factor were substantial.

Table 14. Relative bias of criterion-related validity estimates for the data generation bi-factor model with positive and negative wording effects

Simulation Conditions		Analysis Model				
$\lambda_g, \lambda_{sp}, \lambda_{sn}$	N_p, N_n	1F	2F_pos	2F_neg	Bi2	Bi1
.6, .6, .6	5, 5	-.21	-.53	-.53	.00	-.28
	7, 3	-.27	-.53	-.53	.00	-.28
.6, .6, .3	5, 5	-.23	-.76	-.26	.00	-.27
	7, 3	-.26	-.76	-.26	.00	-.27
.6, .3, .3	5, 5	-.05	-.50	-.50	.00	-.06
	7, 3	-.07	-.52	-.49	.00	-.08
.3, .6, .6	5, 5	-.48	-.63	-.63	.00	-.54
	7, 3	-.53	-.63	-.62	.01	-.54
.3, .6, .3	5, 5	-.51	-.76	-.37	-.01	-.53
	7, 3	-.53	-.75	-.37	.00	-.54

Note. λ_g = item loadings on the general factor. λ_{sp} = item loadings on the positive specific factor. λ_{sn} = item loadings on the negative specific factor. N_p = number of positively worded items. N_n= number of negatively worded items. 1F = unidimensional model. 2F_pos = positive trait factor from two-factor CFA. 2F_neg = negative trait factor from two-factor CFA. Bi2 = bi-factor with positive and negative wording effects. Bi1 = bi-factor with a negative wording effect.

Table 15. Relative bias of criterion-related validity estimates for the data generation bi-factor model with negative wording effect

Simulation Conditions		Analysis Model				
λ_g, λ_{sn}	N_p, N_n	1F	2F_pos	2F_neg	Bi2	Bi1
.6, .6	5, 5	-.23	.01	-1.01	.01	.00
	7, 3	-.12	.00	-1.00	.01	.00
.6, .3	5, 5	-.04	.00	-1.00	.00	.00
	7, 3	-.01	.01	-1.01	.00	.00
.3, .6	5, 5	-.51	.01	-1.01	.01	.00
	7, 3	-.43	.01	-1.01	.02	.00

Note. λ_g= item loadings on the general factor. λ_{sn} = item loadings on the negative specific factor. N_p= number of positively worded items. N_n = number of negatively worded items. 1F = unidimensional model. 2F_pos = positive trait factor from two-factor CFA. 2F_neg = negative trait factor from two-factor CFA. Bi2 = bi-factor with positive and negative wording effects. Bi1 = bi-factor with a negative wording effect.

4.6 POWER AND TYPE I ERROR RATES

The statistical power in detecting the criterion-related validity coefficient (when the true criterion-related validity coefficient was nonzero) and the Type I error rate (when the true

criterion-related validity coefficient was zero) for the correct model and misspecified models by the three data generation models was examined. For all conditions across the three data generation models, the unidimensional model had the best level of power, followed by the bi-factor model with negative wording effect (see Table 16). No difference in Type I error was present among the analysis models across all conditions; Type I error rates were all acceptable (see Table 17).

Table 16. Power by data generation model

Data Generation Model	Simulation Conditions				Analysis Model				
	Criterion-related Validity Coefficient (Positive, Negative)	Item Loadings	N_p, N_n	r	1F	2F _pos	2F _neg	Bi2	Bi1
2F	.5, .5	λ_p, λ_n .6, .6	5, 5	.4	1.00	**1.00**	**1.00**	.62	1.00
				.7	1.00	**.89**	**.90**	.72	1.00
			7, 3	.4	1.00	**1.00**	**1.00**	.64	1.00
				.7	1.00	**.88**	**.84**	.68	1.00
		.6, .3	5, 5	.4	1.00	**.99**	**.97**	.64	1.00
				.7	1.00	**.60**	**.49**	.78	1.00
			7, 3	.4	1.00	**.90**	**.84**	.63	1.00
				.7	1.00	**.48**	**.22**	.71	.95
	.5, .1	.6, .6	5, 5	.4	1.00	**1.00**	**.24**	.49	1.00

Table 16 continued

				.7	1.00	**.94**	**.12**	.62	1.00
			7, 3	.4	1.00	**1.00**	**.26**	.57	1.00
				.7	1.00	**.92**	**.10**	.61	1.00
			5, 5	.4	1.00	**1.00**	**.16**	.57	1.00
				.7	1.00	**.71**	**.07**	.66	.99
		.6, .3	7, 3	.4	1.00	**.98**	**.10**	.55	.98
				.7	1.00	**.57**	**.03**	.61	.91
		$\lambda_g, \lambda_{sp}, \lambda_{sn}$	5, 5		1.00	1.00	1.00	**1.00**	1.00
		.6, .6, .6	7, 3		1.00	1.00	1.00	**1.00**	1.00
		.6, .6, .3	5, 5		1.00	.79	1.00	**.98**	1.00
			7, 3		1.00	.76	1.00	**.99**	1.00
		.6, .3, .3	5, 5		1.00	.95	.95	**.99**	1.00
			7, 3		1.00	.90	.92	**.98**	1.00
		.3, .6, .6	5, 5		1.00	1.00	1.00	**1.00**	1.00
			7, 3		1.00	1.00	1.00	**1.00**	1.00
		.3, .6, .3	5, 5		1.00	.90	1.00	**.99**	1.00
Bi2	.5		7, 3		1.00	.85	1.00	**1.00**	1.00
		λ_g, λ_{sn}	5, 5		1.00	1.00	.05	.78	**1.00**
		.6, .6	7, 3		1.00	1.00	.06	.78	**1.00**
		.6, .3	5, 5		1.00	.98	.04	.82	**1.00**
Bi1	.5		7, 3		1.00	.96	.04	.76	**1.00**

Table 16 continued

			5, 5		1.00	1.00	.04	.81	**1.00**
		.3, .6	7, 3		1.00	1.00	.04	.79	**1.00**

Note. λ_p = item loadings on the positive trait factor. λ_n = item loadings on the negative trait factor. λ_g = item loadings on the general factor. λ_{sp} = item loadings on the positive specific factor. λ_{sn} = item loadings on the negative specific factor. N_p= number of positively worded items. N_n = number of negatively worded items. r = inter-factor correlation. 1F = unidimensional model. 2F_pos = positive trait from two-factor CFA. 2F_neg = negative trait from two-factor CFA. Bi2 = bi-factor with positive and negative wording effects. Bi1 = bi-factor with a negative wording effect. Values in bold indicate the pooled means when the analysis model matched the true model.

Table 17. Type I error rates by data generation model

Data Generation Model	Simulation Conditions			Analysis Model				
	Item Loadings	N_p, N_n	r	1F	2F_pos	2F_neg	Bi2	Bi1
2F	λ_p, λ_n .6, .6	5, 5	.4	.04	**.05**	**.04**	.02	.05
			.7	.05	**.04**	**.04**	.03	.05
		7, 3	.4	.04	**.06**	**.05**	.01	.05
			.7	.04	**.06**	**.05**	.02	.05
	.6, .3	5, 5	.4	.06	**.04**	**.03**	.03	.06
			.7	.06	**.03**	**.03**	.04	.06
		7, 3	.4	.05	**.03**	**.05**	.02	.05
			.7	.04	**.02**	**.02**	.02	.04
Bi2	$\lambda_g, \lambda_{sp}, \lambda_{sn}$ .6, .6, .6	5, 5		.06	.05	.05	**.06**	.08
		7, 3		.05	.05	.06	**.05**	.07

Table 17 continued

	.6, .6, .3	5, 5		.05	.05	.05	**.05**	.06
		7, 3		.04	.05	.07	**.04**	.05
	.6, .3, .3	5, 5		.06	.05	.05	**.06**	.06
		7, 3		.05	.05	.06	**.05**	.06
	.3, .6, .6	5, 5		.05	.05	.04	**.05**	.07
		7, 3		.04	.04	.07	**.04**	.05
	.3, .6, .3	5, 5		.06	.05	.04	**.06**	.08
		7, 3		.05	.05	.05	**.05**	.05
Bi1	λ_g, λ_{sn} .6, .6	5, 5		.06	.06	.05	.07	**.06**
		7, 3		.05	.05	.05	.06	**.05**
	.6, .3	5, 5		.05	.05	.05	.04	**.05**
		7, 3		.04	.05	.04	.03	**.04**
	.3, .6	5, 5		.07	.06	.06	.08	**.07**
		7, 3		.05	.04	.04	.06	**.05**

Note. λ_p = item loadings on the positive trait factor. λ_n = item loadings on the negative trait factors. λ_g = item loadings on the general factor. λ_{sp} = item loadings on the positive specific factor. λ_{sn} = item loadings on the negative specific factor. N_p = number of positively worded items. N_n = number of negatively worded items. r = inter-factor correlation. 1F = unidimensional model. 2F_pos = positive trait from two-factor CFA. 2F_neg = negative trait from two-factor CFA. Bi2 = bi-factor with positive and negative wording effects. Bi1 = bi-factor with a negative wording effect. Values in bold indicate the pooled means when the analysis model matched the true model.

5.0 DISCUSSION

The main purpose of this study was to assess the impact of misspecifying the model when using negatively worded items. Three data generation models were simulated: 1) a two correlated factor CFA with positively worded items on one factor and negatively worded items on the other factor, 2) a bi-factor CFA with two specific factors representing method factors related to positive and negative wording effects, and 3) a bi-factor CFA with one specific factor representing a method factor related to a negative wording effect. In addition to these three models, the unidimensional model was fitted in each data generation structure to examine the impact of wording effects on the validity evidence for criterion relationships.

Three research questions posed in chapter 3 were addressed in this study:

1) How well do model fit indices perform in identifying the correct model for negative wording effects?
2) What are the effects of negative wording on the estimates of internal-consistency coefficients?
3) What are the effects of negative wording on the validity evidence for criterion relationships and the internal structure of the measure?

This chapter interprets results in light of the research questions and discusses the findings in conjunction with other literature. This chapter also presents the limitations of interpretation, followed by implications for practice and recommendations for further research.

5.1 SUMMARY OF RESULTS/FINDINGS

Results regarding non-convergence of the analysis bi-factor model with positive and negative wording effects across three generation models imply that the analysis model containing the estimation of the validity coefficient is overparameterized when the data generation model simulated the criterion-related validity coefficient as zero. Findings in terms of model fit evaluation and impact on factor loadings, internal-consistency coefficients, and the validity evidence for criterion relationships and the internal structure of the measure when misspecifying the models for the negative wording effect were organized by data generation model in the order of the three research questions. Also, implications of the results were presented by integrating results with relevant literature to discuss consistencies and inconsistencies with results of those studies cited in the literature.

5.1.1 RQ1: How well do model fit indices perform in identifying the correct model for negative wording effects?

When the true underlying model was the two-factor CFA, all approximate indices do not work in identifying the correct model because results showed that all approximate indices identified the true model and misspecified bi-factor models as models with good fit. Chi-square, CFI, and TLI identified the misspecified unidimensional model associated with poor fit in more than half of the conditions and RMSEA and SRMR identified poor fit in a few conditions. In contrast to the performance of approximate indices, information criteria such as AIC, BIC, and SABIC functioned well in identifying the true model in the simulated balanced conditions only. This

may be because information criteria penalize the bi-factor models more than the two-factor CFA which is more parsimonious.

When the true underlying model was one of the bi-factor models, all approximate indices and information criteria performed poorly in identifying the true model except the chi-square statistics identified the bi-factor model with positive and negative wording effects most frequently. Each approximate index identified the true model with good fit, but not the only model with good fit. In this study, both the two-factor CFA and bi-factor models with negative wording effect (when the data generation model was bi-factor with positive and negative wording effects) or bi-factor model with positive and negative wording effects (when the data generation model was bi-factor with negative wording effect) were models being intentionally misspecified, but in almost all conditions fit values judged those models to be fitting well.

In sum, model fit indices performed poorly in selecting the true structural model among multidimensional models (two-factor and bi-factor models) as model fit indices suggested that the misspecified multidimensional models also provided a good fit under a variety of conditions. Monte Carlo studies (Gu et al., 2017; Reise et al., 2013) concluded that model fit indices (CFI, RMSEA, and/or SRMR) were not informative in model selection between the true bi-factor and misspecified unidimensional models, even though they tend to correctly identify the misfit of the unidimensional model. In addition, according to Morgan et al. (2015), when the correlated two-factor CFA was the true underlying structure, model fit indices favored the true model or bi-factor model dependent on simulated conditions. Such a preference was made because authors reported and compared mean values of each index. The model flagged as the best fitting model was the model with the highest CFI and TLI and the lowest RMSEA, SRMR, and information criteria. Percentage of each of these indices meeting the criteria for indicating good fit reported

in this study does not allow to select the model with optimal fitting among a set of good-fitting models. In practice, selecting one model among a set of good-fitting models based solely upon mode fit indices is generally not advisable. Instead, selection should be done on substantive grounds.

5.1.2 RQ2: What are the effects of negative wording on the estimates of internal-consistency coefficients?

Each true model produced unbiased estimates of strength indices (i.e., ECV, composite reliability, and homogeneity coefficient in this study) for its corresponding data generation model. When the data generation model was a bi-factor model with positive and negative wording effect and the analysis model was a bi-factor model with negative wording effect, estimates of composite reliability were unbiased which indicates that the estimation of variance attributed to both the general and specific factors was unbiased across all conditions. Both estimates of ECV and homogeneity coefficient were substantially biased in conditions except for conditions with a true ECV of .8. For those conditions wherein ECV was greater than .75, the relative bias was at acceptable levels.

When the data generation model was a bi-factor model with negative wording effect, and the analysis model was a bi-factor model with positive and negative wording effects, homogeneity coefficients were estimated without noticeable bias. Trivial bias existed in the estimation of composite reliability in most conditions except when mean item loadings on the specific factor were higher than on the general factor. These inflated biases might be a result of the inappropriate specification of the model structure. All biases in ECV were negative and substantial, indicating that the bi-factor model with positive and negative wording effects

generally underestimated the relative strength of the general factor to the specific factors; in other words, the inclusion of the positive wording effect was redundant.

Findings of bias of internal-consistency coefficients differed from conclusions drawn by Gu et al. (2017). In Gu et al., the misspecified unidimensional modeloverestimates the homogeneity coefficient, while slightly underestimates the composite reliability. This is because the misspecified model was the unidimensional model while the current study compared the internal consitency measures between two bi-factor models. This study did not compare the bias of internal-consistency coefficients for the unidimensional model because the homogeneity coefficient and the composite reliability are the same when there are no specific factors. However, there were some findings from this studywere consistent with Gu et al. in that when ECV was larger (>.80), the internal-consistency bias was smaller when the misspecified model is underparameterized.

5.1.3 RQ3: What are the effects of negative wording on the validity evidence for criterion relationships and the internal structure of the measure?

Validity evidence for Criterion relationship

When the data generation model was the two-factor CFA, those estimated validity coefficients in the analysis two-factor CFA were unbiased. In the condition when the criterion-related validity coefficient related to positively and negatively worded items was both .5, the estimated validity coefficients in the analysis of all misspecified models were inflated, indicating an overestimated prediction from the trait variables to an external criterion variable. This is not surprising as the validity evidence from two separate factors were imposed onto only one factor in the misspecified models. In the condition when the criterion-related validity coefficient related to

positively worded items and negatively worded items was .5 and .1, respectively, the estimated validity coefficients in the analysis unidimensional model and the bi-factor model with negative wording effect were overestimated as expected, but underestimated in the analysis bi-factor model with positive and negative wording effects, which might be due to the addition of one more latent factor in this model when compared to the two-factor data generation model.

When the data generation model was the bi-factor model with positive and negative wording effects, all misspecified models had negative and nontrivial biases, indicating that ignoring the wording effect severely underestimated the prediction from the trait variables. Further, the relative biases under the condition wherein the factor loadings on the general, positive specific, and on the negative specific factor was .6, .3, and .3, respectively, were acceptable (around 5%) in both analysis unidimensional model and the bi-factor model with negative wording effect. This is because ECV was very high in these conditions. The relative biases under conditions when factor loadings on the general factor was .3 were relatively larger than biases under other conditions in analysis models. This is because ECV was low when the general factor loadings were smaller than the specific factor loadings. This is consistent with Reise et al. (2013)'s finding that ECV negatively correlated with the bias of the validity evidence for criterion relationships. The relative biases of criterion-related validity coefficient depend on the presence of a strong general factor. As general factor loadings increase and the specific factor loadings decrease, relative bias in criterion-related validity coefficient decreases (Reise et al., 2013).

When the data generation model was the bi-factor model with negative wording effect, fitting the true model accurately estimated the prediction across all conditions. All biases were close to zero in the misspecified bi-factor model with positive and negative wording effects.

Most biases in the unidimensional model were substantial and all were negative, consistent with Gu et al. (2017)'s finding that ignoring wording effect underestimated the relation of the measure to the criterion.

All analysis models had acceptable Type I error rate for the criterion-related validity coefficient. The unidimensional model had the highest level of power detecting the validity evidence for criterion relationships across all conditions for all three generation models. This is not surprising when the two-factor model is the data generation model as the unidiemsnional model overestimated this coefficient. When the data generation model is one of the bi-factor models, the misspecified models have acceptable power under most conditions, even though they tend to underestimate the criterion-related validity coefficient. This may be due to the large sample size in this study. No relationship between statistical power and ECV was found in this study while Gu et al. (2017) claimed that the statistical power for detecting the validity evidence for criterion relationships positively correlated with ECV.

Internal structure

As expected, when the two-factor CFA was the true underlying model, the inter-factor correlation was high and when fitting the bi-factor model with negative wording effect or unidimensional model, factor loadings were close to the true value. Across all conditions, the pattern of factor loadings from the bi-factor model with negative wording effect suggests that the negative specific factor may be interpreted substantially and would suggest using a two-factor CFA in practice given the true underlying model is unknown.

When the data generation model was a bi-factor model with positive and negative wording effects, it is not surprising that the pooled means of factor loadings related to positively

worded items and/or negatively worded items were inflated in the analysis unidimensional model. This is because part of the item variance contributed by the specific factors were imposed on the one latent factor in the unidimensional model. For a similar reason, pooled means of factor loadings for both positive trait and negative trait factors were inflated in the analysis two-factor CFA model. For the analysis bi-factor model with negative wording effect, negatively worded items tended to load more on the specific factor than on the general factor, leading to underestimated general loadings related to negatively worded items, but overestimated general loadings related to positively worded items and overestimated specific factor loadings related to negatively worded items. It seems that when there was both positively and negatively worded items, biased factor loadings would result from not taking into account any wording effect.

When the data generation model was a bi-factor model with negative wording effect, the analysis bi-factor model with positive and negative wording effects produced unbiased factor loadings of general factors and correctly specified negative wording factor, while the factor loadings were negligible for the misspecified positive wording factor (all less than .20). The positively worded items' loadings on the general factor were far larger than those on the corresponding specific factor, suggesting that the positive wording effect could be small to negligible. This implies that overfitting a bi-factor model with more specific factors has no impact on factor loadings of the general factor and other specific factors. When fitting the two-factor CFA, factor loadings of the positive trait factor was estimated unbiasedly while the factor loadings of negative trait factor inflated due to ignoring the negative wording effect.

5.2 IMPLICATIONS AND LIMITATIONS

A few practical suggestions are provided based on the results from this study. First, researchers should be very cautious when using approximate fit indices or information criteria to select analysis models. Even though under some conditions these model fit indices correctly identify the misfit of the unidimensional model, they are not able to distinguish among the three multidimensional models for these analyses of negatively worded items. Researchers have to rely on substantive and conceptual grounds in model selection. Second, researchers are recommended to fit different models for possible wording effect, and examine carefully the internal structure (i.e., factor loadings) of different models. Between the two bi-factor models that differ in the addition of a positive wording effect, it is recommended to fit the bi-factor model with both positive and negative wording effects. When there are both positive and negative wording effects, omitting one or two specific factors would result with underestimated criterion-related validity and biased factor loadings. When there is only negative wording effect, over-fitting with an additional specific factor has no impact on the criterion-related validity coefficient or factor loadings. It is suggested, given the existence of negative worded items, both specific factors related to positive and negative wording effects should be considered. Only when a specific factor has negligible loadings (such as <.2), this specific factor is a candidate for removing. Third, ECV, the composite reliability (omega), and the homogeneity coefficient (*omega*H) should be computed and evaluated. High estimated ECV (such as >.80) justifies the use of specific factors for wording effects, while moderate to low ECV would make it difficult to decide whether the negatively worded items should be considered as a method effect or a substantive factor (another trait factor).

In sum, with percentages of model fit indices not working well in model selection, the model building strategy would be suggested as below. Researchers are suggested to first look at model fit indices and select those models being identified with fitting well. Next, by comparing the item loadings on the general factor and the specific factor(s), if loadings on the specific factor(s) are far higher than corresponding loadings on the general factor, it is an indicator that bi-factor model is not appropriate. The bi-factor model with both positive and negative wording specific factors should be estimated first. Only when loadings on one specific factor are negligible, this specific factor could be removed.

This study has several limitations. First, as any simulation study, this study is limited in the simulation conditions considered. The total number of items is fixed with only two ratios of positively worded items versus negatively worded items considered. The pattern of factor loadings is limited, not allowing complexity in real data, such as residual correlation or item cross-loading. It is impossible to simulate all possible real-world modeling violations in one single study. Second, results in terms of model fit percentages did correctly identify the data generation model as a model with a good fit, though percentages of model fit indices identified misspecified models as fitting well. This could occur when there was a small difference in absolute magnitudes of model fit indices across models. Future research should evaluate the absolute value of model fit indices when comparing across models for negatively worded items. Third, this study only focuses on bi-factor models with one general trait factor, and no specific domain factor (i.e., a substantive specific factor). If the target construct consists of multiple correlated dimensions, there might be a hybrid of the specific domain and method factors representing the specific factors. In this case, the bi-factor model will include 1) a general factor on which all items load 2) several specific factors shared by different sets of items whose content

are highly similar and 3) method factors taking into account wording effect(s). Future research is needed to examine the effect of misspecifying the model for the wording effects on the general trait factor and other domain specific factors.

APPENDIX A

ROSENBERG SELF-ESTEEM SCALE (RSES; ROSENBERG, 1965) ITEMS

Item	Content	Wording
1	I feel that I have a number of good qualities.	P
2	I wish I could have more respect for myself.	N
3	I feel that I'm a person of worth, at least on an equal plane with others.	P
4	I feel I do not have much to be proud of.	N
5	I take a positive attitude toward myself.	P
6	I certainly feel useless at times.	N
7	All in all, I'm inclined to feel that I am a failure.	N
8	I am able to do things as well as most other people.	P
9	At times I think I am no good at all.	N
10	On the whole, I am satisfied with myself.	P

Note. Response categories for items are: (1) Never true, (2) Seldom true, (3) Sometimes true, (4) Often true, (5) Almost always true. P = positively worded item. N = negatively worded item.

APPENDIX B

PERCENTAGES OF AIC, BIC, AND SABIC, CORRECTLY IDENTIFYING THE TRUE MODEL BY DATA GENERATION MODEL AND SIMULATION CONDITIONS

Data Generation Model	Simulation Conditions				Information Criteria		
	Criterion-related Validity Coefficient (Positive, Negative)	Item Loadings	N_p, N_n	r	AIC	BIC	SABIC
2F	.5, .5	λ_p, λ_n .6, .6	5, 5	.4	.84	1.00	1.00
				.7	.81	1.00	.99
			7, 3	.4	.13	.00	.03
				.7	.10	.00	.03
		.6, .3	5, 5	.4	.81	1.00	1.00
				.7	.81	.84	.97
			7, 3	.4	.11	.00	.02
				.7	.10	.00	.02
	.5, .1	.6, .6	5, 5	.4	.85	1.00	1.00

				.7	.82	1.00	1.00
			7, 3	.4	.12	.00	.03
				.7	.11	.00	.02
			5, 5	.4	.82	1.00	.99
				.7	.82	.79	.95
		.6, .3	7, 3	.4	.13	.00	.02
				.7	.11	.00	.02
Bi2	.5	$\lambda_g, \lambda_{sp}, \lambda_{sn}$	5, 5		.05	.00	.00
		.6, .6, .6	7, 3		.05	.00	.00
		.6, .6, .3	5, 5		.05	.00	.00
			7, 3		.05	.00	.00
		.6, .3, .3	5, 5		.07	.00	.00
			7, 3		.07	.00	.00
		.3, .6, .6	5, 5		.06	.00	.00
			7, 3		.05	.00	.00
		.3, .6, .3	5, 5		.06	.00	.00
			7, 3		.05	.00	.00
Bi1	.5	λ_g, λ_{sn}	5, 5		.11	.00	.01
		.6, .6	7, 3		.21	.02	.10
		.6, .3	5, 5		.14	.00	.02
			7, 3		.23	.02	.12

			5, 5		.12	.00	.01
		.3, .6	7, 3		.23	.02	.13

Note. λ_p = item loadings on the positive trait factor. λ_n = item loadings on the negative trait factor. λ_g = item loadings on the general factor. λ_{sp} = item loadings on the positive specific factor. λ_{sn} = item loadings on the negative specific factor. N_p = number of positively worded items. N_n = number of negatively worded items. r = inter-factor correlation. 2F = two-factor CFA. Bi2 = bi-factor with positive and negative wording effects. Bi1 = bi-factor with a negative wording effect.

APPENDIX C

PERCENTAGE OF EACH APPROXIMATE INDEX MEETING THE CRITERIA FOR INDICATING GOOD FIT FOR THE UNIDIMENSIONAL MODEL BY DATA GENERATION MODEL AND SIMULATION CONDITIONS

	Simulation Conditions				Approximate Indices				
Data Generation Model	Criterion-related Validity Coefficient (Positive, Negative)	Items Loadings	N_p, N_n	r	χ^2	CFI	TLI	RMSEA	SRMR
2F	.5, .5	λ_p, λ_n .6, .6	5, 5	.4	.00	.00	.00	.00	.24
				.7	.00	.00	.00	.14	1.00
			7, 3	.4	.00	.00	.00	.00	1.00
				.7	.00	.47	.16	.96	1.00
		.6, .3	5, 5	.4	.00	.38	.18	1.00	1.00
				.7	.38	.99	.97	1.00	1.00

			7, 3	.4	.26	1.00	1.00	1.00	1.00
				.7	.77	1.00	1.00	1.00	1.00
	.5, .1	.6, .6	5, 5	.4	.00	.00	.00	.00	.18
				.7	.00	.00	.00	.11	1.00
			7, 3	.4	.00	.00	.00	.00	1.00
				.7	.00	.48	.16	.97	1.00
		.6, .3	5, 5	.4	.01	.59	.35	1.00	1.00
				.7	.44	1.00	.97	1.00	1.00
			7, 3	.4	.50	1.00	1.00	1.00	1.00
				.7	.83	1.00	1.00	1.00	1.00
Bi2	.5	$\lambda_g, \lambda_{sp}, \lambda_{sn}$ .6, .6, .6	5, 5		.00	.00	.00	.00	.00
			7, 3		.00	.00	.00	.00	.00
		.6, .6, .3	5, 5		.00	.00	.00	.00	.04
			7, 3		.00	.66	.10	.00	1.00
		.6, .3, .3	5, 5		.00	.35	.06	.20	1.00
			7, 3		.00	.98	.85	.95	1.00
		.3, .6, .6	5, 5		.00	.00	.00	.00	.00
			7, 3		.00	.00	.00	.00	.02
		.3, .6, .3	5, 5		.00	.00	.00	.00	1.00
			7, 3		.00	.72	.36	.97	1.00
Bi1	.5	λ_g, λ_{sn}	5, 5		.00	.00	.00	.00	1.00
			7, 3		.00	.00	.00	.00	1.00

		.6, .6							
		.6, .3	5, 5		.03	1.00	1.00	1.00	1.00
			7, 3		.12	1.00	1.00	1.00	1.00
		.3, .6	5, 5		.00	.15	.03	.98	1.00
			7, 3		.00	.00	.00	.65	1.00

Note. λ_p = item loadings on the positive trait factor. λ_n = item loadings on the negative trait factors. λ_g = item loadings on the general factor. λ_{sp} = item loadings on the positive specific factor. λ_{sn} = item loadings on the negative specific factor. N_p = number of positively worded items. N_n = number of negatively worded items. r = inter-factor correlation. 2F = two-factor CFA. Bi2 = bi-factor with positive and negative wording effects. Bi1 = bi-factor with a negative wording effect. χ^2 = percentage of non-significant chi-square. CFI = percentage of CFI equals to or greater than .95. TLI = percentage of TLI equals to or greater than .95. RMSEA = percentage of RMSEA equals to or less than .06. SRMR = percentage of SRMR equals to or less than .08.

BIBLIOGRAPHY

Aguado, J., Campbell, A., Ascaso, C., Navarro, P., Garcia-Esteve, L., & Luciano, J. V. (2012). Examining the Factor Structure and Discriminant Validity of the 12-Item General Health Questionnaire (GHQ-12) Among Spanish Postpartum Women. *Assessment, 19*(4), 517-525. doi:10.1177/1073191110388146.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723. doi:10.1109/TAC.1974.1100705

Alessandri, G., Vecchione, M., Donnellan, B. M., & Tisak, J. (2013). An Application of the LC-LSTM Framework to the Self-esteem Instability Case. *Psychometrika, 78*(4), 769-792. doi:10.1007/s11336-013-9326-4

Alessandri, G., Vecchione, M., Eisenberg, N., & Laguna, M. (2015). On the factor structure of the Rosenberg (1965) General Self-Esteem Scale. *Psychol Assess, 27*(2), 621-635. doi:10.1037/pas0000073

Alessandri, G., Vecchione, M., Tisak, J., & Barbaranelli, C. (2011). Investigating the nature of method factors through multiple informants: Evidence for a specific factor? *Multivariate Behavioral Research, 46*(4), 625. doi:10.1080/00273171.2011.589272

Ang, R. P., Neubronner, M., Oh, S.-A., & Leong, V. (2006). Dimensionality of rosenberg's self-esteem scale among normal-technical stream students in Singapore. *Current Psychology, 25*(2), 120-131. doi:10.1007/s12144-006-1007-3

Bagley, C., Bolitho, F., & Bertrand, L. (1997). Norms and Construct Validity of the Rosenberg Self-Esteem Scale in Canadian High School Populations: Implications for Counselling. *Canadian Journal of Counselling, 31*(1), 82.

Bagozzi, R. P. (1993). Assessing Construct Validity in Personality Research: Applications to Measures of Self-Esteem. *Journal of Research in Personality, 27*(1), 49-87. doi:10.1006/jrpe.1993.1005

Barnette, J. J. (1999). *Likert response alternative direction: SA to SD or SD to SA: Does it make a difference?* Paper presented at the American Educational Research Association.

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. doi:10.1177/00131640021970592

Bassili, J. N., & Scott, B. S. (1996). Response latency as a signal to question problems in survey research. *The Public Opinion Quarterly, 60*(3), 390-399.

Baumgartner, H., & Jan-Benedict, E. M. S. (2001). Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research, 38*(2), 143-156. doi:10.1509/jmkr.38.2.143.18840

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. doi:10.1111/j.1745-3984.1985.tb01061.x

Billiet, J. B., & McClendon, M. J. (2000). Modeling Acquiescence in Measurement Models for Two Balanced Sets of Items. *Structural Equation Modeling: A Multidisciplinary Journal, 7*(4), 608-628. doi:10.1207/S15328007SEM0704_5

Boduszek, D., Hyland, P., Dhingra, K., & Mallett, J. (2013). The factor structure and composite reliability of the Rosenberg self-esteem scale among ex-prisoners. *Personality and Individual Differences, 55*(8), 877-881. doi:10.1016/j.paid.2013.07.014

Boduszek, D., Shevlin, M., Mallett, J., Hyland, P., & O'Kane, D. (2012). Dimensionality and construct validity of the Rosenberg self-esteem scale within a sample of recidivistic prisoners. *Journal of Criminal Psychology, 2*(1), 19. doi:10.1108/20093821211210468

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K. A., & Paxton, P. (1998). Detection and Determinants of Bias in Subjective Measures. *American Sociological Review, 63*(3), 465-478.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review, 111*(4), 1061-1071. doi:10.1037/0033-295X.111.4.1061

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

Byrne, B. M., & Goffin, R. D. (1993). Modeling MTMM data from additive and multiplicative covariance structures: An audit of construct validity concordance. *Multivariate Behavioral Research, 28*(1), 67-96.

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1997). Beyond Bipolar Conceptualizations and Measures: The Case of Attitudes and Evaluative Space. *Personality and Social Psychology Review, 1*(1), 3-25. doi:10.1207/s15327957pspr0101_2

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment* (Vol. no. 07-017.). Beverly Hills: Sage Publications.

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. doi:10.1207/s15327906mbr4102_5

Chessa, A. G., & Holleman, B. C. (2007). Answering attitudinal questions: modelling the response process underlying contrastive questions. *Applied Cognitive Psychology, 21*(2), 203-225. doi:10.1002/acp.1337

Clark, H. H. (1976). *Semantics and comprehension* (Vol. 187). The Hague: Mouton.

Cloud, J., & Vaughan, G. M. (1970). Using balanced scales to control acquiescence. *Sociometry, 33*(2), 193-202.

Coleman, C. M. (2013). *Effects of negative keying and wording in attitude measures: A mixed-methods study*. Retrieved from http://commons.lib.jmu.edu/diss201019/73

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

Conway, J. M., Lievens, F., Scullen, S. E., & Lance, C. E. (2004). Bias in the Correlated Uniqueness Model for MTMM Data. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(4), 535-559. doi:10.1207/s15328007sem1104_3

Corwyn, R. F. (2000). The Factor Structure of Global Self-Esteem among Adolescents and Adults. *Journal of Research in Personality, 34*(4), 357-379. doi:10.1006/jrpe.2000.2291

Cronbach, L. J. (1942). Studies of acquiescence as a factor in the true-false test. *Journal of Educational Psychology, 33*(6), 401-415. doi:10.1037/h0054677

Cronbach, L. J. (1946). Response Sets and Test Validity. *Educational and Psychological Measurement, 6*(4), 475.

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* (Vol. 1, pp. 112-132). New York: Routledge.

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling: A Multidisciplinary Journal, 13*(3), 440-464. doi:10.1207/s15328007sem1303_6

DiStefano, C., & Motl, R. W. (2009). Personality correlates of method effects due to negatively worded items on the Rosenberg Self-Esteem scale. *Personality and Individual Differences, 46*(3), 309-313. doi:10.1016/j.paid.2008.10.020

Dobson, C., Goudy, W. J., Keith, P. M., & Powers, E. (1979). Further analysis of of Rosenberg's Self-Esteem Scale. *Psychological Reports, 44*(2), 639-641. doi:10.2466/pr0.1979.44.2.639

Donnellan, M. B., Ackerman, R. A., & Brecheen, C. (2016). Extending Structural Analyses 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. doi:10.1080/00223891.2015.1058268

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 Years After Likert: Thurstone Was Right. *Industrial and Organizational Psychology, 3*(4), 465-476. doi:10.1111/j.1754-9434.2010.01273.x

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika, 65*(2), 241-261. doi:10.1007/BF02294377

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating Trait Effects From Trait-Specific Method Effects in Multitrait-Multimethod Models: A Multiple-Indicator CT-C(M−1) Model. *Psychological methods, 8*(1), 38-60. doi:10.1037/1082-989X.8.1.38

Flora, D. B., & Curran, P. J. (2004). An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis With Ordinal Data. *Psychological methods, 9*(4), 466-491. doi:10.1037/1082-989X.9.4.466

Gana, K., Saada, Y., Bailly, N., Joulain, M., Hervé, 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*(4), 406-416. doi:10.1016/j.jrp.2013.03.011

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist, 46*(2), 107-119. doi:10.1037//0003-066X.46.2.107

Gnambs, T., Scharl, A., & Schroeders, U. (2018). The Structure of the Rosenberg Self-Esteem Scale: A Cross-Cultural Meta-Analysis. *Zeitschrift für Psychologie, 226*(1), 14-29.

Goldsmith, R. E., & Desborde, R. (1991). A Validity Study of a Measure of Opinion Leadership. *Journal of Business Research, 22*(1), 11.

Gray-Little, B., Williams, V. S. L., & Hancock, T. D. (1997). An Item Response Theory Analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 23*(5), 443-451. doi:10.1177/0146167297235001

Greenberger, E., Chen, C., Dmitrieva, J., & Farruggia, S. P. (2003). Item-wording and the dimensionality of the Rosenberg Self-Esteem Scale: do they matter? *Personality and Individual Differences, 35*(6), 1241-1254. doi:10.1016/S0191-8869(02)00331-8

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. doi: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*, 1-11. doi:10.1080/10705511.2017.1286228

Hensley, W. E., & Roberts, M. K. (1976). Dimensions of Rosenburg's self-esteem scale. *Psychological Reports, 38*(2), 583-584.

Holleman, B. (1999). Wording Effects in Survey Research Using Meta-Analysis to Explain the Forbid/Allow Asymmetry. *Journal of Quantitative Linguistics, 6*(1), 29-40. doi:10.1076/jqul.6.1.29.4145

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2*(1), 41-54.

Horan, P. M., DiStefano, C., & Motl, R. W. (2003). Wording Effects in Self-Esteem Scales: Methodological Artifact or Response Style? *Structural Equation Modeling: A Multidisciplinary Journal, 10*(3), 435-455. doi:10.1207/S15328007SEM1003_6

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. doi:10.1080/10705519909540118

Huang, C., & Dong, N. (2011). Factor Structures 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

Hughes, G. D. (2009). The Impact of Incorrect Responses to Reverse-Coded Survey Items. *Research in the Schools, 16*(2), 76.

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

Kam, C. C. S. (2016). Why Do We Still Have an Impoverished Understanding of the Item Wording Effect? An Empirical Examination. *Sociological Methods & Research*. doi:10.1177/0049124115626177

Kamoen, N., Holleman, B., Mak, P., Sanders, T., & van den Bergh, H. (2011). Agree or Disagree? Cognitive Processes in Answering Contrastive Survey Questions. *Discourse Processes, 48*(5), 355-385. doi:10.1080/0163853x.2011.578910

Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology, 12*(3), 247-252. doi:10.1016/0022-1031(76)90055-X

Knowles, E. S., & Condon, C. A. (1999). Why People Say "Yes": A Dual-Process Theory Of Acquiescence. *Journal of Personality and Social Psychology, 77*(2), 379-386. doi:10.1037/0022-3514.77.2.379

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*(1), 537-567. doi:10.1146/annurev.psych.50.1.537

Lance, C. E., Baranik, L. E., Lau, A. R., & Scharlau, E. A. (2009). If it ain't trait it must be method: (Mis)application of the multitrait-multimethod design in organizational research. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: doctrine, verity and fable in the organizational and social sciences* (pp. 337-360). New York: Routledge.

Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait-correlated method and correlated uniqueness models for multitrait-multimethod data. *Psychological methods, 7*(2), 228-244. doi:10.1037//1082-989X.7.2.228

Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., Raudsepp, L., Liukkonen, J., . . . Sport, S. (2012). Method Effects: The Problem With Negatively Versus Positively Keyed Items. *Journal of Personality Assessment, 94*(2), 196. doi:10.1080/00223891.2011.645936

Magazine, S. L., Williams, L. J., & Williams, M. L. (1996). A Confirmatory Factor Analysis Examination of Reverse Coding Effects in Meyer and Allen's Affective and Continuance Commitment Scales. *Educational and Psychological Measurement, 56*(2), 241-250. doi:10.1177/0013164496056002005

Marsh, H. W. (1989). Confirmatory Factor Analyses of Multitrait-Multimethod Data: Many Problems and a Few Solutions. *Applied Psychological Measurement, 13*(4), 335-361. doi:10.1177/014662168901300402

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. doi:10.1037/0022-3514.70.4.810

Marsh, H. W., & Bailey, M. (1991). Confirmatory Factor Analyses of Multitrait-Multimethod Data: A Comparison of Alternative Models. *Applied Psychological Measurement, 15*(1), 47-70. doi:10.1177/014662169101500106

Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: Sage.

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.supp

Marshall, G. N., Wortman, C. B., Kusulas, J. W., Hervig, L. K., & Vickers, R. R. (1992). Distinguishing Optimism From Pessimism: Relations to Fundamental Dimensions of Mood and Personality. *Journal of Personality and Social Psychology, 62*(6), 1067-1074. doi:10.1037/0022-3514.62.6.1067

Messick, S. (1991). Psychology and Methodology of Response Styles. In R. E. Snow & D. E. Wiley (Eds.), *Improving Inquiry in Social Science: A Volume in Honor of Lee J. Cronbach.* (pp. 161-200). Hillsdale, NJ: Lawrence Erlbaum Associates.

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development and validation of the penn state worry questionnaire. *Behaviour Research and Therapy, 28*(6), 487-495. doi:10.1016/0005-7967(90)90135-6

Michaelides, M. P., Koutsogiorgi, C., & Panayiotou, G. (2016a). 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. doi:10.1080/00223891.2015.1089248

Michaelides, M. P., Koutsogiorgi, C., & Panayiotou, G. (2016b). Method/Group Factors: Inconsequential but Meaningful—A Comment on Donnellan, Ackerman, and Brecheen (2016). *Journal of Personality Assessment*, 1-2. doi:10.1080/00223891.2016.1233560

Michaelides, M. P., Zenger, M., Koutsogiorgi, C., Brähler, E., Stöbel-Richter, Y., & Berth, H. (2016). Personality correlates and gender invariance of wording effects in the German version of the Rosenberg Self-Esteem Scale. *Personality and Individual Differences, 97*, 13-18. doi:10.1016/j.paid.2016.03.011

Morgan, G., Hodge, K., Wells, K., & Watkins, M. (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

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), 116-124. doi:10.1080/10705511.2014.961800

Motl, R. W., & DiStefano, C. (2002). Longitudinal Invariance of Self-Esteem and Method Effects Associated With Negatively Worded Items. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(4), 562-578. doi:10.1207/s15328007sem0904_6

Mulaik, S. A. (1971). *The foundations of factor analysis*. New York: McGraw-Hill.

Myers, N. D., Martin, J. J., Ntoumanis, N., Celimli, S., & Bartholomew, K. J. (2014). Exploratory bifactor analysis in sport, exercise, and performance psychology: A substantive-methodological synergy. *Sport, Exercise, and Performance Psychology, 3*(4), 258-272. doi:10.1037/spy0000015

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

Ory, J. C. (1982). Item Placement and Wording Effects on Overall Ratings. *Educational and Psychological Measurement, 42*(3), 767-775. doi:10.1177/001316448204200307

Owens, T. J. (1993). Accentuate the Positive-and the Negative: Rethinking the Use of Self-Esteem, Self-Deprecation, and Self-Confidence. *Social Psychology Quarterly, 56*(4), 288-299.

Owens, T. J. (1994). Two Dimensions of Self-Esteem: Reciprocal Effects of Positive Self-Worth and Self-Deprecation on Adolescent Problems. *American Sociological Review, 59*(3), 391-407.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (Vol. 1, pp. 17-59). San Diego, CA: Academic Press.

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

Pohl, S., & Steyer, R. (2010). Modeling common traits and method effects in multitrait-multimethod analysis. *Multivariate Behavioral Research, 45*(1), 45-72. doi:10.1080/00273170903504729

Pullmann, H., & Allik, J. (2000). The Rosenberg Self-Esteem Scale: Its dimensionality, stability and personality correlates in Estonian. *Personality and Individual Differences, 28*(4), 701-715. doi:10.1016/S0191-8869(99)00132-4

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

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*(8), 1597-1607. doi:10.1016/j.paid.2006.10.035

Ray, J. J. (1983). Reviving the problem of acquiescent response bias. *Journal of Social Psychology, 121*(1), 81-96. doi:10.1080/00224545.1983.9924470

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

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. doi:10.1080/00273171.2016.1243461

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, 16*(S1), 19-31. doi:10.1007/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.

Riley-Tillman, T. C., Chafouleas, S. M., Christ, T., Briesch, A. M., & LeBel, T. J. (2009). The impact of item wording and behavioral specificity on the accuracy of direct behavior ratings (DBRs). *School Psychology Quarterly, 24*(1), 1-12. doi:10.1037/a0015248

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*(1), 3-32. doi:10.1177/01466216000241001

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 27*(2), 151-161. doi:10.1177/0146167201272002

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

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, N.J U6 - ctx_ver=Z39.88-2004&ctx_enc=info%3Aofi%2Fenc%3AUTF8&rfr_id=info%3Asid%2Fsummon.serialssolutions.com&rft_val_fmt=info%3Aofi%2Ffmt%3Akev%3Amtx%3Abook&rft.genre=book&rft.title=Society+and+the+adolescent+selfimage&rft.au=Rosenberg%2C+Morris&r

ft.date=1965&rft.pub=Princeton+University+Press&rft.externalDocID=585615&paramdict=en-US U7 - Book: Princeton University Press.

Roszkowski, M. J., & Soven, M. (2010). Shifting gears: consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education, 35*(1), 113-130. doi:10.1080/02602930802618344

Salerno, L., Ingoglia, S., & Lo Coco, G. (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. doi:10.1016/j.paid.2017.02.063

Sauley, K. S., & Bedeian, A. G. (2000). Equity sensitivity: Construction of a measure and examination of its psychometric properties. *Journal of Management, 26*(5), 885-910. doi:10.1016/S0149-2063(00)00062-3

Sauro, J., & Lewis, J. (2011, 2011). *When designing usability questionnaires, does it hurt to be positive?*

Savalei, V., & Falk, C. F. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate Behav Res, 49*(5), 407-424. doi:10.1080/00273171.2014.931800

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

Schmitt, D. P., & Allik, J. (2005). Simultaneous administration of the Rosenberg Self-Esteem Scale in 53 nations: Exploring the universal and culture-specific features of global self-esteem. *Journal of Personality and Social Psychology, 89*(4), 623-642. doi:10.1037/0022-3514.89.4.623

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

Schriesheim, C. A., & Eisenbach, R. J. (1995). An exploratory and confirmatory factor-analytic investigation of item wording effects on the obtained factor structures of survey questionnaire measures. *Journal of Management, 21*(6), 1177-1193. doi:10.1016/0149-2063(95)90028-4

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. doi:10.1177/0013164491511005

Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. *Educational and Psychological Measurement, 41*(4), 1101-1114. doi:10.1177/001316448104100420

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461-464. doi:10.1214/aos/1176344136

Schweizer, K. (2012). On correlated errors. *European Journal of Psychological Assessment, 28*(1), 1-2. doi:10.1027/1015-5759/a000094

Shevlin, M. E., Bunting, B. P., & Lewis, C. A. (1995). Confirmatory factor analysis of the Rosenberg self-esteem scale. *Psychological Reports, 76*(3), 707-710.

Solís Salazar, M. (2015). The dilemma of combining positive and negative items in scales. *Psicothema U6 - ctx_ver=Z39.88-2004&ctx_enc=info%3Aofi%2Fenc%3AUTF-8&rfr_id=info%3Asid%2Fsummon.serialssolutions.com&rft_val_fmt=info%3Aofi%2Ffmt%3Akev%3Amtx%3Ajournal&rft.genre=article&rft.atitle=The+dilemma+of+combining+positive+and+negative+items+in+scales&rft.jtitle=Psicothema&rft.au=Sol%C3%ADs+Salazar%2C+Mart%C3%ADn&rft.date=2015&rft.eissn=1886144X&rft.volume=27&rft.issue=2&rft.spage=192&rft_id=info%3Apmid%2F25927700&rft.externalDocID=25927700&paramdict=en-US U7 - Journal Article, 27*(2), 192.

Sonderen, v. 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), 1-7.

Supple, A. J., Su, J., Plunkett, S. W., Peterson, G. W., & Bush, K. R. (2013). Factor structure of the Rosenberg Self-Esteem Scale. *Journal of Cross-Cultural Psychology, 44*(5), 748-764. doi:10.1177/0022022112468942

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research, 45*(1), 116-131. doi:10.1509/jmkr.45.1.116

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*(4), 529-554. doi:10.1086/214483

Tomas, J. M., & Oliver, A. (1999). Rosenberg's self-esteem scale: Two factors or method effects. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 84-98. doi:10.1080/10705519909540120

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

Vasconcelos-Raposo, J., Fernandes, H. M., Teixeira, C. M., & Bertelli, R. (2012). Factorial validity and invariance of the Rosenberg Self-Esteem Scale among Portuguese youngsters. *Social Indicators Research, 105*(3), 483-498. doi:10.1007/s11205-011-9782-0

Vecchione, M., Alessandri, G., Caprara, G. V., & Tisak, J. (2014). Are method effects permanent or ephemeral in nature? The case of the Revised Life Orientation Test. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(1), 117-130. doi:10.1080/10705511.2014.859511

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: A Multidisciplinary Journal, 8*(2), 275-286. doi:10.1207/S15328007SEM0802_6

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

Wang, Y., Kong, F., Huang, L., & Liu, J. (2016). Neural correlates of biased responses: The negative method effect in the Rosenberg Self-Esteem Scale is associated with Right Amygdala Volume: Neural correlates of method effect. *Journal of Personality, 84*(5), 623-632. doi:10.1111/jopy.12185

Watson, D., & Clark, L. A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063-1070.

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. doi:10.1080/0260293032000130234

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *JMR, Journal of Marketing Research, 49*(5), 737.

Weijters, B., Geuens, M., & Schillewaert, N. (2009). The proximity effect: The role of inter-item distance on reverse-item bias. *International Journal of Research in Marketing, 26*(1), 2-12. doi:10.1016/j.ijresmar.2008.09.003

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*(1), 1-26. doi:10.1177/014662168500900101

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

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for Confirmatory Factor Analysis. *Journal of Psychopathology and Behavioral Assessment, 28*(3), 186-191. doi:10.1007/s10862-005-9004-7

Wu, Y., Zuo, B., Wen, F., & Yan, L. (2017). Rosenberg Self-Esteem Scale: Method effects, factorial structure and scale invariance across migrant child and urban child populations in China. *Journal of Personality Assessment, 99*(1), 83-93.

Yang-Wallentin, F., Jöreskog, K. G., Luo, H., Humanistisk-samhällsvetenskapliga, v., Uppsala, u., Samhällsvetenskapliga, f., & Statistiska, i. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling: A Multidisciplinary Journal, 17*(3), 392-423. doi:10.1080/10705511.2010.489003

Zhang, X., & Savalei, V. (2016). Improving the factor structure of psychological scales: the expanded format as an alternative to the Likert Scale format. *Educational and Psychological Measurement, 76*(3), 357.