ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

CARELESS RESPONDING AND WORDING FACTORS IN LIKERT-TYPE SCALES: MODELLING AND FIXING THE PROBLEM

JEFFREY C. KENNEDY

NANYANG BUSINESS SCHOOL

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy

2009

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

ACKNOWLEDGEMENTS

To my supervisor, Ang Soon, who brought me to Singapore and provided me the opportunity to complete this long overdue academic journey – thank you for providing me the freedom to pursue my interests, and for guidance toward a successful end result.

I also want to express special thanks to Sasha Chernyshenko, for the unique combination of Ukrainian and Kiwi pragmatism underpinning his advice, and to Ringo Ho for his insightful feedback and suggestions. I value your support, friendship and advice.

Karl Pajo, Greet Van Hoye, and Cheryl Tay graciously allowed me to use their samples of CSES data – thank you for your generosity.

Finally, a big thank you to Nicky, Rosemary, Andrew and Kate for keeping me grounded, and to my parents, Miles and Betsy Kennedy, for their unfailing optimism that I might one day get around to completing a PhD.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Table of Contents

LIST OF TABLES iii

LIST OF FIGURES iv

ABSTRACT v

CHAPTER ONE
INTRODUCTION 1

Rationale for Using Negatively-Keyed Items 2

Problems Introduced by Negatively-Keyed Items 3

Wording Factors and Confirmatory Factor Analysis 7

Overview of Dissertation 9

CHAPTER TWO
LITERATURE REVIEW 11

Likert's Technique for Attitude Scaling 11

Acquiescent Responding 14

Problems Introduced by Negatively-Keyed Items 17
Psychometric Problems Associated with Negatively Worded Items 18
Item Wording Factors 20
RSES Wording Factor as Artefact 21
RSES Wording Factor as Substantive Factor 25

The Source of Item Wording Factors 28
Item Wording Factors are Meaningful Constructs 29
Item Wording Factors are Just Method Effects 32
Differing Item Response Distributions 35
Cognitive Complexity of Negative Items 37
Carelessness 42

The Need for Further Research into Careless Responding 49

CHAPTER THREE
ITEM WORDING EFFECTS IN THE CORE SELF-EVALUATIONS SCALE 53

Core Self-Evaluations 53
Measuring Core Self-Evaluations 54
Item Wording Effects in the CSES 56

Method 61
Instruments 61
Samples 63
Statistical Methods 64

Results 74
Factor Structure 74
Structural Models 83

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Discussion 85

Conclusions 89

Chapter Summary 90

CHAPTER FOUR
MONTE CARLO SIMULATION OF CARELESS RESPONDING 91

Item Wording Factors 91

Careless Responding 92

Study One – Patterns of Careless Responding 93
Method 97
Results 98
Influential Cases in the Two-Factor (Positive and Negative) Model 98
Cases Which Influence Individual Item Parameters 99
Discussion 107

Study Two – Monte Carlo Simulation 108
Method 109
Data Sets 109
Determining the Number of Factors 113
CFA Model Fit 114
Results 115
Discussion 125

CHAPTER FIVE
IDENTIFYING CARELESS CASES IN EMPIRICAL DATA 129

Identifying Careless Cases in the Monte Carlo Data 130
Method 131
Results 132

Using Simulated Data as a Guide to 'Cleaning' Empirical Data 135
Method 138
Results 139

Discussion 143

CHAPTER SIX
SUMMARY 147

Summary of Key Findings 149

Implications 151
Identify and Remove Careless Responders 152
Model the Wording Effects Explicitly 154

Areas for Future Research 155

REFERENCES 159

APPENDICES:
A. Python Program Syntax Written for this Research 173
B. MAP and One-Factor Model Fit Indices for 5-Point Scale 195
C. Phi Values for Samples at Each Level of Carelessness 200

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

LIST OF TABLES

Table 1	*The Core Self-Evaluations Scale (CSES)*	57
Table 2	*Rotated Factor Loadings (Two-Factors Extracted)*	75
Table 3	*Fit Indices for Alternative Models of Negative Item Wording Effects in the CSE Scale*	76
Table 4	*Standardised CFA Parameter Estimates for Alternative CSE Scale Models*	79
Table 5	*Fit Indices for Alternative Structural Models Incorporating CSE Scale*	84
Table 6	*Most Influential Cases in the Two Factor Model (Ranked in Descending Order of Phi)*	99
Table 7	*Most Influential Cases in the Two Factor Model (Ranked in Descending Order of LX(2,1) and Ascending Order of LX(2,2)*	101
Table 8	*Most Influential Cases in the Two Factor Model (Ranked in Descending Order of LX(4,1) and Ascending Order of LX(4,2)*	102
Table 9	*Most Influential Cases in the Two Factor Model (Ranked in Descending Order of LX(6,1) and Ascending Order of LX(6,2))*	103
Table 10	*Most Influential Cases in the Two Factor Model (Ranked in Descending Order of LX(8,1) and Ascending Order of LX(8,2))*	104
Table 11	*Most Influential Cases in the Two Factor Model (Ranked in Descending Order of LX(10,1) and Ascending Order of LX(10,2))*	105
Table 12	*Most Influential Cases in the Two Factor Model (Ranked in Descending Order of LX(12,1) and Ascending Order of LX(12,2))*	106
Table 13	*Thresholds used to generate skewed distributions*	112
Table 14	*Percentage of Samples for which Velicer's MAP Indicated a Single Factor Underlying the Interitem Correlation Matrix (7-Point Response Scale)*	117
Table 15	*Percentage of Non-Significant Satorra-Bentler Chi-square Values for a One-Factor Model Fitted to Samples with Varying Levels of Careless Responding (7-Point Response Scale)*	121
Table 16	*SRMR Values Averaged Over 200 Samples for One-Factor Model Fitted to Samples with Varying Levels of Careless Responding (7-Point Response Scale)*	123
Table 17	*NNFI Values for One-Factor Model Fitted to Samples with Varying Levels of Careless Responding*	124
Table 18	*Fit Indices for One-Factor Model Fitted to Sample 3 with Varying Percentages of Outlying Cases Trimmed from Sample*	140

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

LIST OF FIGURES

Figure 1.	Alternative models of the Core Self-Evaluations Scale. Error variables omitted for clarity. Models 1 and 2 incorporate core self-evaluations constructs only. Models 3, 4 and 5 model method (item wording) effects through latent constructs. Models 6 and 7 model method effects through correlated errors.	67
Figure 2.	Structural models comparing one- and two-factor representations of the Core Self-Evaluations Scale (error variables omitted for clarity).	71
Figure 3.	Percentage of samples for which Velicer's MAP indicated a single factor underlying the inter-item correlation matrix (7-point response scale; percentage in each cell based on 200 samples)	118
Figure 4.	Percentage of samples with non-significant Satorra-Bentler chi-square (7-point response scale; percentage in each cell based on 200 samples)	122
Figure 5.	NNFI values for one-factor model averaged over 200 samples, for different levels of careless responding.	125
Figure 6.	Boxplots of mean correlation (phi) between positive item and negative item scales for each level of carelessness in the simulated data.	133
Figure 7.	Histograms of mean correlation (phi) between positive item and negative item scales for jackknife samples at two level of carelessness in the simulated data.	134
Figure 8.	Histogram of mean correlation (phi) between positive item and negative item scales for jackknife samples from Sample 3 data.	135
Figure 9.	Histogram showing frequency of careless (shaded) and non-careless cases; cases ordered according to decreasing correlation between positive and negative item scales.	137
Figure 10.	Plot of mean standardised residuals for same-keyed and differently-keyed item pairs (Sample 3 data, with indicated percentages of outliers removed).	141

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

ABSTRACT

Valid measurement is fundamental to the generation of knowledge, and summated rating scales are the dominant measurement format in I/O psychology and organisational behaviour. Users of these scales have traditionally been advised to include both positively and negatively keyed items, but this practice can give rise to an artifactual wording factor. With increasing use of confirmatory factor analysis and structural equation modelling, this problem has assumed greater significance. The presence of such factors can destroy model fit, and seriously mislead researchers.

I demonstrate the ongoing relevance of this problem by showing how it affects a recently published unidimensional scale, Judge et al.'s (2003) Core Self-Evaluations Scale (CSES). Using four diverse samples from different countries, I show that a one factor model cannot be fitted to CSES data unless wording effects are accounted for.

Having identified the presence of an item wording factor in the CSES samples, I use these data to identify patterns of responding consistent with carelessness. Carelessness has been identified as an important possible explanation for the emergence of item wording effects, but past simulations of careless responding have unrealistically assumed that careless respondents are careless when answering all negatively worded items.

Using the patterns identified in empirical data, I create a realistic model of careless responding by varying both the number of careless respondents, and the proportion of items each such respondent answers carelessly. I then use this Monte Carlo simulation to identify the effect of careless responding on factor structure and model fit in a CFA framework, varying item skew, inter-item correlation, and the number of item rating points. A wording factor emerges at relatively low levels of

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

careless responding, and the effect is slightly more marked in the simulated narrow construct scale (with higher average inter-item correlations). Importantly, high levels of carelessness create an item wording factor in items whose distributions were sampled from a normal population. Thus, even if researchers expect an item to have a normal response distribution, sample variation in actual skew levels can result in an item wording factor.

Finally, I use Rensvold and Cheung's (1999) jackknife technique to demonstrate that extreme careless cases in the Monte Carlo simulation can be reliably identified. I apply this finding to empirical data, illustrating how removal of relatively small numbers of influential cases can markedly increase the correlation between positive-item and negative-item subscales, significantly reduce chi-square, and achieve acceptable fit for a single factor CFA model. This ability to remove the artifactual wording factor in balanced scales by identifying careless respondents has important benefits for both scale developers and scale users.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

CHAPTER ONE

INTRODUCTION

In 1882, the Dutch physicist Heike Kamerlingh Onnes, arguing for the importance of measurement in science, proclaimed a desire to have the motto "By measurement to knowledge [door meten tot weten]" inscribed above the entrance to every laboratory (Kamerlingh Onnes, 1983). The first applied psychology laboratory had been established only three years previously, by Wilhelm Wundt at the University of Leipzig (Fuchs & Milar, 2002), and measurement of mental processes was a central concern of Wundt and his colleagues. Measurement continues to be of critical importance today (Haig & Borsboom, 2008).

In I/O psychology and organisational behaviour, variations on the summated rating scale first developed by Likert (1932) have become the dominant measurement technology. The bulk of our knowledge in core topics such as leadership, motivation, job satisfaction, and organisational commitment is based on the interpretation of data gathered by use of rating scales. The quality and veracity of this knowledge is critically dependent on the quality of rating scales employed—"By measurement to knowledge." Indeed, Hinkin (1998) has stated that the "adequate measurement of abstract constructs [using survey questionnaires] is perhaps the greatest challenge to understanding the behaviour of people in organizations" (p. 104).

Not surprisingly, there exists a large literature on 'best practice' techniques for scale development (recent examples include Clark & Watson, 1995; Hinkin, 1998; Simms, 2008; Spector, 1992; Worthington & Whittaker, 2006). While there is considerable agreement on the steps to be followed in developing and validating scales, one area of controversy relates to the use of negatively-keyed items.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Rationale for Using Negatively-Keyed Items

In his seminal work on the construction of attitude scales, Likert (1932) argued for the use of both negative and positive statements. He stated that this was necessary in order to avoid error (e.g., resulting from stereotypical responses to individual items). This recommendation was followed by early adopters of the Likert technique (e.g., Baumgardner, 1935). The use of balanced items assumed increased importance when Cronbach (1941, 1946) publicised the existence of response sets – the tendency for a person to respond to an item based on the form of the item, rather than its content.

Subsequent writers (Jackson & Messick, 1958; Rorer, 1965) drew a distinction between 'response styles' (the pattern of content-independent responding Cronbach had referred to as a 'set') and 'response sets' (the tendency to respond to item content in a way designed to create a certain impression – e.g. to 'fake good' on a personality inventory). The use of both positive and negative items was seen as a means for reducing the effect of response styles – as Moser and Kalton (1971) explained, "variation between positive and negative items forces the respondent to consider each item carefully, rather than to respond automatically to them all in the same way," (p. 362).

The identification of response styles served to further strengthen the practice of including both positive and negative items in scales. Ironically, the concern about response styles turned out to have been overstated. Rorer (1965) produced a strong critique of the existing evidence for acquiescent responding, and the following debate between his critics (e.g., Bentler, Jackson, & Messick, 1971) and supporters (e.g., Block, 1971) resulted in general acceptance that "most of the concerns about self-

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

inventories being dominated by response styles have proved to be largely false alarms" (Nunnally, 1978, p. 559).

Despite this, use of a balanced set of positive and negative items continued to be recommended as best practice in scale construction. For example, Nunnally (1978) recommended that "the pool of items should be about evenly divided between positive and negative statements" (p. 605) and Spector (1992) included the injunction: "Use both positively and negatively worded items" (p. 24). Of the 159 published organisational survey and research scales summarised by the British Telecom Occupational Psychology Unit (1984a, 1984b), 60% included at least one negatively-keyed item. Among this subset of scales, the modal percentage of reversed items was 50%.

Problems Introduced by Negatively-Keyed Items

Unfortunately, increasing numbers of studies began to identify problems with the use of negative items in scales. Schriesheim and colleagues (Schriesheim, Eisenbach, & Hill, 1991; Schriesheim & Hill, 1981; Schriesheim & Kerr, 1974) carried out a number of studies centred around the Leader Behavior Description Questionnaire (LBDQ; Stogdill, 1963). They concluded that positive items had higher validity than items involving negation. While both positive items and simple negations had similarly high reliability, polar opposites and negated polar opposites had unacceptably low alphas. Schriesheim et al. (1991) recommended that polar opposite items (and their negations) should not be used in scales because of their adverse effect on reliability and validity.

Other studies have also found differences in psychometric properties between positive and negative items. Ahlawat (1985) found that negation (of either positive or

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

negative items) created problems for his young respondents, using the State-Trait Anxiety Inventory (Spielberger, Vagg, Barker, Donham, & Westberry, 1980). Barnette (1997, 2000) compared a positive-item version of a 20-item attitude scale with versions containing a mix of positive and negative items. He found that versions of the scale which contained negative items had markedly lower reliabilities than the positive item version. The various studies are not always in agreement – for example, Ahlawat found lower reliability for negated items, whereas Schriesheim et al. (1991) found that negated (as opposed to polar opposite) items had the same reliability as standard (positively keyed) items. Overall, however, the evidence suggested the need for a more critical stance towards use of negatively keyed items.

A more consistent set of findings holds in regard to the effect of negatively keyed items on scale factor structure. The publication of a study by Carmines and Zeller (1979) focused attention on the influence of item wording on scale factor structures. Carmines and Zeller demonstrated that positive and negative items in the Rosenberg Self Esteem Scale (Rosenberg, 1965) loaded on separate factors, but that these factors each had the same pattern of correlations with other variables. They concluded that the factor structure was a function of response set, acting differently on the positively and negatively worded items. In effect, the scale comprised a single substantive (self esteem) factor, and a factor representing systematic error variance. Subsequent studies confirmed and extended this finding (e.g., Bachman & O'Malley, 1986; Corwyn, 2000; Greenberger, Chen, Dmitrieva, & Farruggia, 2003; Marsh, 1986; Tomás & Oliver, 1999)

While early work focused on Rosenberg's self esteem scale, item wording factors have been identified in a wide range of scales. Examples include measures of job characteristics (Idaszak & Drasgow, 1987), loneliness (Knight, Chisholm, Marsh,

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

& Godfrey, 1988; Miller & Cleary, 1993), leadership (Schriesheim & Eisenbach, 1995), organisational commitment (Magazine, Williams, & Williams, 1996), and union commitment (Bayazit, Hammer, & Wazeter, 2004; Kelloway, Catano, & Southwell, 1992).

Some studies have suggested the possibility of the item wording factor being more than just a method artefact. For example, Quilty, Oakman, and Risko (2006) found significant correlations between a negative item wording factor and several personality variables. However, most of the researchers cited have concluded that item wording effects lack any substantive content, and are methodological artefacts.

The effect seems to be primarily a function of negatively-worded items; modelling the factor by allowing negative item errors to covary, or by specifying a factor underlying negatively worded items, generally results in better fit than modelling it on positive items (see, for example, Aluja, Rolland, García, & Rossier, 2007; Corwyn, 2000; Gana, Alaphilippe, & Bailly, 2005; Horan, DiStefano, & Motl, 2003; Marsh, 1996; Tomás & Oliver, 1999). There is considerable evidence that negative sentences are harder to process than affirmative sentences (Kaup, Zwaan, & Lüdtke, 2007). People take longer to verify negative statements than positive statements, and performance on such tasks is correlated with verbal ability (Lansman, cited in Hunt, 1978). This suggests the possibility that verbal ability may interact with item wording to create the item wording factor.

Marsh (1986, 1996) tested this possibility. In his 1986 study, he compared the means of positive and negative items on a 66-item scale, and found the correlation between them to increase with age. This suggested (along with evidence from item means and variance) that younger children were more likely to endorse negative items in a manner inconsistent with their overall positive self-concept. In another sample, a

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

negative item factor correlated significantly with reading ability. His 1996 study demonstrated increasing correlations between positive and negative item means as reading ability increased, suggesting that higher reading ability resulted in greater consistency of responses across both item types. Other studies have reached similar conclusions (e.g., Dunbar, Ford, Hunt, & Der, 2000; Weems, Onwuegbuzie, & Collins, 2006).

In contrast, Cordery and Sevastos (1993) found no effect of education level on item wording factors using a large (N = 3,400) adult sample. They noted that education level could be confounded with motivation to respond accurately, and concluded that item wording effects were "a function of general care exercised in responding to negatively worded items" (p. 143). Careless or confused responding has long been recognised as a potential source of error in questionnaire responses. Nunnally (1967) noted that differences in the degree of carelessness between different subgroups (e.g., adults, children) can differentially affect the score distributions of subgroups; such differences can generate spurious factors (Bernstein & Teng, 1989).

Careless or confused responding can result from misunderstanding items, emotional reactions, a desire to finish the task quickly, or fatigue (Barnette, 1995). Barnette found evidence in two large samples (totalling over 6,000 respondents) that as many as 10% of respondents may not have attended fully to the reverse wording of items.

Schmitt and Stults (1985) investigated whether careless responses to negatively keyed items could account for the emergence of a method factor associated with item wording. Using simulated data, they identified a component associated with negatively keyed items when as few as 10% of respondents were careless, and the prominence of the factor increased with increasing numbers of negative items in the scale.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Woods (2006) used a confirmatory factor analysis framework to investigate carelessness. She found the fit of a one-factor model began to decline markedly as the percentage of careless respondents reached 10%. However, she used dichotomous items – a format which is very rare in organisational studies. Furthermore, she modelled careless responding in the same way as Schmitt and Stults – with careless respondents answering all negatively keyed items carelessly. Woods noted this as a limitation, and stated the need for future research to investigate more plausible scenarios, whereby respondents answer carelessly in various degrees.

This brief summary of research has highlighted the problem of spurious item wording factors. Over recent years, the existence of such wording factors has taken on much greater importance. Variance due to item wording will not prevent a researcher from calculating a correlation or regression coefficient, but it may invalidate a structural equation or confirmatory factor analysis model. Advances in quantitative methodology have led to increasing adoption of such latent variable approaches, and they are now the preferred approach for many research problems. Inability to fit a single factor measurement model to a purportedly unidimensional scale has important ramifications – especially if the researcher is not aware of the potential existence of a wording factor.

Wording Factors and Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) is a powerful technique which can be used to evaluate psychometric properties of scales (e.g., dimensionality and item-factor relationships), construct validity, the existence of method effects, and measurement invariance across time or different populations. For these reasons, it has become "one of the most commonly used statistical procedures in applied research" (T. A. Brown,

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

2006, p. 1). Because of this, it is critical that we develop a better understanding of the way in which careless responding can affect the fit of CFA models.

Firstly, we need evidence based on a more plausible model of careless responding. It is unrealistic to assume that a careless respondent is careless with all items. In fact, such extreme carelessness is generally easy to identify by visual inspection and sorting of datasets, and such cases can be removed. We currently have no studies which have tried to identify individual patterns of carelessness in empirical data, or which have used realistic models of carelessness.

Secondly, we need to use more objective means for determining the number of factors. The 'eigenvalue > 1' criterion is not ideal, representing an upper bound for the number of factors rather than an accurate estimate (Preacher & MacCallum, 2003). Alternatives such as parallel analysis and MAP are now available, and published macros make it possible to use them with major statistical software packages (O'Connor, 2000).

Thirdly, we need to understand the effect of careless responding to Likert-type scales on the fit of CFA models. The study by Woods (2006) guides us in terms of dichotomous items, but analysis of 5 or 7 point rating scales is needed.

Fourthly, techniques appropriate to the analysis of item-level ordinal data need to be used. Most past CFA studies of wording factors have employed estimation techniques designed for use with continuous data which can lead to biased estimates and errors (Jöreskog, 2002).

Fifthly, we need to determine whether the levels of skew typically found in organisational measures (Micceri, 1989) influence the emergence and effects of wording factors resulting from careless responding.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Finally, we need to evaluate possible techniques for identifying careless cases in empirical data, and the feasibility of removing such cases in order to improve the validity of our measures.

Overview of Dissertation

In this dissertation, I seek to address these gaps in our current knowledge.

Following a more detailed review and critique of relevant research (Chapter 2), I begin by analysing a recently published measure, the Core Self-Evaluations Scale (CSES; Judge et al., 2003). This scale comprises balanced numbers of positive and negative items, and I use it to illustrate the ongoing relevance of the item wording factor problem. I evaluate the effect of this factor on CFA model fit in Chapter 3, using four different samples, and demonstrate how serious a problem it is.

In Chapter 4 I use the CSES samples to identify inconsistencies in responses to positive and negative items by individual cases. Nunnally (1978) suggests that such inconsistencies can identify respondents who may be answering carelessly. I demonstrate the high degree of variability in inconsistent responding, and the need for a more complex simulation of careless responding than has been used in studies to date. On the basis of this analysis, I design a realistic Monte Carlo simulation of careless responding. I generate data samples representing a wide range of carelessness levels, and evaluate the factor structure and model fit of these samples for different levels of skew and inter-item correlation.

In Chapter 5 I apply a technique for identifying influential cases in structural equation models (Rensvold & Cheung, 1999) to determine whether it is possible to identify careless cases in the simulated samples. Based on this analysis, I apply the

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

technique to empirical data samples, demonstrating its efficacy in improving model fit by removing comparatively low numbers of inconsistently responding cases.

Chapter 6 provides an overall summary and discussion of findings and their implications.

Contribution

In summary, this dissertation contributes to our current understanding by being the first study to comprehensively look at careless responding as the primary reason for emergence of artificial wording factors by:

1. Modelling carelessness realistically. Previous studies have assumed that a careless respondent answers all negatively keyed items carelessly. This type of 'all or nothing' carelessness is relatively easy to identify in datasets, and its use in simulations can exaggerate the effect of careless responding.

2. Using a confirmatory factor analysis framework to identify effects of carelessness on factor structure and model fit of a rating scale. Previous simulations of careless responding have used principal components analysis, or have considered only dichotomous responses.

3. Using defensible techniques (parallel analysis, MAP) to estimate the number of factors explaining inter-item correlations.

4. Exploring the emergence of artifactual factors resulting from carelessness in scales whose item response distributions have been drawn from a normal population (as well as from skewed distributions).

5. Presenting a technique for estimating the amount of careless responding in empirical data, and for identifying cases whose inconsistent responding is contributing most to the creation of a spurious wording factor.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

CHAPTER TWO

LITERATURE REVIEW

This dissertation focuses on undesirable side effects resulting from the use of negatively keyed items in summated rating scales. I begin this chapter by highlighting the significant breakthrough represented by Likert's (1932) development of a simple method for measuring attitudes, and the rapid adoption of his injunction to use both positively and negatively keyed items. I describe how concerns over acquiescent responding reinforced the practice of using both types of items, even though these concerns later turned out to have been overstated. Following a brief summary of research into the psychometric properties of negatively keyed items, I present a detailed review of their influence on scale factor structure. Much of this research has been carried out on Rosenberg's Self Esteem Scale (RSES; Rosenberg, 1965), and I use this research stream to examine the main controversies and methodological issues posed by item wording factors. I finish the chapter by reviewing the various explanations put forward for the emergence of wording factors, and present arguments for the need to examine the role of careless responding in more depth.

Likert's Technique for Attitude Scaling

Research into selection of pilots for the United States Air Service during World War I found that emotional stability correlated most highly with individual measures of flying ability (Henmon, 1919). Emotional stability was measured by discharging a pistol near an unsuspecting applicant, and recording the amplitude of hand tremors and changes in breathing caused by this sudden shock.

In contrast, mental ability was being reliably measured that year by administering paper and pencil questionnaires to large groups of candidates – 200,000

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

draftees a month were undertaking the army alpha test in May 1918 (Kevles, 1968). The test took less than an hour, and was marked within the same day (Yerkes, 1917, cited by Kevles).

In the introduction to their article on the measurement of personality, Allport and Allport (1921) commented favourably on progress made in developing scales for measuring intelligence, and contrasted this with the challenges faced by personality researchers. At the same time, Pressey (1921) was seeking to develop a group test to measure aspects of personality as efficiently in time and labour as existing measures of mental ability.

In the first issue of the Harvard Business Review, Starch (1922) extolled the merits of ability testing, counselled executives away from the services of 'character analysts' (phrenologists, physiognomists, palmists and graphologists), and bemoaned the fact that psychologists had "as yet not devised any satisfactory method of measuring … in an objective way, … qualities of personality or makeup … and attitudes towards superiors or inferiors and toward fellow-workers" (p. 79).

There was clearly enormous demand for a reliable, straightforward and systematic means by which to measure personality, temperament and attitudes. Thurstone made the first significant advance, with his standard technique for developing attitude scales, based on principles from psychophysical theory (Thurstone, 1928; Thurstone & Chave, 1929). Although effective, the technique was cumbersome; Likert (1932) described it as "exceedingly laborious" (p. 6), and proposed a simpler alternative. While there was some initial debate about whether the Likert approach saved time compared with Thurstone scaling (Bird, as cited in A. L. Edwards & Kenney, 1946), Edwards and Kenney concluded that on the available evidence, the Likert approach was "less time-consuming and less laborious than the Thurstone

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

technique" (pp. 82–83). At last researchers had a technique for measuring attitudes with the same ease and efficiency as for mental abilities. The Likert approach subsequently went on to become "what is today the most widely used methodology for attitude scale construction" (Bottom, 2006, p. 17).

In the appendix to his monograph, Likert (1932) presented five criteria for selecting statements for inclusion in scales. One criterion stated the desirability of wording statements so that:

> about one-half of them have one end of the attitude continuum corresponding to the *left* or *upper* part of the reaction alternatives and the other half have the same end of the attitude continuum corresponding to the *right* or *lower* part of the reaction alternatives .... These two kinds of statements ought to be distributed throughout the attitude test in a chance or haphazard manner (p. 46).

Likert (1932) introduced this requirement to guard against two potential problems. Firstly, he was concerned about 'space error'. This stems from psychophysical studies, where the relative positions of a stimulus and a referent could introduce constant error (Masin & Agostini, 1991). For example, weight judgments using the right and left hands might differ as a function of handedness, while line length judgments might differ depending on whether the reference line is to the left or right of the stimulus because of a preference for scanning from left to right based on the usual order of reading or writing. Lewin did not expand on the meaning of this type of error in the context of his scale. His method did not require any explicit comparison of statements – each was presented as an item in its own right. However, he presumably felt that using both types of items would balance out any constant error caused by a preference for one end of the scale, or a preference for one extreme of the measured attitude.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

His second concern was to avoid stereotyping. At the outset of his project, Likert (1932) believed that individual items would tend to measure specific attitudes. For example, while he grouped twelve items into an 'Imperialism' scale, he expected these to measure twelve different attitudes. With this belief, Likert wanted to ensure that responses to individual items reflected these individual attitudes; by mixing positively and negatively keyed items, the tendency to respond in a stereotyped way would be reduced.

Researchers adopting Likert's (1932) approach to attitude scaling also adopted his recommendations for wording items, including the advice to include items at both poles of the attitude. Baumgardner (1935) for example, described Likert's criteria as "axiomatic concepts [to] be kept in mind .... recognised by all workers who have been successful in attitude measurements" (p. 491). Thus, balanced use of positively and negatively keyed items has been an integral part of developing scales using the "method of summated ratings" (Bird, 1940, p. 159) since their inception.

Acquiescent Responding

The use of both positively and negatively keyed items assumed greater importance with the identification of response sets. It had been known for some time that students answering true-false tests tended to guess 'true' much more often than 'false' when unsure of an answer (e.g., see Fritz, 1927). Lorge (1937) suggested that this might reflect a personality trait. Cronbach (1941) explored this possibility, introducing the term 'acquiescent response set' to describe the tendency to guess 'true' more often than 'false' when uncertain. To the extent that this tendency reflected an underlying personality trait (perhaps a tendency towards uncritical thinking), the

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

resulting score would "conceal personality tendencies within an alleged measure of knowledge [and be] apt to mislead both teacher and student" (Cronbach, 1942, p. 401).

The same could be argued for attitude scales – if respondents uncertain about their standing on an item tended to agree more often than disagree, the resulting score would be biased. By having a balanced set of positively and negatively keyed items, the effect of acquiescence would tend to be cancelled out, preventing such respondents from getting extreme scores. While the scores would still contain error, the use of balanced items would result in "far less damage to estimates of means and standard deviations and results of statistical tests" (Spector, 1992, p. 25).

Acquiescence was also implicated by some researchers in the emergence of factors related to item keying. For example, Jackson and Messick (1958) reviewed factor analytic studies of the MMPI. Such studies had typically found that two factors accounted for much of the variance (either at item level, or as second order factors), and that these two factors were largely distinguished by the direction of item keying. Jackson and Messick concluded that acquiescence was interacting with item content (or a response style) to influence the factor structure.

Rorer (1965) reviewed twenty years of research into acquiescence. He considered a range of studies, covering correlations between measures of acquiescence on different tests, correlations between acquiescence scores and non-personality variables (verbal ability; intelligence), and correlations between acquiescence scale scores and observations of acquiescent behaviour. In summary, he concluded that, when forced to guess, respondents answer in a non-random manner, but that no evidence existed to suggest they were forced to guess on non-ability questionnaires (e.g., personality or attitude measures). Content rather than response style was the major determinant of item responses.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Rorer's critique convinced most psychologists that response style was no longer an important issue, and that responses to items could be taken as based primarily on content, albeit subject to influence by response set (Nunnally, 1967).

Interestingly, although concerns about the impact of acquiescent responding had largely abated, use of a balanced set of positive and negative items was still considered best practice in scale construction. Nunnally (1978) recommended that "the pool of items should be about evenly divided between positive and negative statements" (p. 605), while in Europe, Moser and Kalton (1971) counselled scale developers "to have a roughly equal number of positively and negatively worded items in the scale" (p. 362). This advice extends to the present day, with popular guides to scale development including recommendations such as: "Use both positively and negatively worded items" (Spector, 1992, p. 24).

Effectively, the position remained little changed from 1932 when Likert first presented the summated scale development technique. Indeed, Moser and Kalton's (1971) advice ("variation between positive and negative items forces the respondent to consider each item carefully, rather than to respond automatically to them all in the same way," p. 362) echoed Likert's concern with stereotyped responses. However, the series of studies into response styles and sets identified additional reasons to be wary when using positively and negatively worded items. The solution to one problem (response styles) may have generated new concerns – for example, factors related to item keying (Jackson & Messick, 1958) and degraded scale psychometrics (Bentler et al., 1971).

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Problems Introduced by Negatively-Keyed Items

In the earlier discussion of acquiescence research, I discussed evidence that positively and negatively keyed items defined separate scales in factor analysis of inventories such as the MMPI (Jackson & Messick, 1958). This section will cover this issue in more detail, together with research confirming additional psychometric concerns (impact on validity and reliability) relating to the use of negatively keyed items in scales.

The use of negatively keyed items assumes that these items are measuring the same underlying construct as their positively worded counterparts. Researchers studying acquiescence attempted to create items which were perfect opposites of each other in order to disentangle the effects of acquiescence from the content-driven component of item responses (Rorer, 1965; Samelson & Yates, 1967). If two positive items correlate .70, then a negatively keyed version of one of these items should correlate -.70 with the other positive item (and -1 with its positive counterpart). While useful for research purposes, a set of perfectly opposed items would create a scale with excessive redundancy, and an artificially inflated reliability (Loevinger, 1954). In a scale used for applied purposes, each item (positive or negative) should make a unique contribution to measurement of the construct domain. On average, however, the correlations between positive items (or between negative items) should be no larger than the correlation (absolute value) between pairs of positive and negative items.

In practice, it proved difficult to form perfect opposites. As Samelson and Yates (1967) pointed out, many item reversals in acquiescence studies resulted in item pairs whose neutral points (the point of transition from agreement to disagreement) differed in regard to the underlying construct. Some studies used quite extreme reversals (e.g., Jackson & Messick, 1957) while others used more moderate reversals

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

(e.g., Peabody, 1961). Clearly, if the set of negative items differ in their extremity from that of the positive items, correlations within the two sets are likely to be greater than correlations between positive and negative items. This clustering may lead to separate factors based on item keying, and to differential means, standard deviations, reliability and validity of the two sets of items.

Psychometric Problems Associated with Negatively Worded Items

A number of studies have compared the properties of positive and negative items within a scale. Schriesheim and Hill (1981) had undergraduate students use modified versions of the Leader Behavior Description Questionnaire (LBDQ; Stogdill, 1963) scales to rate imaginary leaders (described in scripts). One version comprised all positive items, another all negative, and the third had a mix of the two item types. The positive item version of the questionnaire was found to be more accurate than either the negative or the mixed item versions.

Schriesheim, Eisenbach, and Hill (1991) compared the reliability and validity of different forms of negatively worded items with original (positively worded) items from the LBDQ-XII. Three forms of negation were used—simple negation, polar opposites, and negated polar opposites. They found that positive items (e.g., "He gives group members precise task assignments") and negated versions (e.g., "He does not give group members precise task assignments") both had high reliability (alpha > .8). In contrast, polar opposite items (e.g., "He gives group members vague task assignments") and negated polar opposites (e.g. "He does not give group members vague task assignments") had alphas less than .6. The polar and negated polar items were also less accurate when compared with the criterion (a scenario description of a leader's behaviour). Schriesheim at al. concluded that polar opposite items (and their

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

negations) should not be used in scales because of their adverse effect on scale reliability and validity.

Barnette (1997, 2000) compared responses to six versions of a 20-item survey designed to measure attitudes towards year-round schooling. He found that the survey versions containing negated items had lower reliabilities (.71 to .73 in the 2000 study) than the versions containing only positively worded items (.81 to .82). This finding differs from that of Schriesheim et al., (1991) who found that negated (as opposed to polar opposite) items had the same reliability as standard (positively keyed) items.

Benson and Hocevar (1985) created a 15-item scale designed to measure the attitude of children (4^{th} to 6^{th} grade) towards school integration (the subjects were all students in schools subject to court-ordered integration) together with a parallel version comprising negated versions of the items and a third version with a mix of positive and negative items. On the basis of differences in item means and variances, they concluded that their elementary school respondents "were less likely to indicate agreement by disagreeing with a negatively phrased item than to indicate agreement by agreeing with a positively phrased item" (p. 237).

Benson and Hocevar (1985) analysed the third version of the scale to identify whether the eight positively phrased items loaded onto the same factor as the seven negatively phrased items. They found that a single factor model demonstrated significantly worse fit than a two factor model defined by item wording (chi-square difference = 52.36, df = 1, $p < .05$). They concluded that the two types of items were measuring different constructs. They state that "the estimated correlation between the two factors was zero" (p. 236), but this seems highly unlikely given the common broad domain tapped by the items, and the factor loadings of all items on the single factor

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

models summarised in Table 2. No value (or *p* value) for the correlation between factors is given in the statistical tables.

In spite of some shortcomings identified in the preceding studies, they did serve to maintain focus on psychometric concerns associated with use of negatively-worded items. In addition, the study by Benson and Hocevar (1985) added weight to earlier observations (e.g., Jackson & Messick, 1958) that positive and negative items can define separate factors in a purportedly homogeneous scale. Studies investigating the existence of such factors will now be reviewed.

Item Wording Factors

In the appendix to their book on reliability and validity issues in measurement theory, Carmines and Zeller (1979) discussed the use of factor analysis for assessing construct validity of scales. They emphasised the importance of interpreting factor analytical results in the context of theory, and illustrated this with data from Rosenberg's (1965) measure of self-esteem (RSES). While Rosenberg conceptualised this as a unidimensional measure, factor analysis revealed a two-factor solution with (in the varimax rotated solution) positive items loading on one factor and negative items loading on the second. Carmines and Zeller calculated separate scale scores for positive and negative items, and correlated these with a variety of relevant external variables. There were no statistically significant differences in the correlations between the two scales and the other variables. On the basis of this construct validity evidence, the authors concluded that the factor structure was a function of response set, operating differently on the positively and negatively worded items. In essence, the second factor (in the unrotated solution) represented systematic error variance.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

The Rosenberg scale (and variations of it) became the subject of many subsequent studies aimed at clarifying its factor structure. In order to give a sense of how studies into item wording factors developed over time, I will review the program of research on this instrument before considering work with other scales.

RSES Wording Factor as Artefact

Marsh (1996) analysed self-esteem data from the National Educational Longitudinal Study (NELS), a large-sample longitudinal database, using items derived from the Rosenberg Self Esteem Scale. His aim was to determine whether prior researchers (e.g., Bachman & O'Malley, 1986; Carmines & Zeller, 1979) were justified in dismissing the second factor as a methodological artefact. For example, Kaufman, Rasinski, Lee, & West (1991) had suggested that the two observed factors might represent two substantive factors – a general evaluation of oneself (all four positive items together with one negative item), and transient self-evaluations (the remaining two negative items). Kaufman et al. argued that the subtlety of the distinction between their two factors meant that the correlation between them would decrease with increasing reading ability. In contrast, Marsh hypothesised that younger and less verbally able students would find it harder to respond to negative items, resulting in a more strongly defined second factor (if the factor was a spurious artefact of item wording). Thus, an increased correlation between the two factors with increased reading ability would support the methodological artefact interpretation, while a decreased correlation would support Kaufman et al.'s model.

Marsh (1996) evaluated six different models within a confirmatory factor analysis framework: Model 1 (a single latent self esteem factor); Model 2 (Kaufman et al.'s two factor model); Model 3 (one latent for positive items, a second for negative

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

items); Model 4 (a single self esteem latent factor, with correlated uniquenesses among negative items); Model 5 (a single self esteem latent factor, with correlated uniquenesses among positive items); and Model 6 (correlated uniquenesses between one pair of positive items was added to Model 4; not all positive uniquenesses could covary due to identification constraints). Model 1 had poor fit to the data, while the models with correlated uniquenesses had the best fit. Correlated uniquenesses among negative items (Model 4) resulted in better fit than for positive items (Model 5). Model 6 had the best fit of all. Marsh concluded that the items measured a single construct, together with "substantively irrelevant method effects" (p. 815). This was further supported in his subsequent assessment of the effect of reading ability on the correlation between the two factors in Models 2 and 3.

Tomás and Oliver (1999) used a Spanish language version of Rosenberg's complete scale, administered to a sample of 640 high school students (average age 15.8) in Valencia. Given the centrality of linguistic issues in emergence of item wording factors, use of a Spanish version would test whether the phenomenon generalised to other languages. Nine different CFA models were tested. Five of these were identical to the first five models in Marsh (1996), as described in the previous paragraph. For their sixth model, Tomás and Oliver allowed all positive item uniquenesses to covary, and all negative item uniquenesses to covary. The remaining models were: Model 7 (one self esteem latent for all items; one method factor for negative items); Model 8 (one self esteem latent for all items; one method factor for positive items); and Model 9 (one self esteem latent for all items; two correlated method factors, for positive and negative items respectively). The single factor model had poor fit, as did the positive and negative self-esteem factor model. Models with method effects modelled on negative items (either covarying uniquenesses or a

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

separate method factor) had good fit (better than modelling method effects on positive items). Modelling both positive and negative item method effects resulted in the best fit of all. Tomás and Oliver concluded that the Rosenberg scale measured a single global self esteem factor, but it was necessary to include method effects to achieve acceptable model fit. In their discussion, however, they raised the possibility that this wording effect might measure something substantively important, and argued for more investigation (relating the factors to other trait measures).

Corwyn (2000) essentially replicated the work of Marsh (1996) and Tomás and Oliver (1999), but did so with the full Rosenberg scale (Marsh had used a shorter version) and with United States students using an English language scale (whereas Tomás and Oliver had used a Spanish version). Corwyn found similar results, with inclusion of method effects leading to increased fit, and stronger method effects for negative items. Corwyn also found that correlation between positive and negative self esteem factors increased with increased verbal ability – a finding consistent with Marsh's conceptualisation of the second factor as being a method artefact rather than a substantive factor.

Dunbar, Ford, Hunt, and Der (2000) reached a very similar conclusion, based on the full Rosenberg scale in two Scottish adult samples (cohorts in the West of Scotland Twenty-07 study, Macintyre, 1987) – one comprising 832 adults around 40 years of age, the other 812 adults around 60 years in age. They found that the correlation between the positive item factor and the negative item factor was lower in subjects with lower verbal ability (based on a median split of the older cohort sample using a verbal reasoning test). There was no difference in correlation when the median split was computed on the basis of a numerical reasoning score. In contrast with Marsh(1996) and Corwyn (2000), Dunbar et al. found that the wording effect was

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

more pronounced among positive items than negative items – a model with correlated positive uniquenesses had better fit than with correlated negative uniquenesses in both of the samples (although the difference in fit was quite small).

A number of replicative studies have confirmed the existence of a wording factor in samples from other populations, and have also found the factor to relate more to negatively worded items. Gana, Alaphilippe, and Bailly (2005) used a French language version of the scale (Vallières & Vallerand, 1990) in a sample of 864 French retirees (average age 72.7), concluding that the scale "is a one-dimensional self-esteem instrument, in spite of the presence of method effects" (p. 175). Aluja, Rolland, García, and Rossier (2007) reached the same conclusion in a sample of French students.

Greenberger, Chen, Dmitrieva, and Farruggia (2003) set up a direct test of the extent to which item wording accounts for the emergence of a second factor in the Rosenberg scale. They created positive versions of negative items, and negative versions of positive items, so as to be able to create an all-positive scale and an all-negative scale (together with the original balanced scale). In addition to assessing factorial structure, they investigated the construct validity of the new versions using measures known to relate to overall self esteem scores. The sample consisted of 741 ethnically diverse undergraduates at a United States university; each person completed one version of the scale (randomly assigned). A two-factor model provided a significantly better fit to the original (mixed positive and negative items) scale than a one-factor model. For each of the two new scales (all positive, or all negative items), there was no significant difference between the fit of a one-factor and a two-factor model (although a one-factor model showed better fit than it did in the original scale). Exploratory factor analysis also suggested that the two revised scales each had only

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

one factor. Correlations between each of the three versions of the self-esteem scale and a variety of other variables (construct validity) showed a very similar pattern. Accordingly, Greenberger et al. concluded that all three versions measured the same construct, and that the second factor in the Rosenberg scale was an artefact of item wording direction.

RSES Wording Factor as Substantive Factor

While there now appeared to be general consensus that the second factor in the Rosenberg Self Esteem Scale was a method effect, Horan, DiStefano, and Motl (2003) speculated that it might also have a substantive interpretation. Referring to Bentler et al.'s (1971) conceptualisation of response style as a personality variable or trait, Horan et al. set out to determine whether item wording factors existed across different substantive areas (using scales with a mix of positive and negative items on the same sample of respondents), and whether these factors were related to each other. They also used longitudinal data to assess the stability of such factors over time. Three scales from the National Educational Longitudinal Study (NELS) were studied – a seven-item version of Rosenberg's Self Esteem Scale (with 3 negatively worded items), a ten item measure of attitudes towards school (3 negative items) and a six item locus of control measure (5 negative items). Earlier results were replicated (single factor models fitted poorly; fit improved by modelling wording effects, with greatest improvement coming from modelling a negative item wording effect). Furthermore, these findings applied to the school attitude and locus of control measures, indicating existence of stability over different construct domains.

Horan et al. (2003) also modelled the three measures simultaneously. Wording factors had to be included in order to get acceptable fit. A model with a single wording

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

factor (for all negatively worded items across the three scales) led to improved fit, but better fit was obtained by modelling three correlated wording factors (one for each scale). The three negative item wording factors had significant correlations with each other (ranging from .31 to .43). This correlation across different substantive areas was taken as evidence in support of a personality trait underlying the observed response effects. Furthermore, longitudinal analysis indicated a high degree of stability in the negative wording factor of the self esteem scale over two consecutive 2-year periods (correlations in excess of .8). It was not possible to explicitly test for personality correlates of the negative wording factor, as no such measures were available in the NELS database. However Horan et al.'s findings suggest the possibility that item wording factors might reflect an underlying personality trait.

Quilty, Oakman, and Risko (2006) explored this possibility, investigating the relationship between wording factors in the Rosenberg Self Esteem Scale, and several personality measures. They used Carver and White's (1994) BIS/BAS scales to measure approach and avoidance motivation (as related to the two underlying biological processes of the behavioural inhibition system (BIS) and behavioural activation system (BAS)). Ten-item measures of the Big Five personality scales were used to assess emotional stability, extraversion, intellect, agreeableness and conscientiousness (Goldberg, 1999). As in previous studies, Quilty et al. confirmed the existence of item wording factors in the Rosenberg scale. Using a correlated trait correlated method model of the self esteem scale (with method factors for both positive and negative wording factors) they found a significant correlation between avoidance motivation (BIS) and the negative wording factor (r = -.27) meaning that individuals with stronger avoidance motivation were more likely to endorse negatively keyed items.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

With regard to the Big Five factors, Quilty et al. (2006) found significant (albeit small) correlations between the negative wording factor and conscientiousness (.14) and emotional stability (.25). More conscientious, and more emotionally stable individuals, were less likely to endorse negatively keyed items. Noting the correlation between self-esteem and stability, Quilty et al. conclude that the method effect in the self-esteem scale serves to inflate the scores of emotionally stable individuals, exaggerating their actual level of self-esteem. Similarly, the self-esteem of emotionally unstable respondents is likely to be underestimated.

A similar approach was taken by DiStefano and Motl (2006) using a slightly different set of personality measures. They began by replicating Horan et al's (2003) finding that an item wording factor can generalise across different substantive areas, within the same set of respondents. Using the Rosenberg Self Esteem Scale (original ten items) and the Social Physique Anxiety Scale (Hart, Leary, & Rejeski, 1989) they modelled a single negative wording factor underlying all negative items across both scales. While this led to improved fit over a model with no wording factor, even better fit was obtained with a single method factor for each scale. The correlation between these two factors was -.37. (The substantive constructs, self esteem and social physique anxiety, also correlated negatively). DiStefano and Motl concurred with Horan et al. in concluding that the wording factor had the characteristics of a response style, perhaps related to underlying personality traits. On this basis, they went on to investigate the relationship between the Rosenberg Self Esteem Scale negative wording factor and various personality scales—Greenwald and Satow's (1970) short version of the Marlowe–Crowne Social Desirability Scale (Crowne & Marlowe, 1960); Carver and White's (1994) BIS/BAS scales; a short form of the Fear of Negative

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Evaluation scale (Leary, 1983); and a measure of self consciousness (Fenigstein, Scheier, & Buss, 1975).

DiStefano and Motl (2006) found no relationship between the wording factor and social desirability. While the path model had acceptable fit, the path from social desirability to the wording factor was non-significant. Similarly, no significant relationship existed between the BIS/BAS measures and the wording factor – a finding in conflict with Quilty et al. (2006), who found a significant relationship between BIS and the RSES negative wording factor. However, DiStefano and Motl did find significant relationships between the measures of Fear of Negative Evaluation and Self Consciousness and the wording factor. Respondents high on these two constructs were less likely to show the existence of method effects. Furthermore, when summed scores on these two scales, together with the Social Physique Anxiety Scale score, were used to create a latent construct representing concern with evaluations by others, the latent also showed a negative relation to the self esteem item wording factor. The authors speculated that concern with others' evaluations may encourage greater reflection and hence more accurate insights into one's behaviour, making such respondents less susceptible to the influence of item phrasing.

The Source of Item Wording Factors

So far, I have focused on discussion of item wording effects in the Rosenberg Self Esteem Scale. This particular instrument has been at the centre of concern with wording factors, and the research stream illustrates some of the central debates (e.g., as to whether such factors have substantive interpretations, or are mere methodological artefacts). The existence of wording factors has also been demonstrated in other scales, covering a wide range of substantive domains. I will now use some of these studies, as

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

well as references to work on the Rosenberg scale, to review the various processes posited as causes for the emergence of such factors. These include the opposing positions of item wording factors as substantively meaningful, versus item factors as mere method effects, and the role of item response distributions, cognitive ability, item complexity and carelessness in generating such factors.

Item Wording Factors are Meaningful Constructs

Rosenberg presented the early version of his scale as a Guttman scale (Rosenberg, 1962) indicating a belief in its unidimensionality. Kohn and Schooler (1969), in one of the early studies to factor analyse these items (together with other items measuring related constructs) identified two factors which cleanly separated positive and negative items. They did not discuss the possibility that these factors might be an artefact; they labelled the positive item factor 'Self-confidence' and the negative item factor 'Self-deprecation'.

Owens (1993, 1994) also took the position that negative items defined one facet of self esteem, with positive items defining a second facet. He demonstrated the goodness of fit of a two-factor model compared with a single factor model, using a scale comprising ten items (four negatively and six positively worded), similar to Rosenberg's (1965) scale. Owens labeled the negative item scale 'self-deprecation' and the positive item scale 'self-confidence'. He demonstrated differential relationships between these subscales and relevant variables, such as depression and school grades (Owens, 1994). However, he did not explore a model containing a single substantive factor (all items) paired with either correlated uniquenesses or a wording factor to capture method variance. It is therefore not possible to determine whether his

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

conceptualisation results in a better fitting model than one which explicitly models a wording factor, in addition to the substantive latent construct.

Chang (1995) investigated item wording effects in Scheier and Carver's (1985) Life Orientation Test (LOT). This scale comprises eight items measuring dispositional optimism; four of the items describe a pessimistic orientation, and are reverse-scored. Chang created alternative versions of the scale, allowing him to compare factor structures of the scale with all items worded optimistically, all worded pessimistically, or a balanced set of optimistic and pessimistic items. On the basis of confirmatory factor analysis, he concluded that pessimistic and optimistic items each measured different constructs; in all cases, a two-factor model fit the data better than a one-factor model. However, Chang limited his two-factor models to the case where all items describing an optimistic orientation loaded on one factor, and all those describing a pessimistic orientation loaded on the second factor. He did not test a model with a single substantive latent, combined with a method factor underlying one set of connotatively consistent items.

Ang, Neubronner, Oh, and Leong (2006) followed a similar approach to Owens (1994), using a construct validity framework to test the interpretability of positive item and negative item factors in the Rosenberg Self Esteem Scale. Their sample comprised 153 Singaporean school students (mean age 12.4 years) at risk of academic failure. The ten-item Rosenberg scale was used, but item 8 was dropped because of a high missing data rate. The two subscales therefore had five positive items and four negative items. The reliability of the overall scale was reported as .71; no reliabilities were given for the shorter subscales but they were presumably less than .7. Only two models were tested with confirmatory factor analysis—a single factor model and a two-factor model (positive item scale and negative item scale)—and the two-factor model fitted best.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Multiple regression was used to assess the relationship between positive and negative self esteem, and the dependent variables (mastery goal orientation, academic self-efficacy, and disruptive behaviour taken from Midgley et al.'s (2000) Patterns of Adaptive Learning Scales). Positive self esteem was a significant predictor of goal orientation and self-efficacy, while negative self esteem was a significant predictor of disruptive behaviour. Ang et al. concluded that positive and negative self esteem are different constructs. Unfortunately, Ang et al. did not directly compare their models with an alternative, comprising a single substantive self esteem factor and a negative wording method factor. The disruptive behaviour scale used as a dependent variable consists of negative statements about the self, and thus may be confounded with method variance common to the negative items of the Rosenberg scale. Modeling this method factor separately would have provided clearer evidence as to the substantive difference between the self esteem facets than the regression approach used by Ang et al.

The studies described so far in this section have made substantive interpretations of two factors emerging from scales which included negatively-keyed items. However, the studies have compared scale facets based on item keying with a single factor model; they have not tested models with a substantive latent (all items) and a method factor. The recent studies on Rosenberg's Self Esteem Scale described earlier (DiStefano & Motl, 2006; Quilty et al., 2006) explicitly model a method factor, with one substantive factor. Their findings suggest that the negative item wording factor has attributes of a consistent response style, and may be related to aspects of personality involving concern with evaluation by others. The evidence is tantalising, but conflicting – for example, Quilty et al. found a significant correlation between

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

avoidance motivation (BIS) and the negative wording factor, while DiStefano and Motl found no relationship.

Item Wording Factors are Just Method Effects

In my review of studies concerning the factor structure of the Rosenberg Self Esteem Scale, I discussed several examples where the two-factor structure was attributed to a substantive general self esteem factor, conflated with a substantively irrelevant method factor (e.g., Bachman & O'Malley, 1986; Carmines & Zeller, 1979; Marsh, 1996). Researchers investigating other scales with a mix of positively and negatively keyed items have reached similar conclusions.

Idaszak and Drasgow (1987) reviewed conflicting evidence regarding the factor structure of the Job Diagnostic Survey (Hackman & Oldham, 1975). Designed to measure five factors with 15 items, prior studies had identified solutions ranging between one and five factors. In the first part of their study, Idaszak and Drasgow identified six factors; five of these matched the a priori pattern of item loadings on JDS dimensions, while a sixth factor (the third to emerge in the analysis) had loadings from all of the negatively keyed items. This factor (which they labelled an 'artifact factor') correlated an average of .12 with the other five factors (compared with an average correlation of .42 between the substantive factors). The same pattern was found in a second (independent) sample. After rewording the five negatively-keyed items so that they were scored in the same direction as the rest of the survey, administration of the revised scale to a new sample resulted in a factor structure consistent with the JDS model.

The UCLA Loneliness Scale (Russell, Peplau, & Ferguson, 1978) was developed as a unidimensional measure of the bipolar loneliness construct. Originally,

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

all twenty items were worded in the same ('loneliness') direction. Concerns about possible response (acquiescence) bias, and relatively poor discriminant validity with related measures (e.g. depression) led Russell and his colleagues to create a revised version of the scale, in which ten of the twenty items were reverse scored (Russell, Peplau, & Cutrona, 1980). A number of researchers subsequently identified wording factors in this revised version (e.g., Knight et al., 1988; Miller & Cleary, 1993), calling into question the unidimensionality of the loneliness construct. In order to clarify the issue, Russell (1996) compared two alternative confirmatory factor analysis models with the single factor model—a two (oblique) factor model, with factors defined by direction of item keying, and a three factor model comprising a substantive loneliness factor (all items) and two orthogonal method factors (positive and negative items respectively). This last model fitted the data best across four diverse samples of data (students, nurses, teachers, and a sample of people over the age of 65). Accordingly, Russell concluded that the scale measured a "global loneliness factor along with two orthogonal method factors" (p. 37).

As part of a program investigating psychometric properties of the LBDQ, Schriesheim and Eisenbach (1995) used exploratory and confirmatory factor analyses to investigate scales combining regular, negated, polar opposite and negated polar opposite items. Details of the scale used, and examples of items, are included in the previous discussion of psychometric properties of negative items (see p. 18). Extracting two factors from the correlation matrix revealed a clear method factor, comprising the set of negated polar opposite items; all other items loaded on the substantive factor. In confirmatory factor analysis, a model which included four separate correlated method factors (one for each item type) and a substantive factor had the best fit. The method factors corresponding to negated items and polar opposite

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

items correlated significantly; combining these factors into one method factor resulted in a model with acceptable fit. The negated polar opposite items had the lowest average amount of trait variance. These items (e.g., "He does not give group members vague task assignments") lacked clarity and are unlikely to be considered acceptable for use in any well constructed scale. However, the existence of variance associated with method effects is clear, and it acts to undermine effectiveness of the scale for measuring the primary construct (Initiating Structure).

Magazine, Williams, and Williams (1996) investigated item wording factors in Meyer and Allen's (1984) Affective and Continuance Commitment Scales (ACS and CCS). They compared the standard two substantive factor model and an alternative three substantive factor model (McGee & Ford, 1987; Meyer, Allen, & Gellatly, 1990) with the same models incorporating a wording factor (for recoded items). In both cases, chi-square difference tests confirmed the superiority of the model containing a wording factor. The factor was considered irrelevant to substantive content, and a potential threat to validity if it correlated with similar method variance in other measures.

Another scale which has been scrutinised with respect to item wording effects is Gordon et al.'s (1980) union commitment scale. One of the facets in this scale comprises all negative items; while Gordon et al. acknowledged that it might be "purely an artefact of negatively worded questions" (p. 487), they argued for its substantive interpretation as a 'belief in unionism' facet. In a comparison of confirmatory factor analysis models of 20-item and 30-item versions of the union commitment scale, Kelloway, Catano, and Southwell (1992) reached a different conclusion. They obtained their best fitting model by using negative items as indicators for a method factor, orthogonal to the substantive factors. Kelloway et al.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

recommended that all the negative items be dropped from the scale. Bayazit, Hammer, and Wazeter (2004) argued that negative items might still contain valuable variance associated with the substantive construct. Using an 18 item subset of Gordon et al.'s original scale, they found that three oblique factors (loyalty to the union, responsibility to the union, and willingness to work for the union) together with an orthogonal factor for reverse-coded items provided the best fit. Loadings of the six reverse-coded items on the substantive factor were largely unchanged after including the methods factor, suggesting that they continued to tap relevant variance, and that variance associated with the method wording had no substantive meaning.

All these researchers have concluded that the factor associated with item wording has no substantive content. However, describing factors related to item wording as 'artefacts' or 'nonsubstantive' does not explain the mechanism by which they arise. I now turn to the reasons which have been put forward as possible processes underlying these factors.

Differing Item Response Distributions

Spector, Van Katwyk, Brannick, and Chen (1997) acknowledged that factors associated with item wording might relate to independent constructs in some scales. However, they felt such factors were more likely to be artefacts based on the way in which people respond to items phrased in different directions. People are more likely to agree with items that are close to their own position on a particular trait or attitude, and to disagree with items that are further away, in either direction ('ideal point' principle, Andrich, 1988; 'unfolding model,' Cliff, Collins, Zatkin, Gallipeau, & McCormick, 1988; 'tolerance' range, Thurstone, 1928). An enthusiastic person is likely to agree with items consistent with this self-view (e.g. 'I feel enthusiastic') and

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

disagree with inconsistent items ("I feel lethargic"). However, someone with a moderate mood may well disagree with both types of items. A mix of such respondents in a sample will result in differing item correlations – correlations within item type (all positive or all negative) will tend to be higher than correlations between different item types. This pattern of correlations will give rise to different factors, related to item wording. Spector et al. created a questionnaire comprising a mix of positively and negatively worded job satisfaction items, covering both extreme and moderate attitudes. They demonstrated that respondents would disagree with items which were more extreme than their own position, even when in the same direction; i.e., satisfied people would often disagree with the extreme positive items (and dissatisfied people with the extreme negative items). They then used a simulation study to generate data consistent with such response patterns, and demonstrated the existence of two factors (based on positive versus negative item wording) when extreme items were used. When only moderate items were used, a single factor model fitted as well as the two (oblique) factor model. Thus, they concluded that some observed instances of item wording factors are likely to be artefacts resulting from extreme item wording, combined with respondents who vary between moderate and extreme positions on the construct being measured.

McPherson and Mohr (2005) sought to confirm Spector et al.'s (1997) findings with the use of real responses to a scale, rather than simulated data. They used the Life Orientation Test (LOT; Scheier & Carver, 1985). Scheier and Carver found that positive and negative items loaded onto separate factors, but as these correlated .64, they argued that the scale should be treated as unidimensional. Other researchers have argued for a two-factor (optimism and pessimism) interpretation (e.g., Marshall, Wortman, Kusulas, & Hervig, 1992). Thus the LOT provides a good vehicle for

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

investigating the extent to which the mechanism outlined by Spector et al. could be creating an artifactual factor.

As an illustration of potential responding consistent with ideal point or unfolding model perspectives, McPherson and Mohr (2005) presented two items from the LOT, presumed to measure different poles of the optimism construct: "I'm always optimistic about the future" and "Things never work out the way I want them to." While a highly optimistic person would agree with the first and disagree with the second, a person with a mid-range level of optimism could logically disagree with both statements. In their study, McPherson and Mohr used the original scale together with a modified version in which items were rephrased so as to be more moderate (for example, by changing "I'm always optimistic about the future" to "I'm usually optimistic about the future"). They identified an interaction between item extremity and item keying (consistent with Spector et al., 1997), with lower negative-item means and higher positive-item means in the moderate version of the LOT than in the original version. The factor analysis results were less clear cut. A two factor model (based on item wording direction) provided the best fit to both scale versions, although the chi-square difference between a one and a two factor model was smaller for the moderate version than for the original version. McPherson and Mohr concluded that item extremity appeared to be "amplifying differences between oppositely keyed items and increasing the level of multidimensionality" (p. 128).

Cognitive Complexity of Negative Items

One of the most important requirements for a good scale item is that it be easy to understand. Likert (1932) emphasised that each item should be stated "in *clear, concise, straight-forward statements.* Each statement should be in the simplest possible

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

vocabulary. No statement should involve double negatives or other wording which will make it involved and confusing" (p. 45, emphasis in original), and this advice is echoed in more recent guides to scale construction (e.g., Clark & Watson, 1995; Hinkin, 1998; Spector, 1992).

Studies of speed and accuracy in language processing highlight the risks of using complex negative items. Kaup, Zwaan, and Lüdtke (2007) summarised research into the effect of negation on sentence processing difficulty. In all studies reviewed, it was found that "negative sentences were harder to process than affirmative sentences, as is evidenced by longer processing times and/or higher error rates for negative sentences compared to affirmative sentences" (p. 260). Their summary showed this to be true for regular negation ("The circle is not present"), polar opposites ("The circle is absent"), and without regard to the location of the negation operator. Furthermore, sentence verification tasks (where respondents have to determine if a positively or negatively worded sentence is an accurate statement) consistently show differences in processing time. It takes longer to verify negative statements than positive statements, and true negatively-worded statements take longer to process than false negatively-worded statements. Lansman (cited in Hunt, 1978) found a significant correlation between verbal ability and performance on such tasks.

The difficulty of processing negated statements, and the relationship with verbal ability, has led some researchers to suggest this as a cause for the emergence of factors related to item wording in scales. Benson and Hocevar (1985) proposed that difficulty in understanding negative items might partly explain their finding of different means and variances between positive and negative items. Similarly, Ahlawat suggested that negative or double negative items create "cognitive complexity for the students, and end up measuring nothing but ambiguity and confusion" (p. 98).

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Marsh (1986, 1996) sought to confirm this suggested relationship between verbal (reading) ability and the emergence of item wording factors. Marsh (1986) administered his 66-item Self Description Questionnaire to a sample of 654 schoolchildren (aged from 7 to 10 years old). The questionnaire measures seven different facets (e.g. enjoyment of sports, attractiveness, relationship with peers, relationship with parents) and includes ten negatively-keyed items. Given the overall scale reliability, one could expect a correlation of .8 between the mean score on positive items, and the mean score on negative items, if they were measuring the same construct. In fact, Marsh found a correlation of .27. It was virtually zero (-.02) for the youngest children, increasing to .60 for the older children. Furthermore, investigation of item means and standard deviations identified evidence of children endorsing negative items, inconsistent with their overall positive self-concept. This was more frequent with younger children, suggesting that the problem with negatively keyed items is linked to cognitive development.

In a new sample of 559 children, Marsh (1986) used confirmatory factor analysis to demonstrate that negative items defined a method factor (as well as reflecting variance in the substantive self-concept construct). By including two measures of verbal ability (comprehension and word knowledge), and teachers' ratings of reading ability, Marsh was able to incorporate a latent construct for reading ability into his models. The negative item factor correlated significantly with this factor (r = .42), indicating that children with poorer reading skills were more likely to answer negatively-keyed items in a manner inconsistent with their responses to positive items.

Marsh's (1996) work using self-esteem data from the NELS (discussed previously on page 21) produced a similar finding. The correlation between scale scores calculated from negative items and from positive items increased steadily with

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

increased reading ability. Dividing students into quartiles on the standardised reading test in the NELS data, Marsh found correlations of .49, .64, .69, and .74. Thus, students with higher reading ability showed much less difference between responses to negative and positive items than did less able students. Dunbar, Ford, Hunt, and Der (2000) replicated Marsh's study using a sample of Scottish adults, and also found that measures of negative and positive self esteem were less correlated in subjects with lower verbal ability.

Weems, Onwuegbuzie, and Collins (2006) used a different approach to investigate the relationship of reading ability to emergence of item wording effects. They used a sample of 153 graduate students (in psychology and education) who could be expected to have high levels of reading ability. Reading ability was measured with the Nelson Denny Reading Test (NDRT; Brown, Fischo, and Hanna, cited in Weems et al.), and participants also completed the Library Anxiety Scale (LAS; Bostick, cited in Weems et al.). Four of the LAS scales contain both positive and negative items, and these scales were used in the study. There was no significant difference between the mean score for positive items and the mean score for negative items on three of the four LAS scales. However, Weems et al. proceeded to conduct a canonical correlation analysis, comparing absolute difference between positive item and negative item scores on each scale with the two NDRT subscale scores (Reading Comprehension and Reading Vocabulary). This analysis identified a significant canonical root ($r = 0.32$).

While Weems et al. (2006) interpreted this as evidence supporting the view that reading ability contributes to item wording effects, there are some problems with their study. The lack of significant differences between positive and negative item scores on three of the four scales suggests the effect does not apply to negative items in general – it was limited to only one scale. Furthermore, the canonical correlation result was

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

based on difference scores. Such scores confound the effect of the two component scores, thereby making it impossible to interpret the finding unambiguously (J. R. Edwards, 1994). Finally, use of positive item and negative item scores ensured that substantive variance related to participants' attitudes to libraries was included in both scores. It is quite likely that attitudes to libraries correlate with reading ability. It would have been more useful to model a separate wording factor (orthogonal to the substantive library scales) and test the relationship between this factor and reading ability. Without such an analysis, Weems et al.'s findings provide only limited support for the hypothesised relationship between reading ability and difficulties associated with negatively worded items.

Using a large sample (N = 3,400) of employed adults, Cordery and Sevastos (1993) also sought to confirm the relationship between education level and emergence of item wording effects in the Job Diagnostic Survey (JDS; Hackman & Oldham, 1975). They used the original version of the JDS, but also included the five positively worded replacement items developed by Idaszak and Drasgow (1987). This enabled them to identify whether education level related to dimensionality of the JDS in the original version (which included five negatively worded items) as well as in the revised (positively worded) version. The fit of a five factor model (consistent with the JDS model) was unacceptable for both elementary and high school graduates using the original scale items. Using the revised version (replacing negatively worded with positively worded items) resulted in acceptable fit in both samples. Education level was therefore not the underlying reason for the negative wording factor. Cordery and Sevastos noted that education level could be confounded with motivation to respond accurately, and concluded that item wording effects were "a function of general care exercised in responding to negatively worded items" (p. 143).

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Careless responding has long been recognised as a potential source of error in questionnaire responses. Nunnally (1967) gives examples of the "ample evidence of the effects of carelessness and confusion in the responses to many different types of psychological measures" (p. 617). He notes that carelessness and confusion introduce an additional source of randomness (measurement error), and also act to bias scores towards the chance level (mid-point on an attitude scale). If individual differences exist among respondents in regard to the degree of carelessness or confusion, this bias can differentially affect the score distributions of subgroups. For example, children could be expected to exhibit more carelessness than adults (Nunnally, p. 618) and, as Cordery and Sevastos (1993) suggest, this might account for past findings of an association between reading ability and the emergence of item wording effects in scales. Evidence in support of carelessness as a possible cause of negative wording factors will now be discussed.

Carelessness

If careless or confused responding is equally likely to affect both positively worded and negatively worded items, it will not lead to emergence of factors related to item wording. I will therefore begin this section by discussing evidence that negatively worded items may be more prone to careless responding.

Evidence relating to the difficulty of processing negated statements (Kaup et al., 2007) has already been discussed; such statements require longer processing time and result in higher error rates than affirmative sentences. Furthermore, linguistic studies into the 'Lexical Invisibility Hypothesis' (Low, 1996) indicate that questionnaire respondents sometimes overlook words used as modifiers in questionnaire items. For example, words such as 'very', 'never', 'seems' and 'tends'

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

can be used to amplify or attenuate the strength of an assertion in a questionnaire. Low used a 'thinking aloud' protocol, in which students provided detailed verbal commentary on their thought processes relating to completion of questionnaire items. His analysis showed that a number of lexical qualifiers and hedges in item stems appeared to be overlooked. None were overlooked by all subjects, and average 'attention rates' varied between 50 and 90 percent. The only negation modifier included by Low was 'never,' and this had a relatively low visibility with Low's subjects. Although far from conclusive, the study suggests the possibility that subjects may be more likely to overlook negating qualifiers when reading questionnaire items.

Barnette (1995) discusses a variety of reasons that can lead to careless or 'nonattending' behaviours among questionnaire respondents. These include misunderstanding items (which may partly be a function of verbal ability); low motivation to respond accurately; anger, frustration or other emotions; a desire to finish the task quickly; and fatigue. Barnette used three large samples (1,240 primary school students, 3,541 high school students, and 2,688 teachers) to identify patterns of responses consistent with nonattentive responding. The latter two samples were administered questionnaires which included negatively-keyed items. By collapsing 'strongly agree' and 'agree' responses into one category, and 'strongly disagree' and 'disagree' into a second category, Barnette was able to conduct a 2 X 2 (reversed/non-reversed items by agree/disagree) chi-square analysis. He assumed that people would have similar response distributions for the two types of items – e.g., a respondent who agreed with 20% of the positively-keyed items could be expected to also agree with 20% of the (recoded) negatively-keyed items. In the student sample, 17.7% of respondents had significant differences between their response distributions to the two types of items ($p < .01$). For teachers, 10.3% had different responses ($p < .01$). By

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

chance, around 1% of respondents could be expected to have significant differences at $p < .01$; the obtained percentages (17.3% and 10.3%) therefore indicate inconsistency between responses to positive and reverse-worded items, suggesting that a noticeable proportion of respondents may not have attended to the reverse wording of survey items.

Holden and Fekken (1990) studied the 220-item Basic Personality Inventory (BPI) with a view to identifying characteristics of individual items which were associated with 'item goodness'. One criterion of goodness was item stability, based on the percentage of individuals who responded identically to the same item on two testing occasions separated by a one month interval. While a number of processes might lead to such inconsistent responding, careless or inattentive responding (perhaps linked to motivation or fatigue) is one of the more important factors. Carelessness is not deterministic – an item answered carelessly (and inaccurately) on one occasion, may well be answered more carefully and accurately on another occasion. An item which is answered carelessly on one occasion, and accurately on another, will show low item stability. A very stable item is one that is less likely to be affected by carelessness on either occasion. Holden and Fekken calculated a negative item score by summing the number of negations (explicit negation, implicit negatives, and negative qualifiers) in each item. They found a significant negative relationship ($r = -.26$, $p < .01$) between negative wording and item stability. This supports the view that negatively-worded items are more vulnerable to careless responding than positively worded items.

These studies indicate clear differences between the processing of negative and positive items. There is a greater likelihood of carelessness or confusion affecting responses to negative items, but it is certainly not the case that all respondents are

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

always careless. Are levels of carelessness consistent with empirical observations (e.g., Barnette, 1995) sufficient to explain the prevalence of an item wording factor in balanced scales?

Schmitt and Stults (1985) explored this, demonstrating that careless responding to negatively-keyed items by small proportions of respondents can cause the emergence of a wording-related factor. Using three different correlation matrices of 30 items, calculated from real data, they computed the complete factor loading matrix for each. These matrices were then used to generate simulated raw data (sample sizes of 400), after reversing the signs of factor loadings for 4, 8 or 12 items to simulate negatively-keyed items. The simulated data had means of 5 (positive items) and 3 (negative items—translating into a mean of 5 once recoded) and a standard deviation of 1.2. The data were converted to integers (1 to 7) by truncating decimals, and collapsing numbers greater than 7 or less than 1 into the end scale points. Principal components analysis with varimax rotation was used to extract factors (eigenvalue > 1 criterion) after leaving varying subsets of cases (0%, 5%, 10%, 15%, and 20%) unrecoded, to simulate carelessness.

For all three datasets, a component associated with negatively worded items was apparent when negatively-keyed items were left unrecoded for as few as 10% of cases. The prominence of the factor increased with increasing numbers of negative items (from 4 to 12).

The approach used by Schmitt and Stults (1985) in their study mirrored common analytical procedures at the time, but is no longer consistent with recent advances. In particular, the study's use of principal components analysis, determination of the number of components using the 'eigenvalues > 1' criterion, and

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

varimax rotation of components is inconsistent with current recommended practice (Preacher & MacCallum, 2003).

Principal components analysis is a data reduction technique designed primarily for use with continuous data, and can give misleading results with categorical data (Bernstein & Teng, 1989; Gilley & Uhlig, 1993; Mayer, 1971). It is also designed to create a small set of composite variables which retains as much information as possible from the original (larger) set of measured variables. Unlike factor analysis, it does not seek to explain relationships between variables by identifying the pattern of relationships between latent factors (Fabrigar, Wegener, MacCallum, & Strahan, 1999). In many situations, the two approaches give very similar results, particularly for high communalities and a large number of variables (Gorsuch, 1990). Schmitt and Stults's study used 30 variables, but Snook and Gorsuch (1989) found common factor analysis was more accurate than component analysis in reproducing original factor loadings, even with 36 variables. An example of real data where principal components gave very different results from factor analysis is presented in Gorsuch's 1988 *Handbook of multivariate experimental psychology* chapter (cited in Gorsuch, 1990); Loehlin (1990) provides a synthetic example.

Thus, Schmitt and Stults's (1985) findings cannot automatically be generalised to scale analysis involving latent factors (e.g. confirmatory factor analysis), or analysis methods designed for ordinal data. Furthermore, the eigenvalue criterion relied upon by Schmitt and Stults for determining the number of components is a less valid measure than Velicer's (1976) minimum average partial correlation (MAP) method or parallel analysis (Hayton, Allen, & Scarpello, 2004; J. L. Horn, 1965).

Two other aspects of Schmitt and Stults's (1985) study deserve comment—their use of skewed item distributions, and the way in which they modeled

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

carelessness. When generating items, Schmitt and Stults assigned a mean of 5 (on a 7-point scale) to positively keyed items and a mean of 3 to negatively-keyed items. Coupled with subsequent recoding into ordered categories, this would result in skewed item distributions. The subset of cases treated as careless would therefore exhibit a very different item distribution (positively-skewed) to the remainder of the sample (negatively-skewed). With sufficient levels of carelessness, the level of heterogeneity in the negative items would lead to the emergence of an additional component (Bernstein & Teng, 1989). If both types of item had been assigned the same mean (4), failure to recode negative items would not result in a different distribution (other than due to sampling variation). Thus, Schmitt and Stults's finding addresses only the case where item response distributions are skewed.

Using skewed item response distributions is consistent with empirical data. Taylor and Brown (1988) have summarised evidence confirming the strong tendency of people to make overly positive self-ratings of their personal traits and abilities, resulting in skewed self-appraisals. Micceri (1989) analyzed 125 samples of real data from psychometric measures and found that 84% were at least moderately asymmetric. Simulation of careless responding needs to consider items with skewed distributions, but it is also important to include a condition which models normally distributed item responses.

Schmitt and Stults (1985) assumed that all careless respondents would answer all negatively-keyed items carelessly. They suggested that careless respondents would infer the meaning of the scale from the first few items and answer remaining items in a consistent manner, even when they were worded in an opposite direction. However, the concept of carelessness implies inattention and chance. One respondent might miss the changed orientation of early negatively-keyed items before recognising one and

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

attending more closely to the rest of the questionnaire. Another respondent might correctly answer the first few negatively-keyed items, but miss later ones as tiredness sets in. Thus, the simulation of carelessness used by Schmitt and Stults does not cover situations where respondents answer varying proportions of negatively-keyed items carelessly, and where not all careless respondents answer the same negatively-keyed items carelessly.

Woods (2006) sought to remedy one of these shortcomings (the use of principal components) by using a confirmatory factor analysis framework to investigate careless responding on reverse-worded items. She generated dichotomous items using a two parameter logistic model (Birnbaum, 1968) for a 23-item test (including 10 negatively keyed items). Carelessness was simulated by switching 0 for 1 (and vice versa) for all 10 negatively keyed items, for varying percentages of respondents (0%, 5%, 10%, 20% and 30%) and sample sizes (250, 500, and 1,000) using 1,000 replications for each condition. Both a one-factor and two-factor model (with a correlation of 1 between factors) fitted when there was no careless responding. When 10% of respondents were careless, the fit of the one-factor model declined noticeably for all sample sizes. As carelessness reached 30%, fit of the one-factor model became "abysmal" (p. 191) while the two-factor model had excellent fit (and a correlation of .30 between factors).

Although supporting Schmitt and Stults's (1985) findings, Woods (2006) used only dichotomous items. This format is very rare in organisational studies; fewer than 3% of the scales described in the British Telecom Occupational Psychology Unit survey item bank (1984a, 1984b) have dichotomous scales. Woods also noted that her study simulated an "all or nothing" approach to carelessness; she stated that future research needs to investigate more plausible scenarios, whereby respondents answer carelessly in varying degrees.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

The Need for Further Research into Careless Responding

In this chapter, I have presented a comprehensive review of the reasons underlying the use of balanced scales, and the potential problems emerging from a mix of positively and negatively keyed items. I have focused particularly on the emergence of an item wording factor in scales which include both positively and negatively keyed items. In the past, it has been possible for researchers to ignore such effects. For example, Nunnally and Bernstein (1994) recommended that scales with highly correlated wording factors (i.e. correlations around .7 or higher between positive item and negative item scales) should be treated as unidimensional. While such an approach allows for identification and estimation of regression and correlation coefficients, failure to model the factor structure appropriately in a latent variable model can result in unacceptable fit, precluding any meaningful interpretation. Thus, while the issue of item wording factors is not new, it has assumed greater importance in the context of modern data analytic techniques.

This review has highlighted a number of important gaps in our current knowledge. Carelessness in responding to negatively keyed items appears to be an important process underlying the generation of item wording effects in rating scales. Whether the careless responding is due to confusion, misunderstanding, low motivation or fatigue, the net effects are indistinguishable (Nunnally, 1967). Accordingly, we need to develop a more accurate understanding of the extent to which careless responding on summated rating scales acts to impair our ability to accurately model summated rating scales in a latent variable framework. The only two studies to model careless responding (Schmitt & Stults, 1985; Woods, 2006) have used an unrealistic model of carelessness (assuming that careless respondents will answer all

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

negatively-keyed items in the wrong direction). Furthermore, the Woods study used a dichotomous scale rather than the more typical five or seven item scale used in organisational research.

The majority of studies into item wording factors have used an exploratory factor analysis framework; none of these has used currently accepted methods (parallel analysis or minimum average partial correlation) to identify the appropriate number of factors to extract. Those studies which have employed confirmatory factor analysis have, with few exceptions (Corwyn, 2000; Marsh, 1996), treated ordinal data as continuous. Finally, the Schmitt and Stults study used skewed item distributions; the effect of carelessness was not evaluated for normally distributed responses.

This thesis aims to address these gaps. I begin in Chapter 3 by establishing the existence of an item wording factor in four diverse samples using a recently published scale, the Core Self Evaluations Scale (CSES; Judge et al., 2003). Because method effects vary across populations (Goldsmith, 1986) and need to be established in different scales on a case-by-case basis (Quilty et al., 2006), confirming the existence of wording effects in the CSES will allow me to use these data in subsequent analyses.

In Chapter 4, I use a jackknife technique (Rensvold & Cheung, 1999) to identify inconsistent patterns of responding to positive and negative items (consistent with carelessness) in the CSES samples. On the basis of these observed patterns, I create a Monte Carlo simulation, varying the probability of a respondent answering carelessly and, for each careless respondent, varying the likelihood that a negatively-keyed item will be answered carelessly. Data analytic techniques appropriate for use with ordinal data are used to assess how the fit of a one factor CFA model changes with increased carelessness, and to explore the way in which this fit varies with different levels of item skew and intercorrelation.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

In Chapter 5, I demonstrate the ability of the jackknife technique to identify extreme cases of careless responding in the Monte Carlo samples. Applying the same approach to the empirical CSES samples, I show that the effects of item wording can be dramatically reduced by removing relatively small numbers of cases, identified by jackknife analysis.

The dissertation concludes with an overall summary and discussion of findings in Chapter 6.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

CHAPTER THREE

ITEM WORDING EFFECTS IN THE CORE SELF-EVALUATIONS SCALE

Emerging out of work into dispositional bases of job satisfaction, Judge, Locke, and Durham (1997) proposed a broad personality trait they labeled core self-evaluations. This trait was construed as a fundamental self-evaluative trait, one which would operate subconsciously to influence people's self-appraisals in a wide range of different situations. Judge et al. identified self-esteem, neuroticism, general self-efficacy and locus of control as existing well-researched traits that were likely to act as indicators of core self-evaluations. Early core self-evaluation research operationalised the construct by combining existing individual measures of these traits, but a purpose-designed scale has now been published (Judge et al., 2003). A noteworthy aspect of the scale is its use of equal numbers of positively and negatively worded items. As discussed in the preceding chapter, this approach is consistent with traditional recommendations for scale construction, but can lead to the creation of a separate factor based on item wording. In this chapter, I confirm the existence of item wording effects in the Core Self-Evaluations Scale, and illustrate the importance of these for users of the scale.

Core Self-Evaluations

While other traits may act as indicators of core self-evaluations—dispositional optimism and negative affectivity have been considered (Erez & Judge, 2001; Judge, Locke, Durham, & Kluger, 1998)—most research has been limited to self-esteem, neuroticism, general self-efficacy and locus of control. Judge, Erez, Bono, and

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Thoresen (2002) identified similarities in the role played by these constructs in theory, and ways in which they have been operationalised. Their meta-analysis of 75 published studies found the four measures had an average correlation of .60, and found evidence of a second-order factor accounting for these correlations. Judge et al. note that each individual construct may have value in specific settings, but the existence of a common underlying factor argues for the value of integrating research streams relating to self-esteem, neuroticism, general self-efficacy and locus of control. The higher-order factor (core self-evaluations) may well account for variance hitherto attributed to individual constructs. Furthermore, use of such a broad factor may allow for higher and more consistent validities when predicting broad outcome criteria such as job performance and job satisfaction (Judge, Van Vianen, & De Pater, 2004).

Several studies have demonstrated the relevance of core self-evaluations to important work variables. Research and/or meta-analytic studies have demonstrated its relationship with job satisfaction (Judge, Bono, & Locke, 2000; Judge et al., 1998), job performance (Judge & Bono, 2001), motivation (Erez & Judge, 2001), social ties formed by expatriates (E. C. Johnson, Kristof-Brown, Van Vianen, De Pater, & Klein, 2003), responses to multi-source feedback in work settings (Bono & Colbert, 2005), sales performance (Sager, Strutton, & Johnson, 2006), and attributions made regarding computer technology (R. D. Johnson, Marakas, & Palmer, 2006).

Measuring Core Self-Evaluations

All the studies cited in the previous paragraph measured core self-evaluations indirectly, using existing measures of self-esteem, locus of control, neuroticism, and general self-efficacy. These individual scales were converted to a measure of core self-evaluations in different ways. In some studies, individual scale scores were averaged

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

(e.g. Bono & Colbert, 2005), or scale scores were weighted by their factor loading based on a principal components analysis of the scale scores (E. C. Johnson et al., 2003). In other studies, the first component extracted in a principal components analysis of all individual items from the four scales was used (e.g. Erez & Judge, 2001). Confirmatory factor analysis techniques have also been applied , with core self-evaluations modelled as a second-order latent construct, and parcels of scale items used as indicators for the four first-order latent variables (e.g. Judge et al., 2000).

Judge, Erez, Bono, and Thoresen (2003) discuss the practical and conceptual difficulties resulting from such indirect measurement of core self-evaluations, and present a 12-item scale designed to measure the construct directly—the Core Self-Evaluations Scale (CSES). The items were designed to cover the broad domain defined by the four component constructs, although care was taken to avoid creating a scale based merely on combining items tapping individual core traits. Both student and employed adult samples were used to develop the final measure, which demonstrated high internal consistency (alpha values greater than .8 in four different samples). Acceptable convergent and discriminant validity were demonstrated through correlations with individual measures of the four component constructs, and with Big Five personality measures.

The CSES comprises six positively worded and six negatively worded (reverse-scored) items (see Table 1). Given the consistency with which past studies have demonstrated the existence of a factor created by negatively-worded items, it seems likely that such a factor will also be present in the CSES. Judge et al. (2003) used confirmatory factor analysis (CFA) to demonstrate the scale's unitary factor structure, but there is some ambiguity about the exact factor structure tested. The single factor model estimates 24 parameters (12 error variances and 12 factor loadings), and the 12

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

scale items generate 78 unique elements in the variance-covariance matrix. The model should therefore have 54 degrees of freedom. The model reported by Judge et al. (2003, Table 3, p. 318) has only 48 degrees of freedom (and the other models reported in the table also have six fewer degrees of freedom than expected). Either an additional (unreported) latent variable was specified, or six pairs of error terms were allowed to covary. Similarly, in reporting the results of confirmatory factor analysis of Spanish and Dutch versions of the CSES, Judge et al. (2004, Table 5, p. 338) the models have only 50 degrees of freedom, indicating that four additional parameters were freed, over and above those required for the CFA.

Item Wording Effects in the CSES

It is possible that Judge et al. allowed error terms to covary in order to accommodate common variance between reverse-scored items[1]. In the early days of structural equation modelling, it was a reasonably common practice to allow correlated measurement residuals in order to increase model fit (Fornell, 1983). However, correlations between error terms indicate the existence of systematic error, meaning that the items are measuring something in addition to the construct they are supposed to measure (Jöreskog, 1993). While model fit can be improved by allowing correlated errors, such an approach comes at the cost of "a correspondent loss of the meaning and substantive conclusions which can be drawn from the model" (Gerbing & Anderson, 1984, p. 574).

---

[1] I have been unable to find out from the two lead authors the nature of the model they confirmed with CFA (T. Judge, personal communication, July 31, 2006; A. Erez, personal communications, March 11, 2007; March 13, 2007; May 16, 2007; May 21, 2007.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Table 1

The Core Self-Evaluations Scale (CSES)

Instructions: Below are several statements about you with which you may agree or disagree. Using the response scale below, indicate your agreement or disagreement with each item by placing the appropriate number on the line preceding that item.

1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

1. I am confident I get the success I deserve in life.
2. Sometimes I feel depressed. (r)
3. When I try, I generally succeed.
4. Sometimes when I fail I feel worthless. (r)
5. I complete tasks successfully.
6. Sometimes, I do not feel in control of my work. (r)
7. Overall, I am satisfied with myself.
8. I am filled with doubts about my competence. (r)
9. I determine what will happen in my life.
10. I do not feel in control of my success in my career. (r)
11. I am capable of coping with most of my problems.
12. There are times when things look pretty bleak and hopeless to me. (r)

r = reverse-scored. This measure is non-proprietary (free) and may be used without permission.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Two approaches can be taken to modelling item wording method effects: the correlated trait–correlated method (CTCM) and the correlated trait–correlated uniqueness (CTCU) methods. Marsh and Grayson (1995) describe these in the context of a multitrait–multimethod conceptual framework. The CTCM incorporates a latent variable underlying items of the same response type, thereby assuming that the method (item wording) is unidimensional. The CTCU approach allows the uniqueness ('error') terms to covary among items using the same method. This approach does not assume that the method effect is unidimensional, and (unlike the CTCM method) does not allow for separate estimates of method and error effects (Quilty et al., 2006). Based on a review of each approach's strengths and weaknesses, Lance, Noble, and Scullen (2002) recommended the CTCM method as having both theoretical and substantive advantages over the CTCU method. For example, this approach allows for unbiased estimation of both method and trait variance components. Researchers can remove the item wording variance from scale items through specifying a method factor, thereby controlling for method effects when estimating relationships between constructs.

The CSES is a relatively new measure, and empirical studies using the scale have only begun to appear in recent years. To date, no published articles have presented clear information on the factor structure of the CSES. Most researchers cite the source article as evidence for unidimensionality, and then use scale scores in correlation or regression analyses to test their hypotheses (e.g., Beal, Trougakos, Weiss, & Green, 2006; Blau, 2007; Brunborg, 2008; Kluemper, 2008; Nikolaou & Judge, 2007; Salvaggio et al., 2007; Spence Laschinger & Finegan, 2008; Tsaousis, Nikolaou, Serdaris, & Judge, 2007; Wanberg, Glomb, Song, & Sorenson, 2005).

Only a few studies have used the CSES in structural equation models. Chamorro-Premuzic, Ahmetoglu, and Furnham (2008) used the CSES in a structural

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

equation model, but did not present any confirmatory factor evidence for unidimensionality. Instead, they stated that principal components analysis indicated the existence of only one factor. Capon, Chernyshenko, and Stark (2007) used overall CSES scale scores in their path model.

D. J. Brown, Ferris, Heller, and Keeping (2007) used a structural equation modelling approach, but randomly combined the 12 CSES items into three parcels. Such an approach is appropriate when a measure is unidimensional, but Brown et al. did not test the CSES for unidimensionality. As shown by Bandalos (2002, p. 100), item parcelling of scales which are not unidimensional "can yield misleading results with regard to the actual fit of one's model, as well as biased estimates of other model parameters." The parcelling approach effectively disguises the possible existence of an item wording factor, and the ways in which such a factor affects results.

No published studies have used a latent variable approach incorporating item-level CSES data. It is in this framework that the existence of a wording factor is likely to present most problems for researchers. If the CSES has a two-factor structure, then modelling it as a unidimensional scale will increase the discrepancy between the model-implied covariance matrix, and the sample covariances. If this discrepancy is large enough, the model will be rejected. A failure to recognise the existence of an item wording method factor, and to model it appropriately, may well result in researchers falsely rejecting substantive structural models, when the true problem is in their measurement model. The absence of any published studies using structural equation modelling four years after publication of the CSES suggests the possibility of a problem, though it is clearly not definitive evidence. Informal contacts with researchers using the scale give further circumstantial evidence that lack of unidimensionality may be causing difficulties. For example, a Turkish translation of

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

the CSES gave clear evidence of an item wording factor (M. Bayatzi, personal communication, June 15, 2007). S. Boyd (personal communication, May 22, 2007) reported finding two factors when undertaking principal components analysis of the CSES as part of a study into employee burnout (Boyd, Ensari, Hoffman, & Newman, 2007). K. Mueller (personal communication, April 11, 2008) could not get acceptable fit for a German translation of the CSES until he was made aware of the existence of a wording factor.

It is therefore important to evaluate the extent to which structural models involving the CSES are affected by the decision as to whether or not to model an item wording factor. The extent to which existence of an item wording factor reduces model fit when the scale is modelled as a unidimensional measure is an empirical question. Researchers intending to use the CSES need to know more precisely what wording effects are present, and how they might influence modelled relationships with other variables of interest.

Including such an item wording factor raises an additional issue. Is this factor merely a method artifact, or does it contain substantive information? Most studies have treated such factors as nuisance method effects, but Horan, DiStefano, and Motl (2003) found significant moderate correlations between negative wording effect factors in different measures (self-esteem, locus of control, and attitude towards school). This suggests the possibility that such factors might arise from a type of response style, rather than just a method effect.

In this study, this question will be answered by using three different dependent variables from two different data sets: self-mastery (a construct measured in Sample 1), teachers' in-class performance (Sample 2), and critical thinking (Sample 1). Critical thinking is measured using a scale which has negative items, thereby allowing a

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

negatively worded item factor to be modelled on the endogenous as well as exogenous side of the model. Analysing three different models from two different samples will give greater certainty to conclusions about the nature of the item wording factor and its relationship to other constructs.

Method

Instruments

Core self-evaluations. The CSES items were taken from Judge et al. (2003). The original questionnaire comprises twelve items, with alternating positive and negative items (see Table 1), and respondents use a five-point rating scale (1 = 'strongly disagree' to 5 = 'strongly agree'). In all the analyses described below, odd-numbered items are positively worded, while even-numbered items are negatively worded. Negative items were recoded before analysis so that high scores on all items reflected high core self-evaluations. As described below, samples analysed in this study used several variations on the original rating format. These changes were generally made by researchers with the aim of ensuring consistency within studies (e.g. using 7-point rating scales for all instruments) rather than for any specific psychometric reason.

Self-mastery. This self-report scale comprised five positively-worded items (e.g., "I am able to remain composed under stressful conditions") measured on a seven-point rating scale. The content of the scale covers the ability to control emotions, manage anxiety, and remain focused on tasks in stressful job conditions, and is therefore conceptually related to core self-evaluations. It was completed by participants in Sample 1 as part of the same questionnaire containing the CSES. In order to reduce common method effects, the CSES items and self-mastery items were

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

presented in different sections of the survey questionnaire, and used different rating anchors ('agreement' for CSES and 'accuracy of description' for self-mastery). The scale had high reliability (alpha = .86).

In-class performance. This self-report scale comprised seven positively-worded items from a scale developed by Benjamin (2002). The scale measured perceived performance on a number of important classroom leadership and teaching tasks (e.g., "I manage student behaviour"). Core self-evaluations have been shown to relate to a wide range of job performance and leadership measures (Judge & Bono, 2001; Judge, Bono, Ilies, & Gerhardt, 2002), and CSES would therefore be expected to have a positive relationship with in-class performance. The scale items were presented in a different section of the questionnaire to the CSES items, and used a six-point scale rather than the five-point scale used by the CSES. The scale had high reliability (alpha = .89).

Critical thinking. This six-item self-report scale measured perceived effectiveness at analysing and solving problems through logical analysis. It was included in this study because the scale has two negatively-worded items, and therefore allows for inclusion of an item-wording factor on the endogenous side of the model. An example of a positive item is "I think through problems logically based on facts and reasoning", while a negatively-worded item is "I become confused or overwhelmed by excessive details during problem solving." Responses to the scale will be influenced to an extent by general self-efficacy, suggesting a likely positive relationship with CSES. As with self-mastery, this scale was presented in a different section of the questionnaire than the CSES items, and used a different rating scale. Scale reliability was adequate (alpha = .71).

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Samples

Four different samples were used. In confirming the factor structure of the CSES, each of these samples was analysed separately to identify the presence of item wording effects. Samples 1 and 2 were further used to assess structural models.

Sample 1. This sample comprised 265 officers in the Singapore Armed Forces (96.2% male) whose ages ranged from 19 to 52 with a mean of 28.7 (SD = 6.4). For this sample, the response scale was changed to a seven-point scale, but 'agreement' anchors were retained (1 = 'strongly disagree' to 7 = 'strongly agree'). Data were gathered as part of a larger study into multi-source feedback and leadership development.

Sample 2. This sample comprised 178 Singaporean teachers (31.5% male); ages ranged from 17 to 61 with a mean of 33.5 (SD = 10.4). A five-point rating scale was used, but instead of rating agreement, participants were asked to rate the extent to which each statement was an accurate description of them (1 = 'very inaccurate' to 5 = 'very accurate'). Data were gathered as part of an unpublished study into personality and job performance of teachers (S. L. C. Tay-Lee, personal communication, August 30, 2006).

Sample 3. This sample comprised 2,088 employed adults undertaking part-time tertiary study in New Zealand (41.6% male); ages ranged from 19 to 69, with a mean of 38.5 (SD = 9.5). For this sample, the response scale was changed to a seven-point scale, but 'agreement' anchors were retained (1 = 'strongly disagree' to 7 = 'strongly agree'). Data were obtained as part of a longitudinal study into the effect of core self-evaluations on perceived organisational support, organisational commitment, and conflict between work and non-work roles (Pajo, Guenole, Mallon, & Ward, 2005).

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Sample 4. This sample comprised 1,876 Flemish job seekers (49.1% male); ages ranged from 17 to 58, with a mean of 27.1 (SD = 9.1). The questionnaire was translated into Dutch and data were gathered as part of a study into personality and job search strategies (Van Hoye, 2006). The original five-point rating scale was retained, but the order of items was altered; respondents answered all of the positive items first, followed by the negative items. Minor wording changes were made to two items, 6 and 10. The original item 6 ("Sometimes, I do not feel in control of my work") became "I feel like I have little control of my work" in back-translation. Item 10 (originally "I do not feel in control of my success in my career") was reworded to "I feel like I have little control of my career success."

Statistical Methods

Exploratory factor analysis (EFA) is often used by researchers to identify the number of dimensions in a set of items (Churchill, 1979). Gerbing and Anderson (1988) suggest that exploratory factor analysis can be a useful preliminary technique, but argue that only confirmatory factor analysis (CFA) provides a strict test of unidimensionality. Given that the CSES is designed as a unidimensional scale, and the competing models (based on existence of a wording method factor) are known, CFA is the most appropriate technique to use. However, researchers who find that their *a priori* factor structure is not confirmed by CFA will often turn to EFA to clarify what is happening (see Schriesheim's comments in Hurley et al., 1997). Furthermore, Williams argues that proper "use of eigenvalues as diagnostics in judgments about dimensionality … provides a more direct picture of dimensionality than goodness-of-fit measures used with CFA" (Hurley et al., 1997, p. 674). For these reasons, both EFA and CFA were used to assess dimensionality in each of the four samples.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Exploratory factor analysis framework. Velicer's minimum average partial correlations (MAP) method (Velicer, 1976) and parallel analysis (Hayton et al., 2004; Zwick & Velicer, 1986) were used to identify the number of factors underlying each of the four CSES datasets. Analysis was carried out in SPSS using syntax developed by O'Connor (2000). For parallel analysis, 200 permutations of the raw data set were used to estimate random eigenvalues, and a 95th percentile cutoff point was used to determine the number of factors. Because parallel analysis using principal axis factoring has been shown to be somewhat liberal, occasionally identifying trivial, negligible factors (Buja & Eyuboglu, 1992), principal components analysis was used.

Confirmatory factor analysis. Individual CSES items are measured at the ordered categorical (ordinal) level, and use of estimators such as maximum likelihood can result in incorrect parameter estimates, standard errors, and test statistics (T. A. Brown, 2006). Thus, procedures outlined by Jöreskog (2002) for the analysis of ordinal items were followed. Specifically, a polychoric correlation matrix was analysed, and diagonally weighted least squares (DWLS) estimation was used. This approach has been shown to produce accurate test statistics, parameter estimates and standard errors for ordinal data with sample sizes comparable to that of this study (Flora & Curran, 2004). The PRELIS (version 2.72, Jöreskog & Sörbom, 1996) program was used to generate the polychoric correlation matrix. Given the small amount of missing data, listwise deletion was used; this reduced the effective sample sizes to 263 (Sample 1), 178 (Sample 2), 2067 (Sample 3) and 1,876 (Sample 4).

The LISREL (version 8.72, Jöreskog & Sörbom, 1999) program was used for subsequent analysis. Seven different models (described below) were tested. Different aspects of model fit were evaluated using a set of six fit indices chosen on the basis of recommendations by Gerbing and Anderson (1993), Hu and Bentler (1999), and

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Tanaka (1993): (a) Satorra-Bentler chi-square statistic, (b) Nonnormed Fit Index (NNFI), (c) root mean squared error of approximation (RMSEA), (d) standardised root mean residual (SRMR), (e) expected cross validation index (ECVI), (f) Consistent Akaike Information Criterion (CAIC), and (g) comparative fit index (CFI). Consistent with the approach of Hu and Bentler[2], combinations of cutoff values of .95 for NNFI, .06 for SRMR, and .06 for RMSEA were used for assessing fit. ECVI values are useful for comparing non-nested models, with lower values being preferred to higher values; similarly, CAIC also allows comparison of models (low values implying better fit) but gives more weighting to model parsimony than ECVI.

Seven different measurement models for the CSES were evaluated; path diagrams are shown in Figure 1.

---

[2] My study uses DWLS estimation on polychoric correlations, while Hu and Bentler used maximum likelihood estimation of covariance matrices in their evaluation of fit indices. Flora and Curran (2004) found slight bias in chi-square estimates based on a similar form of estimation to DWLS implemented in Mplus. In the absence of any studies providing clear guidance on fit indices with DWLS, I have used Hu and Bentler's values.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

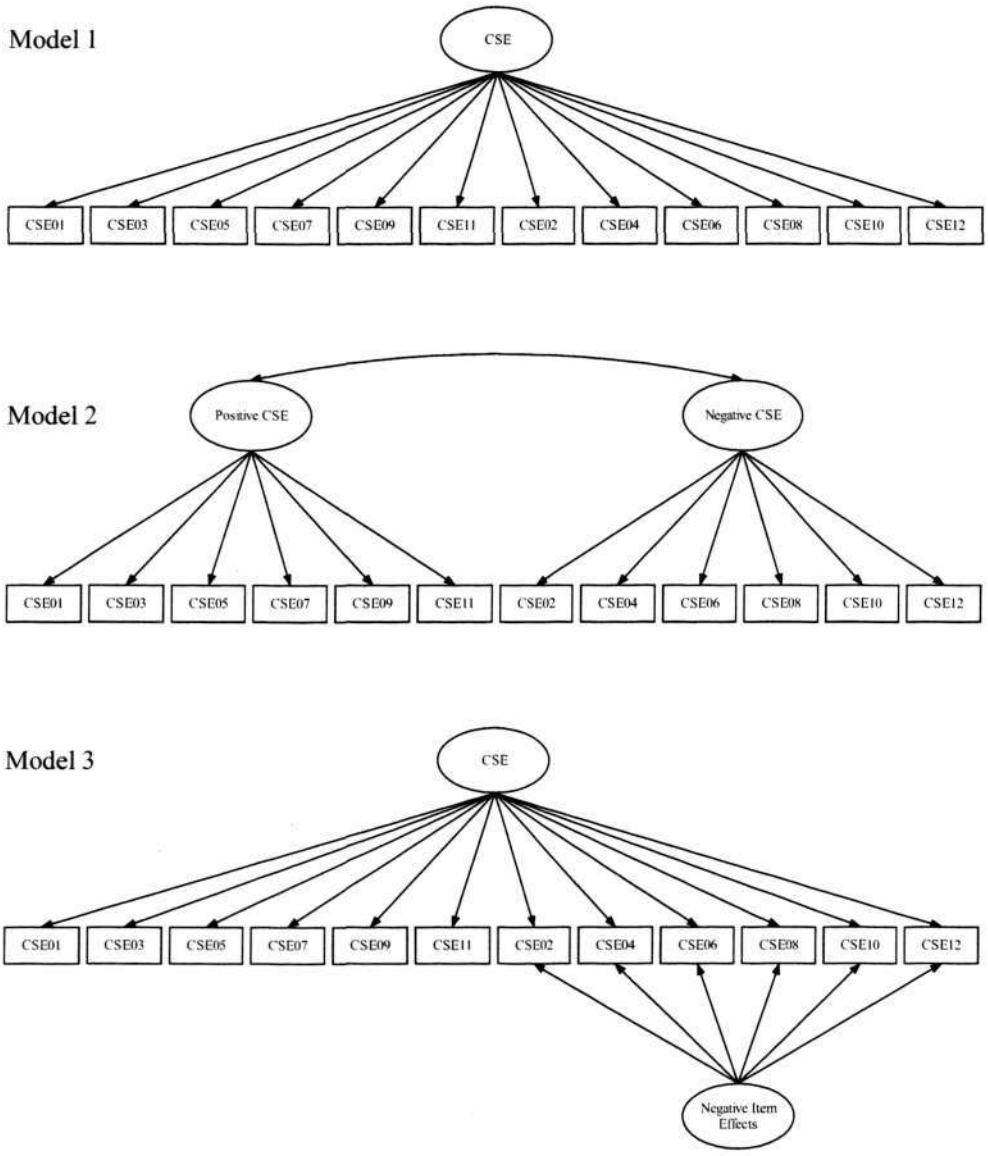


Figure 1. Alternative models of the Core Self-Evaluations Scale. Error variables omitted for clarity. Models 1 and 2 incorporate core self-evaluations constructs only. Models 3, 4 and 5 model method (item wording) effects through latent constructs. Models 6 and 7 model method effects through correlated errors.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

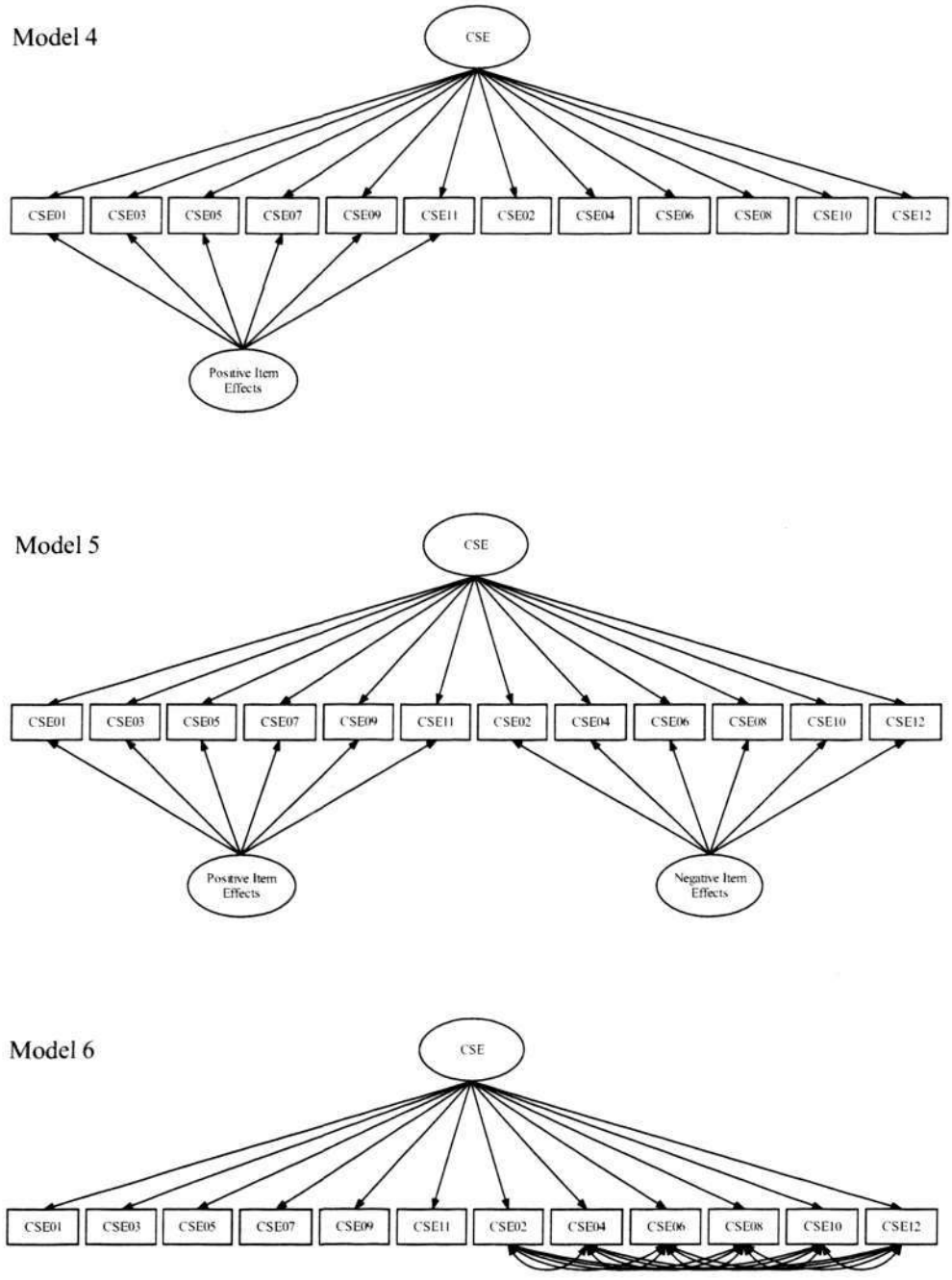


Figure 1 (Continued).

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

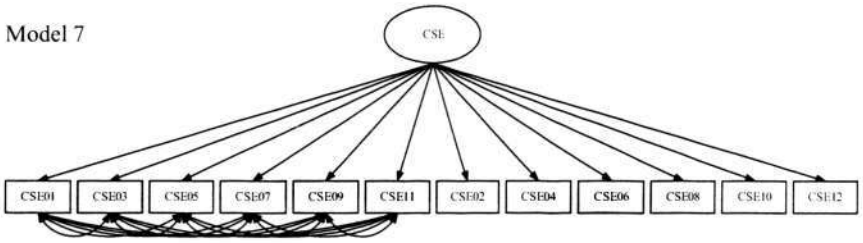


Figure 1 (Continued).

Model 1 posits a single latent construct representing the substantive trait, core self-evaluations. Model 2 includes two oblique factors, with separate factors for positively and negatively worded items. Models 3 and 4 use the CTCM approach, including a core self-evaluations factor (indicated by all items) and a second (orthogonal) factor based on item wording. In Model 3, the method factor is indicated by negative items while in Model 4 it underlies positive items. Model 5 is a combination of the preceding two models, with both a negative item and a positive item methods factor. The last two models (6 and 7) use the CTCU approach, with a single core self-evaluations factor indicated by all items. In Model 6, error terms among negatively worded items are allowed to covary while in Model 7, error terms among positively worded items covary.[3]

Full structural model. Figure 2 illustrates path diagrams for the structural models. In order to assess the need to model wording effects, Models 1 and 3 were compared in structural models involving self-mastery (Model 8), in-class teaching performance (Model 9), and critical thinking (Model 10). Models 8a, 9a and 10a included a single CSES latent variable (Model 1), while Models 8b, 9b, and 10b also

[3] An eighth model, allowing covariation within positive uniquenesses and within negative uniquenesses (combining Models 6 and 7) was considered. However, the large number of parameters estimated resulted in very low parsimony-adjusted fit (PNFI < .4), so it has not been included.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

included the item wording factor (Model 3). The critical thinking scale included two negatively worded items, so a method factor was included for these two items in Model 10b.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

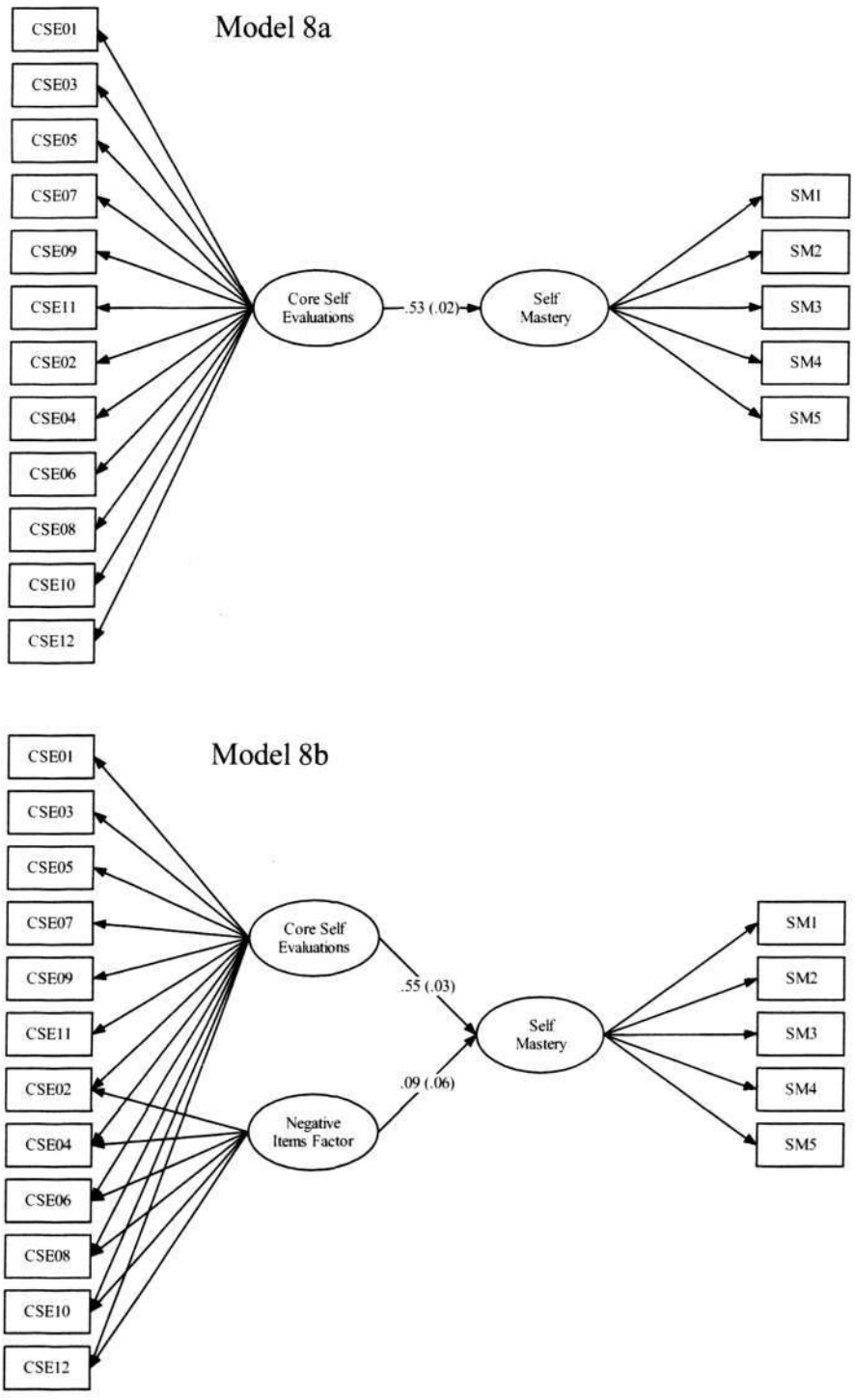


Figure 2. Structural models comparing one- and two-factor representations of the Core Self-Evaluations Scale (error variables omitted for clarity).

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

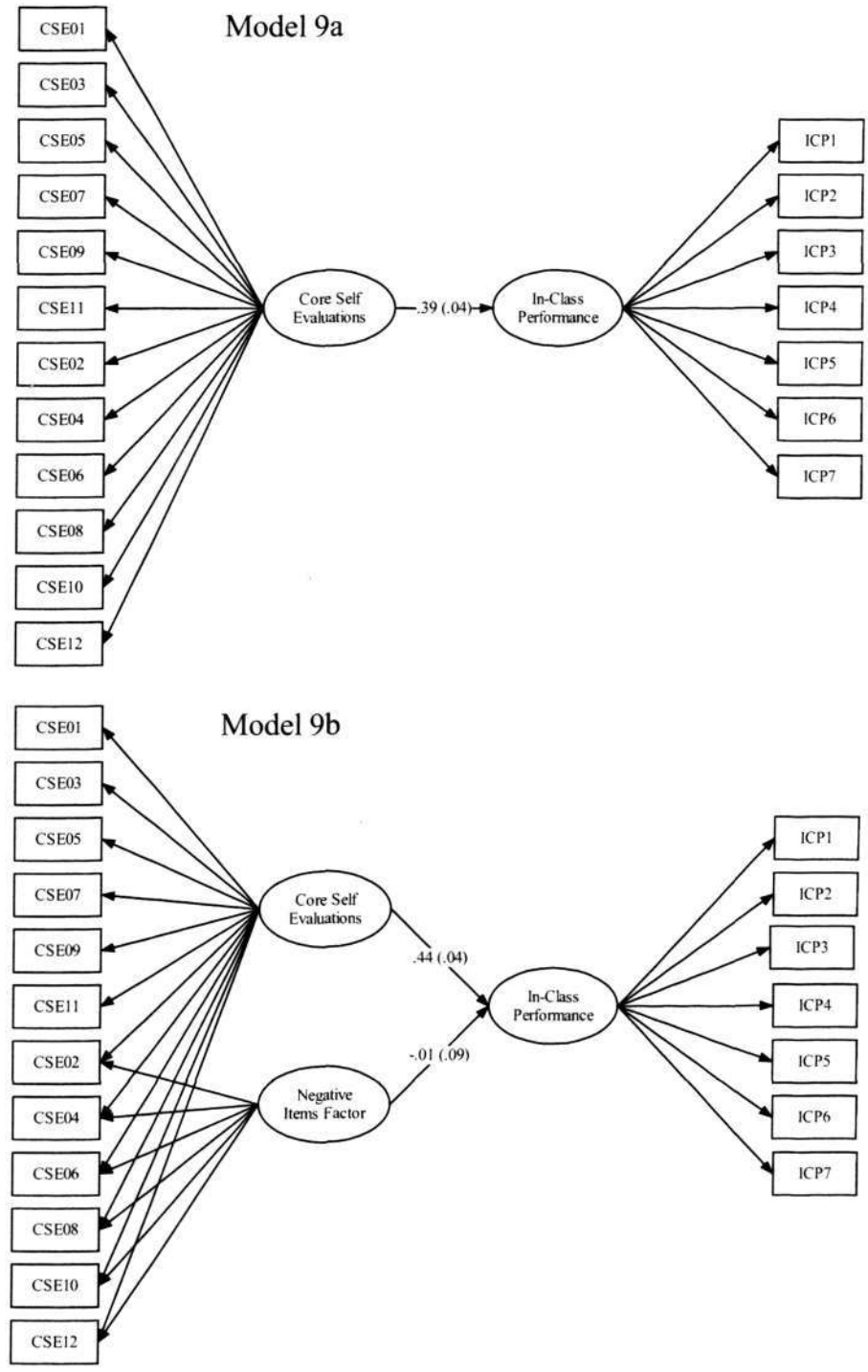


Figure 2 (continued)

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

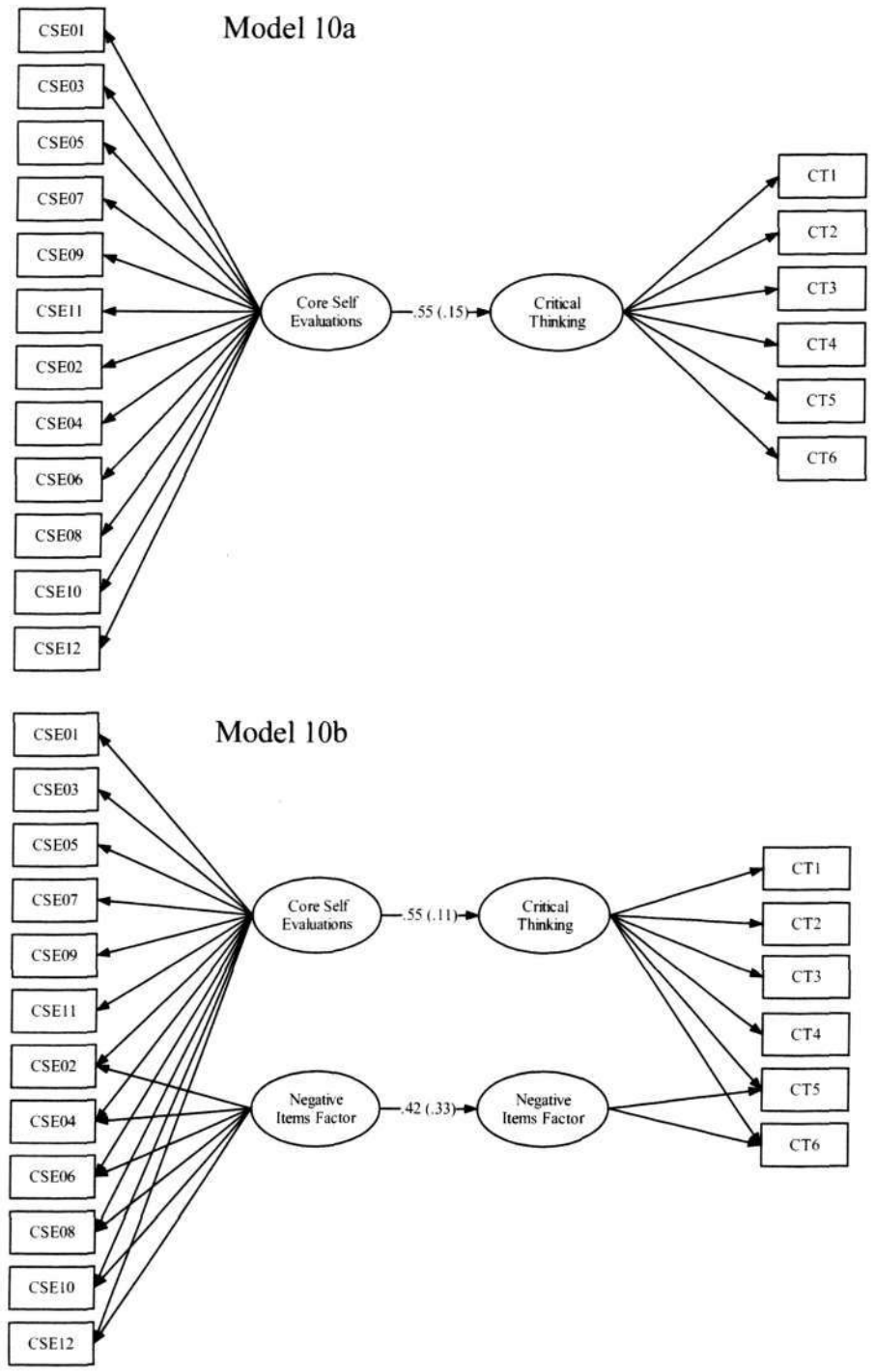


Figure 2 (continued)

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Results

Factor Structure

For all four samples, MAP and parallel analysis indicated the presence of two factors. Exploratory factor analysis (principal axis factoring with varimax rotation and two factors extracted) revealed a clear split between positively worded and reverse-coded items (Table 2). All the positively worded items defined one factor in each sample, with an average item loading .62, and no cross-loadings greater then .40. Similarly, negatively-worded items loaded on a separate factor, with an average item loading of .64. With the exception of Sample 3, no cross-loadings exceeded .40. In Sample 3, item 10 was the only exception; this negatively worded item had its highest loading (.43) on the factor defined by positive items (and had a loading of .35 on the negative item factor).

Fit indices for the seven confirmatory factor analysis models are reported in Table 3. Model 1, with all 12 items reflecting a single core self-evaluations factor showed the worst fit to data in all four samples. Analysis of residuals for Model 1 revealed a striking pattern. The fitted model consistently overestimated the correlations between different types of items (i.e. between a positively worded and a negatively worded item). Out of 36 such correlations, no sample had more than four non-negative residuals. Conversely, the model consistently underestimated the correlations between like items (i.e. between pairs of positively worded items, or pairs of reverse-scored items). Out of 30 such correlations, no sample had more than four negative residuals. The averages of the standardised residuals for fitted correlations between like items were 1.52, 1.47, 3.31 and 3.97 for Samples 1 to 4 respectively. Between different items they were -1.32, -1.37, -3.03, -3.54 respectively.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Table 2

Rotated Factor Loadings (Two-Factors Extracted)

	Sample 1		Sample 2		Sample 3		Sample 4	
Item	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2
CSE01	.70		.64		.60		.52	
CSE03	.73		.76		.67		.62	
CSE05	.70		.68		.48		.61	
CSE07	.54		.71		.55		.62	
CSE09	.60		.60		.55		.52	
CSE11	.68		.75		.54		.58	
CSE02		.67		.76		.68		.69
CSE04		.66		.71		.62		.64
CSE06		.67		.68		.39		.60
CSE08		.68		.60		.54		.59
CSE10		.54		.59	.43			.60
CSE12		.67		.76		.72		.72

Note: The highest loading for each item is shown; for clarity, any other factor loadings < .40 are not shown.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Table 3

Fit Indices for Alternative Models of Negative Item Wording Effects in the CSE Scale

Sample 1 (N = 263):

Model	*SB Chi-Square*	*df*	*NNFI*	*RMSEA (90% CI)*	*ECVI (90% CI)*	*CAIC*	*CFI*	*SRMR*
Model 1	246.03**	54	.88	.12 (.10–.13)	1.12 (.95–1.32)	403.8	.90	.10
Model 2[a]	89.44**	53	.98	.05 (.03–.07)	.53 (.45–.65)	253.7	.98	.06
Model 3	79.00**	48	.98	.05 (.03–.07)	.53 (.45–.64)	276.2	.98	.05
Model 4	84.07**	48	.97	.05 (.03–.07)	.55 (.47–.66)	281.2	.98	.06
Model 5[b]								
Model 6	50.42	39	.99	.03 (.00–.06)	.49 (.45–.58)	306.7	.99	.04
Model 7	66.86**	39	.98	.05 (.03–.07)	.55 (.48–.65)	323.2	.99	.05

Sample 2 (N = 178):

Model	*SB Chi-Square*	*df*	*NNFI*	*RMSEA (90% CI)*	*ECVI (90% CI)*	*CAIC*	*CFI*	*SRMR*
Model 1	243.31**	54	.86	.14 (.12–.16)	1.65 (1.39–1.94)	391.7	.89	.13
Model 2[c]	101.64**	53	.96	.07 (.05–.09)	.86 (.72–1.04)	256.2	.97	.08
Model 3	74.17**	48	.98	.06 (.03–.08)	.76 (.65–.91)	259.6	.98	.06
Model 4	89.06**	48	.97	.07 (.05–.09)	.84 (.71–1.01)	274.5	.98	.07
Model 5	37.85	42	1.00	.00 (.0–.04)	.64 (.64–.72)	260.4	1.00	.05
Model 6	60.67*	39	.98	.06 (.02–.08)	.78 (.69–.93)	301.8	.99	.05
Model 7	62.42*	39	.98	.06 (.03–.08)	.79 (.69–.94)	303.5	.99	.07

Sample 3 (N = 2067):

Model	*SB Chi-Square*	*df*	*NNFI*	*RMSEA (90% CI)*	*ECVI (90% CI)*	*CAIC*	*CFI*	*SRMR*
Model 1	1702.03**	54	.88	.12 (.12–.13)	.85 (.78–.91)	1909.2	.90	.09
Model 2[d]	743.15**	53	.95	.08 (.07–.08)	.38 (.34–.43)	959.0	.96	.06
Model 3	460.19**	48	.97	.06 (.06–.07)	.25 (.22–.29)	719.2	.98	.04
Model 4	630.63**	48	.95	.08 (.07–.08)	.33 (.30–.38)	889.7	.97	.05
Model 5	316.9	42	.97	.06 (.05–.06)	.19 (.16–.22)	627.8	.98	.03
Model 6	373.67**	39	.97	.06 (.06–.07)	.22 (.19–.25)	710.4	.98	.04
Model 7	479.74**	39	.96	.07 (.07–.08)	.27 (.24–.31)	816.5	.97	.05

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Sample 4 (N = 1876):

Model	*SB Chi-Square*	*df*	*NNFI*	*RMSEA (90% CI)*	*ECVI (90% CI)*	*CAIC*	*CFI*	*SRMR*
Model 1	1724.78**	54	.90	.13 (.12–.13)	.95 (.88–1.02)	1929.7	.91	.11
Model 2[e]	522.20**	53	.97	.07 (.06–.07)	.31 (.27–.35)	735.6	.98	.06
Model 3	397.40**	48	.98	.06 (.06–.07)	.24 (.21–.28)	653.5	.98	.04
Model 4	487.87**	48	.97	.07 (.06–.08)	.29 (.26–.33)	744.0	.98	.05
Model 5	266.15**	42	.98	.05 (.05–.06)	.18 (.15–.21	573.5	.99	.04
Model 6	222.23**	39	.98	.05 (.04–.06)	.16 (.14–.19)	555.2	.99	.04
Model 7	385.17**	39	.97	.07 (.06–.08)	.25 (.22–.28)	718.1	.98	.05

* $p < .05$ ** $p < .01$
[a] Correlation between Model 2 factors = .60.
[b] No admissible solution for this model; theta-delta not positive definite.
[c] Correlation between Model 2 factors = .57.
[d] Correlation between Model 2 factors = .68.
[e] Correlation between Model 2 factors = .58.
Notes: SB = Satorra-Bentler; NNFI = non-normed fit index; RMSEA = root-mean-square error of approximation; ECVI = expected cross-validation index; CAIC = consistent Akaike information index; GFI = goodness of fit index SRMR = standardised root-mean-square residual.

The systematic deviations of residuals from their expected value of zero imply a specification error in the model (Jöreskog, 1993). The particular pattern observed suggests that covariation between like items is not being adequately modelled by the single latent variable.

Models 2 through 7 all demonstrated improved fit over Model 1. All six models met or exceeded Hu and Bentler's (1999) suggested minimum of .95 for NNFI (or Tucker Lewis Index) and, in Samples 1, 3 and 4, had values of standardised root-mean-square residual (SRMR) less than or equal to .06. In Sample 2, Models 2, 4 and 7 exceeded the recommended .06 cut-off for SRMR; Models 3, 5 and 6 demonstrated acceptable fit.

Parameter estimates for all models are summarised in Table 4. In Model 2 (correlated factors for positive and negative items), all parameter estimates were positive and significant ($p < .05$) and the factors explained diverse amounts of item variance (R^2 ranged from .17 to .60 in Sample 1; from .23 to .91 in Sample 2; from .19

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

to .55 in Sample 3; and from .35 to .62 in Sample 4). The two factors correlated .60, .57, .68 and .58 in Samples 1 to 4 respectively, suggesting a moderate amount of shared variance.

Models 3 and 4 used the CTCM method to model item wording effects, with a separate factor for negative items (Model 3) and positive items (Model 4). Model 3 displayed better fit, having a smaller SRMR, smaller ECVI, and smaller CAIC than Model 4 in all samples. All Model 3 parameter estimates were positive and significant ($p < .05$) in all samples, with the exception of item CSE02 and the substantive core self-evaluations factor (.14, $t = 1.82$, $p = .07$ in Sample 1, and .14, $t = 1.30$, $p = .19$ in Sample 2). In Model 4, all parameters (in all samples) were positive and significant ($p < .05$) with the exception of CSE09 and the core self-evaluations factor in Sample 2 (.21, $t = 1.59$, $p = 11$). In Model 3, the amount of item variance explained, R^2, ranged from .23 to .57 (Sample 1); from .23 to .89 (Sample 2); from .20 to .61 (Sample 3); and from .34 to .62 (Sample 4). In Model 4 it ranged from .18 to .60 (Sample 1); from .28 to .77 (Sample 2); from .20 to .56 (Sample 3); and from .36 to .58 (Sample 4).

Model 5 (positive and negative item wording factors) was problematic. It did not reach an admissible solution in Sample 1 due to a non positive-definite residual (theta-delta) matrix. Several different parameter starting values were tried without success; in all cases, the interim solution provided indicated that five of the six positively-worded items had loadings close to zero on the positive item factor. This suggests that all the shared variance in positive items was accounted for by the substantive CSE factor.

Table 4
Standardised CFA Parameter Estimates for Alternative CSE Scale Models

Sample 1:

Items	*Model 1*	*Model 2*	*Model 3*	*Model 4*	*Model 5*[a]	*Model 6*	*Model 7*
CSE01	.59	.65	.65	.35 .59		.65	.35
CSE03	.65	.70	.70	.40 .60		.70	.40 .38
CSE05	.66	.72	.72	.46 .54		.72	.45 .26 .37
CSE07	.57	.63	.62	.40 .46		.62	.40 .34 .17[c] .25
CSE09	.52	.57	.57	.33 .47		.57	.34 .33 .22 .23 .19
CSE11	.71	.75	.75	.48 .56		.75	.47 .28 .35 .31 .29 .31
CSE02	.33	.42	.14[b] .55	.42		.13[c]	.42
CSE04	.61	.71	.45 .52	.71		.46 .37	.71
CSE06	.46	.54	.27 .54	.54		.26 .24 .16[c]	.54
CSE08	.66	.77	.50 .53	.77		.50 .20 .30 .37	.77
CSE10	.42	.48	.30 .38	.48		.31 .20 .12[b] .30 .18	.48
CSE12	.59	.69	.44 .51	.69		.44 .33 .30 .25 .22 .19	.69

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Sample 2:

Items	*Model 1*	*Model 2*	*Model 3*	*Model 4*	*Model 5*	*Model 6*	*Model 7*
CSE01	.59	.67	.66	.33[c] .60	.42 .56	.66	.34
CSE03	.69	.76	.76	.35 .76	.49 .71	.76	.35 .50
CSE05	.59	.67	.67	.34 .60	.51 .44[c]	.67	.33 .30 .46
CSE07	.68	.75	.75	.44 .60	.63 .40[c]	.75	.43 .40 .42 .35
CSE09	.41	.48	.48	.21[b] .49	.33[b] .41	.49	.20[b] .32 .35 .30 .23[c]
CSE11	.82	.95	.94	.63 .60	.87 .31[b]	.95	.63 .18[b] .43 .42 .42 .34
CSE02	.48	.56	.14[b] .74	.57	.18[b] .80	.15[b]	.57
CSE04	.64	.74	.44 .58	.73	.54 .48	.42 .45	.73
CSE06	.45	.54	.25[c] .54	.55	.33[c] .48	.25[c] .39 .30	.55
CSE08	.68	.74	.55 .42	.75	.69 .27[b]	.53 .23 .34 .28	.74
CSE10	.68	.73	.51 .46	.73	.61 .36	.51 .33 .26 .20[c] .23[c]	.73
CSE12	.61	.70	.32 .68	.70	.41 .63	.33 .53 .33 .38 .24 .35	.70

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Sample 3:

Items	*Model 1*	*Model 2*		*Model 3*		*Model 4*		*Model 5*		*Model 6*					*Model 7*					
CSE01	.60	.66		.66		.45	.49	.62	.18[c]	.66					.45					
CSE03	.55	.61		.62		.33	.67	.46	1.15	.62					.33	.35				
CSE05	.50	.56		.56		.36	.44	.49	.24	.56					.38	.14	.38			
CSE07	.68	.74		.73		.58	.38	.75	.02[b]	.74					.57	.22	.19	.14		
CSE09	.58	.64		.64		.44	.45	.61	.11[b]	.64					.43	.28	.26	.14	.19	
CSE11	.62	.68		.68		.47	.47	.65	.14[b]	.68					.47	.17	.30	.21	.25	.26
CSE02	.57		.62	.34	.63	.62		.36	.62	.35					.62					
CSE04	.59		.65	.40	.56	.65		.43	.54	.39	.35				.65					
CSE06	.40		.44	.29	.34	.44		.32	.31	.27	.20	.19			.44					
CSE08	.65		.71	.53	.42	.71		.56	.37	.52	.21	.29	.18		.71					
CSE10	.61		.66	.59	.19	.65		.63	.12	.58	.06	.10	.15	.13	.65					
CSE12	.66		.72	.43	.65	.72		.47	.63	.44	.44	.34	.21	.25	.11	.72				

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Sample 4:

Items	*Model 1*	*Model 2*		*Model 3*		*Model 4*		*Model 5*		*Model 6*					*Model 7*					
CSE01	.54	.64		.64		.40	.48	.44	.44	.64					.39					
CSE03	.60	.73		.73		.43	.59	.48	.55	.73					.43	.31				
CSE05	.54	.67		.67		.36	.61	.40	.58	.67					.36	.24	.46			
CSE07	.65	.79		.78		.49	.57	.54	.53	.78					.49	.30	.30	.32		
CSE09	.49	.59		.59		.33	.50	.37	.48	.59					.33	.26	.21	.22	.34	
CSE11	.52	.64		.64		.34	.58	.39	.56	.65					.35	.25	.28	.35	.34	.39
CSE02	.67		.72	.34	.69	.72		.53	.64	.36					.72					
CSE04	.61		.66	.32	.62	.66		.53	.45	.33	.46				.66					
CSE06	.72		.76	.51	.52	.76		.76	.13[c]	.49	.31	.30			.76					
CSE08	.73		.76	.55	.48	.76		.79	.10[b]	.53	.28	.31	.33		.76					
CSE10	.70		.74	.49	.52	.74		.74	.12[b]	.47	.24	.29	.40	.33	.74					
CSE12	.68		.73	.33	.72	.73		.55	.57	.34	.53	.43	.32	.30	.37	.73				

[a] Model 5 did not reach an admissible solution in Sample 1.
[b] Not significant.
[c] $p < .05$
Note: Unless otherwise noted, all parameters significant at $p < .01$. For Models 3 and 4, the first column shows parameters for the substantive CSE factor and the second column shows parameters for the item wording factor. For Models 5 and 6 the first column shows parameters for the CSE factor, while subsequent columns represent the lower half of the error covariance matrix. Even-numbered scale items are negatively worded.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

In addition to failing to reach a solution in Sample 1, Model 5 resulted in offending estimates in Sample 3. Item CSE03 had a standardised loading in excess of 1 on the positive item factor, while three other items (CSE07, CSE09, and CSE11) had zero loadings. In Sample 2, four item loadings were non-significant, and another three barely reached significance ($p < .05$). In contrast, the positive items in Sample 4 all had significant loadings on the positive factor; however, two negative items (CSE08 and CSE10) had zero loadings on the negative item factor, and a third (CSE06) barely reached significance ($\lambda = .13$, $p < .05$).

Models 6 and 7 used the CTCU method to model item wording effects, allowing negative errors (Model 6) or positive errors (Model 7) to covary. In Model 6, all parameters relating items to the core self-evaluations factor were positive and significant, with the exception of item CSE02 in Sample 1 (.15, $t = 1.42$, $p = .16$). The only non-significant correlated errors parameter was that between the error terms of CSE04 and CSE10 in Sample 1 (.12, $t = 1.64$, $p = .10$). In Model 7, all parameters relating items to the CSE factor were positive and significant, with the exception of item CSE09 in Sample 2 (.20, $t = 1.79$, $p = .07$). The only non-significant correlated errors parameter was that between the error terms of CSE01 and CSE11 in Sample 2 (.18, $t = 1.61$, $p = .11$). Correlating negative errors (Model 6) provided a better fit than correlating positive errors (Model 7) in all samples. Model 6 (with Sample 1 data) was the only model for which the hypothesis of exact fit ought not to be rejected ($\chi^2 = 50.42$, $df = 39$, $p = .10$).

Structural Models

Models 8a, 9a, and 10a (which modelled CSES as a unidimensional scale) failed to meet acceptable standards of fit (see Table 5 for fit indices). In addition to

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

failing the exact fit chi-square test, values of RMSEA and SRMR exceeded generally accepted guidelines (Hu & Bentler, 1999). With the introduction of an item-wording factor (models 8b, 9b and 10b), the chi-square test became insignificant, indicating that each of the model-implied correlation matrices no longer diverged significantly from the observed matrices.

Table 5

Fit Indices for Alternative Structural Models Incorporating CSE Scale

Model	*SB Chi-Square*	*df*	*NNFI*	*RMSEA (90% CI)*	*ECVI (90% CI)*	*CAIC*	*SRMR*
Model 8a	241.39**	117	.97	.06 (.05–.07)	1.20 (1.04–1.39)	477.8	.08
Model 8b	120.39	109	1.00	.02 (0–.04)	.80 (.75–.92)	409.4	.05
Model 9a	269.53**	150	.97	.07 (.05–.08)	1.99 (1.75–2.27)	516.6	.10
Model 9b	113.84	142	1.01	0	1.35 (1.35–1.35)	410.3	.06
Model 10a	347.97**	133	.93	.08 (.07–.09)	1.67 (1.47–1.90)	596.5	.08
Model 10b	129.45	123	1.00	.01 (0–.04)	.89 (.86–1.01)	443.4	.05

** $p < .01$

Notes: Listwise deletion reduced sample sizes to 262 (Model 8), 177 (Model 9), and 255 (Model 10). SB = Satorra-Bentler; NNFI = non-normed fit index; RMSEA = root-mean-square error of approximation; ECVI = expected cross-validation index; CAIC = consistent Akaike information criterion; SRMR = standardised root-mean-square residual.

Model 8 tested the relationship between core self-evaluations and self-mastery. Values for structural model parameter estimates (with standard errors in parentheses) are shown in Figure 2. The coefficient of the path from core self-evaluations to self mastery was significant (.55, $p < .01$), but the path from the negative items factor to self mastery was non-significant. The squared multiple correlation (R^2) for self mastery was .47.

Model 9 tested the relationship between core self-evaluations and in-class teaching performance. The coefficient of the path from core self-evaluations to in-class

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

performance was significant (.44, $p < .01$) but that from the item wording factor was non-significant. For in-class performance, $R^2 = .30$.

Model 10 tested the relationship between core self-evaluations and critical thinking. Because the critical thinking measure included two negatively-worded items, model 10b included an item-wording factor on both the exogenous and endogenous sides of the model. Of the two structural paths, only the coefficient for the path from core self-evaluations to critical thinking was significant (.55, $p < .01$). For critical thinking, $R^2 = .58$.

Discussion

This study has clearly demonstrated the existence of item wording effects in the CSES, using two different methods (MAP and parallel analysis in the exploratory factor analysis framework, and comparison of fit of competing models in the CFA framework). All of the models incorporating wording effects showed improved fit over the single factor model (Model 1). Modelling wording effects for negative items produced better fit than modelling positively worded item effects—a finding that is consistent with other research on scales containing positively and negatively worded items (DiStefano & Motl, 2006).

In terms of goodness of fit measures, there is no clear winner between the CTCM (Model 3) and CTCU (Model 6) approach to modelling negatively worded items. However, Lance et al. (2002) identify a number of theoretical and substantive reasons for preferring the CTCM approach. For example, this approach allows for unbiased estimation of both method and trait variance components. Researchers can remove the item wording variance from scale items through specifying a method factor, thereby controlling for method effects when estimating relationships between

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

the substantive core self-evaluations latent variable and other constructs. While modelling both a positive and a negative item wording factor produced an even closer-fitting model (when a solution existed), problems were evident with non-significant parameters. It appears that using just one wording factor is sufficient to successfully model the item covariances.

Items CSE02 ("Sometimes I feel depressed") and CSE06 ("Sometimes I do not feel in control of my work") were the worst indicators of the core self-evaluations factor in Samples 1 to 3, with loadings ranging between .14 and .34. Item CSE02 also had a relatively low loading in Sample 4 (.34), but CSE06 performed better (.51), perhaps because this item was reworded in the Belgian sample. The performance of these items in other samples needs to be assessed; removing or replacing them with items that better tap the substantive construct (after controlling for item wording) is likely to improve the overall scale.

The analysis of models 8, 9, and 10 demonstrated that researchers cannot ignore the item wording factor. The amount of model misspecification introduced by failure to accurately model the item wording factor would have led researchers to reject all three of the models. The poor fit would preclude any interpretation of or reliance on obtained model parameters. Once the wording factor was included, model fit improved across all indices, and chi-square tests of overall fit became insignificant. While the size of structural parameters connecting core self-evaluations with dependent variables remained almost the same across each model pair, the acceptable level of model fit in the item wording factor models now allows these values to be interpreted with confidence.

The item wording factor had no significant relationship with the dependent variable in any of the three models. While the estimated parameter for the path

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

between the two wording factors in model 10b equalled .42, the high standard error of this value (.33) rendered it insignificant. It therefore appears, based on these two samples and three dependent variables, that the item wording factor contains no substantive content.

The generalisability of this finding will need to be tested with additional studies using different dependent variables. There is some evidence that negative wording factors in scales measuring different content areas can be correlated across the different content areas. For example, Horan et al. (2003) modelled negative item factors for three different scales in the same model (self-esteem, attitude towards schooling, and locus of control). Correlations among these wording factors were all significant (ranging from .31 to .43) and comparable in size to two of the three correlations between the substantive factors.

Calculation of a single CSES score based on the sum of all items will include variance from both the item wording factor and the core self-evaluations factor. Subsequent use of this score in correlation or regression analysis will confound the two variance components. CSES scores may therefore demonstrate upwardly biased estimates of correlations with other scales which include item wording effects (if these aren't modelled separately). In testing convergent validity of the CSES, Judge et al. (2003) correlated the CSES with Rosenberg's (1965) self-esteem scale, which has been shown to have a strong negative-wording factor (Marsh, 1996), and with a generalised self-efficacy scale (Judge et al., 1998) which included negatively-worded items. It is therefore possible that the CSES's correlation with these scales was overestimated due to correlations between unmodeled item wording factors.

While use of item parcelling to combine CSES items into a smaller set of indicators will allow better fit of structural models, the improved fit will be misleading

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

(and parameter estimates may be biased) because of the presence of both a wording factor and a substantive factor in the scale (Bandalos, 2002).

Quilty et al. (2006) found significant correlations between a negatively-worded item factor and both conscientiousness (.14, $p < .05$) and emotional stability (.25, $p < .01$). Thus, more conscientious and emotionally stable respondents were less likely to endorse negative items. If such a relationship also holds with the CSES scale, then respondents low on emotional stability are likely to have their level of core self-evaluations deflated below the true level as a result of the contribution of the item wording factor.

In this study, dependent variables were measured in the same questionnaire as the CSES items. While response scales were varied, and the scales presented in different sections of the questionnaire to reduce common method bias (Podsakoff, Mackenzie, Lee, & Podsakoff, 2003), some bias is likely to remain. Any such bias, common to all items in CSES and the other questionnaires, will tend to inflate the estimate of the path from substantive core self-evaluations to the dependent variable in each model. The estimates reported in this paper may therefore overestimate the relationship between core self-evaluations and self-mastery, in-class performance, and critical thinking respectively. However, the possible existence of method bias does not invalidate the model comparisons. If present, it will operate the same way in each of the six models. Thus, the finding of significant differences in the fit of structural models with and without an item wording factor is not undermined by possible common method bias.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Conclusions

The CSES has a clear, stable negative wording factor; it emerged in four samples that differed markedly in terms of gender composition, nationality and occupation. The four samples also used different scales (five points versus seven points), One was a translated (Dutch language) version, and used different item ordering (grouping positive and negative items together, rather than alternating them). Failure to explicitly model the wording factor in structural equation models undermines model fit, and may lead researchers to falsely reject hypothesised relationships. This factor does not invalidate the CSES, but argues for care in its use and interpretation. In particular, users of the scale must separately model the negative wording factor in future studies, so that relationships between core self-evaluations and other variables are estimated accurately.

The low validity of items CSE02 and CSE06 as measures of core self-evaluations in these samples (after controlling for wording effects) suggests they may contribute little to the scale. If this finding is confirmed in new samples, the items should be replaced.

Some researchers have argued that negatively-worded items should not be included in a scale; Marsh (1996), for example, suggested keeping them in a questionnaire to reduce acquiescence bias, while dropping them from subsequent analysis. Unfortunately, dropping the negative items from the CSES will reduce its coverage of the core self-evaluations domain. Furthermore, with the exception of item 2 ('Sometimes I feel depressed') in Samples 1 and 2, all of the negatively-worded items continued to have significant loadings on the substantive core self-evaluations construct even when the negative item factor was included in the model. This argues against dropping negatively-keyed items from the scale.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Modelling an item wording factor is not without cost; it uses up degrees of freedom in a model, and adds to the complexity of the relationships specified. It is therefore important to develop a better understanding of the processes underlying the emergence of such a factor.

Chapter Summary

This chapter has demonstrated the existence of an item wording factor in the recently published Core Self-Evaluations Scale. The factor is present in samples from different populations, and is associated more with negatively worded than with positively worded items. The factor is sufficient to destroy the fit of a single factor model in both confirmatory factor analysis and structural equation models.

Having established the presence of a wording method factor in the CSES samples, it is now possible to investigate whether careless responding might be the process underlying the emergence of this factor. In the next chapter, I identify cases in these samples who exhibit different patterns of inconsistent responding consistent with carelessness. I then use this information to develop a realistic simulation of carelessness in the context of a Monte Carlo study of the effect of careless responding on CFA model fit.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

CHAPTER FOUR

MONTE CARLO SIMULATION OF CARELESS RESPONDING

Why do scales which include negatively-keyed items often produce a spurious factor, related to item wording? Careless responding, perhaps induced by the complexity of negative items, is a potential cause. Simulation studies (Schmitt & Stults, 1985; Woods, 2006) have provided insights into this process, but have assumed that a careless respondent will be careless when answering all negatively-keyed items. This assumption is unrealistic, and as the first study described in this chapter will illustrate, more complex response patterns are needed.

The second study described in this chapter will use a Monte Carlo simulation to demonstrate how careless responding creates spurious factors, and will examine their effect on CFA model fit.

Item Wording Factors

Several explanations have been proposed for the mechanism by which negatively-keyed items give rise to a method factor. Samelson and Yates (1967) suggested that negated items may have different 'neutral points' than supposedly-equivalent positive items, and Spector (1992) demonstrated how such differences in distributional properties of scale items can give rise to artifactual factors. Personality has also been suggested as a possible underlying cause (DiStefano & Motl, 2006), while Marsh (1986) identified the linguistic complexity of some negated items as a contributing factor.

While each of these mechanisms may contribute to the problem in specific scales, careless or confused responding is a simple and more general alternative explanation. If a subset of respondents fail to recognise or accurately interpret

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

negatively-keyed items, and respond to them in the same way they respond to positive items, negatively-keyed items are likely to exhibit greater response-distribution heterogeneity than positive items. If a negatively-keyed item measured on a 7 point rating scale has a mean response of 3 and is positively-skewed, the mean of the reverse-coded scale will be 5 (and the distribution will be negatively-skewed). For the subset of careless respondents, reverse-coded scores will have a mean of 3 and positive skew. As shown by Bernstein and Teng (1989), items with such mixed distributions are likely to form a separate factor from the more homogeneously distributed positive items.

Careless Responding

In Chapter 2, I summarised evidence regarding careless responding in relation to negatively worded items (e.g., Barnette, 1995 ; Holden & Fekken, 1990). Schmitt and Stults (1985) demonstrated that careless responding to negatively-keyed items can cause the emergence of a wording-related principal component in item-level factor analysis, but (as noted on p. 46) their early methodological approach precludes generalising their findings to the CFA framework.

Woods (2006) used confirmatory factor analysis to investigate the effect of careless responding to reverse-worded items, but did not address all of the limitations in the Schmitt and Stults (1985) study. Her study used dichotomous items, which are seldom used in organisational research scales[4]. She modeled carelessness in the same way as Schmitt and Stults (i.e., a careless respondent answered all negative items carelessly). Recognising this as a limitation of her study, she suggested that future

[4] Fewer than 3% of the scales described in the British Telecom Occupational Psychology Unit survey item bank (1984a, 1984b) have dichotomous scales.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

research needs to investigate the effect of respondents answering carelessly in varying degrees.

The studies in this chapter directly address limitations of the above studies. In the first study, empirical data are used to identify examples of response patterns consistent with carelessness or confusion. This information is then used to guide the simulation in the second study. Carelessness is modeled in a realistic manner, with varying proportions of careless respondents differing in the probability with which they answer a negatively-keyed item carelessly. Normal and skewed distributions of item responses (on 5 and 7 point scales) are analyzed using techniques appropriate for non-normal data. Objective identification of the number of factors underlying the item correlations is made, and confirmatory factor analysis is used to compare the fit of one and two factor solutions.

Study One – Patterns of Careless Responding

Woods (2006) stated the need for simulations which used more realistic models of careless responding, allowing for respondents who answer with varying degrees of carelessness. In order to do this, it seems necessary to start with real sets of data, to get a sense of how carelessness might be reflected in actual response patterns.

Carelessness (and what Barnette 1995, has called 'nonattending responding') can affect data in many different ways. My focus is on carelessness which manifests itself by generating responses to negatively worded items that are inconsistent with responses to positively worded items in the same scale. Nunnally (1978) suggests looking for evidence of incompatible responses in summated rating scales as a technique for identifying carelessness (although he points out that such evidence is

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

only circumstantial—people's responses, in the absence of carelessness, are never perfectly uniform).

While visual sorting and scanning of the dataset can identify extreme cases, less obvious cases (and larger datasets) require the use of statistical methods. For example, Barnette (1995) used a chi-square test to compare agree versus disagree responses to positive and negatively keyed items, finding a greater than chance proportion of respondents answering inconsistently between positively and negatively worded items. However, this approach to identifying carelessness with respect to negative wording involves considerable simplification (e.g. collapsing the four-point rating scale into two categories), and cannot be used with an odd number of rating points.

An alternative approach comes from consideration of the likely effect of careless responses on the fit of confirmatory factor analysis models. If the artifactual negative wording factor (in a scale measuring only one substantive construct) is a function of careless responding to negatively worded items, then removing cases with careless responses will result in improved fit for the one factor model. Individual cases which result in the biggest improvement in fit when left out of the analysis are clearly less consistent with the hypothesised model, and are likely to include anomalous (careless or confused) responses.

Rensvold and Cheung (1999) describe a method for implementing this technique using the jackknife procedure. They discuss several procedures which have been proposed for identifying influential cases in structural equation modelling, but point out that such techniques typically rely on comparing each case with the other cases, or assessing their effect on the covariance matrix. Rensvold and Cheung argue

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

that it is necessary to identify influential cases (cases which disproportionately influence model fit or parameters) in the context of the actual model being tested.

The jackknife procedure begins with a sample of *N* observations. By dropping the first case, a new sample (of *N-1* cases) is created. The second case is then dropped from the original sample, creating another sample of *N-1* cases. The process continues, dropping a different case each time, until *N* new samples have been created, each consisting of *N-1* cases. The appropriate confirmatory factor analysis model is fitted to each of these samples in turn. The *N* models can then be rank ordered according to model fit. Jackknife samples ranked near the top and bottom of this list identify potential influential cases. Rensvold and Cheung (1999) demonstrated this approach by using simulated data, generated to match a predetermined model (3 correlated factors, each indicated by 4 items measured on a 7-point rating scale). Two items for each factor were purportedly measured with reverse-keyed items; one anomalous case was introduced into the sample, by providing ratings of '2' on these (recoded) items and '6' for the other two items on each factor. This simulated a respondent who answered '6' to all items, regardless of item wording direction. A population of 25,000 cases was generated, and 250 separate samples (of 250 cases each, plus the anomalous case) were used in their analysis. The jackknife procedure successfully identified the introduced anomalous case as the most influential one in each sample.

Rensvold and Cheung (1999) were interested in identifying influential cases of any kind – those that inflated as well as reduced overall model fit. In my study, the focus is on identifying inconsistent responding to positive and negative items. This requires use of model parameters rather than overall fit (as the overall fit index will be influenced by any case which conflicts with the hypothesised model, not just the inconsistent response patterns I am interested in).

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

If factors are defined by item wording (a positive item factor correlated with a negative item factor), then perfect correlation between the two factors results in a model which is equivalent to the one factor model. At the opposite extreme, zero correlation between the two factors indicates that the two sets of items are measuring independent constructs. By rank ordering jackknife samples according to phi (the correlation between factors), the cases which most clearly define separate positive and negative item factors will be identified. The highest jackknife-sample phi values will be associated with cases which have consistency within like items, but inconsistent responses between positive and negative items – the cases which (when left out of the analysis) shift phi closer towards the one-factor model value.

The use of phi will favour identification of cases with a high proportion of negatively worded items answered anomalously. In order to find whether respondents have answered only one or two items inconsistently, it is necessary to consider item parameter values in a two factor model. When factors are defined by positive versus negative items, both factors contain variance related to the substantive construct, as well as wording effects. In order to separate out the wording effects, it is necessary to model a substantive construct which is indicated by all scale items, and a separate (orthogonal) factor which captures variance unique to the negative items. Jackknife models can then be ranked in order of Λ_x values for negative items. Specifically, for an individual item, those jackknife samples ranked high according to values of Λ_x for the substantive factor, and ranked low on Λ_x for the wording factor identify influential cases in regards to that item's definition of the wording factor.

In summary, I will use two models to identify inconsistent response patterns. The value of phi (correlation between a positive item and negative item factor, Model 2 in *Figure 1*) will identify the most influential cases with inconsistent responses. To

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

determine whether some cases answer only one or two negative items inconsistently, I will use individual item loadings on the wording and substantive factors (Model 3 in *Figure 1*).

Method

The jackknife procedure was applied to all four Core Self-Evaluations Scale (CSES) datasets described in Chapter 3. Detailed analysis will be provided for the largest sample (Sample 3). Standard techniques for identifying careless responses (e.g., sorting the cases on responses and looking for patterns) become more difficult with large samples (Rensvold & Cheung, 1999, p. 305). Furthermore, a large sample is likely to include the full range of inconsistent patterns likely to be encountered in practice.

Sample 3 comprised 2,088 employed adults undertaking part-time tertiary study in New Zealand (41.6% male); ages ranged from 19 to 69, with a mean of 38.5 (SD = 9.5). Responses to the CSES were made on a seven-point scale (Pajo et al., 2005). Cases with missing values for any item (1% of sample) were deleted before analysis, resulting in a final sample of 2,067 cases.

I wrote a Python program[5] to manage the generation of jackknife samples, submission of data and syntax to PRELIS and LISREL, and the collation of LISREL fit indices and parameter values by SPSS. With 2,067 cases, a total of 2,067 separate jackknife samples were created. PRELIS was used to generate polychoric correlation and asymptotic covariance matrices for each sample, and confirmatory factor analysis

---

[5] An open source computer programming language (www.python.org); SPSS Inc. have created modules which allow integration of python programming and SPSS syntax (www.spss.com/devcentral). Python programs used for this dissertation are included in Appendix A.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

was conducted in LISREL using diagonally weighted least squares estimation. As discussed in the preceding section, two models were estimated – a two factor model with factors defined by positive versus negative items, and a two factor model with one substantive factor (all items) and one orthogonal wording factor (negatively worded items). These correspond to the path diagrams for Models 2, and 3 in Figure 1 (p. 67). Summaries of model fit indices and parameter values were imported into SPSS to facilitate subsequent analysis and case identification.

Results

In discussing the pattern of responding in cases, I will present individual responses to the 12 CSES items as a string of digits corresponding to responses on each item (from 1 to 12). The numbers represent ratings after negatively keyed items have been recoded; odd-numbered items are positively keyed, while even-numbered items are negatively keyed. Thus, a respondent who answered all items '6' on the original 7-point scale would be represented by the response pattern '6 2 6 2 6 2 6 2 6 2 6 2', indicating considerable inconsistency between responses to positive and negative items.

Influential Cases in the Two-Factor (Positive and Negative) Model

Table 6 lists the fifteen most influential cases in regards to correlation between the positive item and negative item factors. The first few cases are consistent with careless or confused responding to most of the negative items (e.g. Case 1343: '6 1 7 1 7 1 5 3 5 1 7 1' and Case 1435: '5 1 7 1 7 1 7 2 6 2 6 1 6 1'). As is typical with measures of self esteem (Taylor & Brown, 1988), most positive item ratings are above the mid-point of the scale, but there are exceptions. For example, the 9th ranked case

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

has responses '4 7 3 7 3 5 4 5 2 4 4 6.' Here, several of the negatively-worded items are inconsistent by suggesting higher levels of core self-evaluations than suggested by positive item responses.

Table 6
Most Influential Cases in the Two Factor Model[a] *(Ranked in Descending Order of Phi)*

		CSES Items[b]												
Rank	Case	1	2	3	4	5	6	7	8	9	10	11	12	Phi[c]
1	1343	6	1	7	1	7	1	5	3	5	1	7	1	.6840
2	1435	5	1	7	1	7	2	6	2	6	1	6	1	.6840
3	377	7	3	7	2	7	3	7	6	7	2	7	2	.6839
4	1046	7	2	7	6	7	4	7	7	7	1	7	4	.6830
5	202	6	7	5	7	6	3	5	7	1	7	4	7	.6826
6	1476	6	7	6	7	6	2	2	7	2	6	2	7	.6825
7	1528	7	3	7	3	6	1	7	2	5	6	7	3	.6825
8	1856	2	6	4	4	7	7	2	1	1	7	1	7	.6825
9	43	4	7	3	7	3	5	4	5	2	4	4	6	.6824
10	1348	6	3	6	2	6	3	6	1	7	1	6	3	.6824
11	1683	6	2	7	3	6	2	6	2	7	3	6	3	.6823
12	891	6	3	7	3	6	1	7	3	6	4	6	2	.6823
13	1294	7	2	6	3	6	1	7	3	6	4	6	3	.6823
14	592	5	6	4	7	5	7	6	7	5	7	5	7	.6823
15	618	1	7	6	7	5	3	6	7	4	7	6	7	.6822

[a] One factor defined by positively worded items (1,3,5,7,9,11) and the second by negatively worded items (2,4,6,8,10,12).
[b] Odd numbered items are positively worded; even numbered are negatively worded. All items were rated on a 7-point scale, and negatively worded items have been recoded.
[c] Correlation between the two latent factors.

Cases Which Influence Individual Item Parameters

Ranking cases according to their impact on item loadings in the two-factor (substantive and method) model allows for identification of cases which have inconsistent responses to only one item. Table 7 to Table 12 list the fifteen most influential cases for each of the negative items in turn. In every table, there is at least one item which has a single anomalous response on the respective item; for example:

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Item 2	Case 1034	7 **1** 7 6 6 6 6 7 7 7 7 7
Item 4	Case 936	7 6 7 **1** 6 6 6 4 7 7 7 7
Item 6	Case 1988	7 6 7 6 6 **1** 6 6 7 6 7 7
Item 8	Case 514	7 6 7 7 7 6 7 **2** 7 7 7 7
Item 10	Case 227	6 6 7 6 6 6 6 6 7 **1** 6 7
Item 12	Case 1516	7 7 6 7 6 7 7 7 7 7 7 **1**

Between these extremes (all negatively keyed items inconsistent with positive items, and only one negatively keyed item inconsistent), it is possible to identify cases with 2, 3, 4 or 5 inconsistent negative responses. For example:

Table 7	Case 1951	6 **1** 7 7 7 6 7 7 6 4 7 **1**
Table 7	Case 840	6 **1** 7 **3** 6 **2** 7 7 7 7 7 7
Table 9	Case 459	7 **3** 7 **2** 7 **3** 6 5 7 6 7 **3**
Table 6	Case 377	7 **3** 7 **2** 7 **3** 7 6 7 **2** 7 **2**

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Table 7

Most Influential Cases in the Two Factor Model[a] *(Ranked in Descending Order of LX(2,1)*[b] *and Ascending Order of LX(2,2)*[c]

		CSES Items													
Rank	Case	1	2	3	4	5	6	7	8	9	10	11	12	LX(2,1)	LX(2,2)
1	374	7	1	7	7	7	7	7	7	7	7	7	1	.3399	.6263
2	197	7	1	7	5	7	7	7	7	6	7	7	4	.3393	.6280
3	1951	6	1	7	7	7	6	7	7	6	4	7	1	.3388	.6270
4	1920	7	1	7	7	5	7	7	7	7	7	7	6	.3387	.6304
5	840	6	1	7	3	6	2	7	7	7	7	7	7	.3387	.6288
6	43	4	7	3	7	3	5	4	5	2	4	4	6	.3386	.6274
7	1798	3	7	5	3	4	3	3	5	3	2	4	6	.3386	.6282
8	412	7	1	7	4	6	1	7	6	6	6	7	6	.3386	.6280
9	368	7	1	7	7	6	4	7	2	7	6	6	1	.3385	.6266
10	1034	7	1	7	6	6	6	6	7	7	7	7	7	.3382	.6312
11	1648	7	1	7	1	6	3	7	7	3	7	7	7	.3379	.6284
12	1476	6	7	6	7	6	2	2	7	2	6	2	7	.3378	.6277
13	362	2	7	5	6	5	6	2	2	3	3	7	6	.3378	.6279
14	911	7	2	7	2	6	2	7	6	7	7	7	4	.3377	.6279
15	1661	2	7	3	4	5	3	4	6	4	2	6	7	.3377	.6280

[a] One factor defined by all items (1–12) and the second (orthogonal) factor by negatively worded items (2,4,6,8,10,12).

[b] Loading of CSES item 2 on the substantive CSE latent factor.

[c] Loading of CSES item 2 on the negative wording method latent factor.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Table 8
Most Influential Cases in the Two Factor Model[a] *(Ranked in Descending Order of LX(4,1)*[b] *and Ascending Order of LX(4,2)*[c]

		CSES Items													
Rank	Case	1	2	3	4	5	6	7	8	9	10	11	12	LX(4,1)	LX(4,2)
1	43	4	7	3	7	3	5	4	5	2	4	4	6	.4034	.5588
2	1215	7	4	7	1	6	3	6	7	7	7	7	7	.4029	.5618
3	1141	6	3	7	1	5	4	7	7	7	7	7	5	.4028	.5610
4	377	7	3	7	2	7	3	7	6	7	2	7	2	.4027	.5592
5	1648	7	1	7	1	6	3	7	7	3	7	7	7	.4027	.5599
6	936	7	6	7	1	6	6	6	4	7	7	7	7	.4027	.5616
7	116	2	5	3	7	6	2	3	7	1	5	6	7	.4025	.5592
8	83	1	2	5	7	6	2	4	3	2	2	5	6	.4024	.5616
9	1016	1	1	6	7	6	1	1	7	1	1	6	1	.4024	.5630
10	911	7	2	7	2	6	2	7	6	7	7	7	4	.4024	.5595
11	397	7	4	7	1	5	2	7	5	5	5	7	1	.4024	.5588
12	1904	7	3	7	2	7	3	6	4	7	7	7	6	.4024	.5596
13	459	7	3	7	2	7	3	6	5	7	6	7	3	.4023	.5595
14	1639	7	3	7	1	6	6	5	3	7	7	7	6	.4023	.5603
15	760	4	3	6	7	3	1	4	6	1	6	5	4	.4022	.5611

[a] One factor defined by all items (1–12) and the second (orthogonal) factor by negatively worded items (2,4,6,8,10,12).
[b] Loading of CSES item 4 on the substantive CSE latent factor.
[c] Loading of CSES item 4 on the negative wording method latent factor.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Table 9

Most Influential Cases in the Two Factor Model[a] *(Ranked in Descending Order of LX(6,1)*[b] *and Ascending Order of LX(6,2)*[c]*)*

		CSES Items													
Case	Rank	1	2	3	4	5	6	7	8	9	10	11	12	LX(6,1)	LX(6,2)
1	1537	2	4	5	2	1	7	1	3	4	1	6	6	.2955	.3340
2	1856	2	6	4	4	7	7	2	1	1	7	1	7	.2945	.3327
3	1040	2	1	2	2	2	6	2	6	2	6	1	7	.2944	.3325
4	503	4	1	6	1	6	7	1	1	1	1	6	2	.2940	.3391
5	2029	7	6	6	3	6	1	7	7	7	7	7	7	.2936	.3351
6	412	7	1	7	4	6	1	7	6	6	6	7	6	.2936	.3324
7	109	7	5	7	6	6	1	6	7	7	7	7	6	.2936	.3351
8	334	7	3	7	3	6	1	6	6	7	3	7	4	.2932	.3328
9	1988	7	6	7	6	6	1	6	6	7	6	7	7	.2932	.3355
10	42	7	3	7	3	5	1	6	6	7	7	7	3	.2932	.3333
11	402	7	6	7	7	7	2	7	7	7	7	7	6	.2932	.3349
12	1528	7	3	7	3	6	1	7	2	5	6	7	3	.2931	.3323
13	647	6	6	7	6	6	1	6	4	7	7	7	7	.2930	.3355
14	376	7	7	7	6	7	2	7	7	7	6	7	7	.2930	.3354
15	1705	6	1	7	3	6	1	5	7	7	7	7	5	.2929	.3333

[a] One factor defined by all items (1–12) and the second (orthogonal) factor by negatively worded items (2,4,6,8,10,12).
[b] Loading of CSES item 6 on the substantive CSE latent factor.
[c] Loading of CSES item 6 on the negative wording method latent factor.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Table 10

Most Influential Cases in the Two Factor Model[a] *(Ranked in Descending Order of LX(8,1)*[b] *and Ascending Order of LX(8,2)*[c]*)*

		CSES Items													
Rank	Case	1	2	3	4	5	6	7	8	9	10	11	12	LX(8,1)	LX(8,2)
1	1287	1	2	2	4	3	4	4	7	1	2	4	1	.5314	.4159
2	486	7	3	7	7	7	7	7	2	7	7	7	7	.5313	.4170
3	514	7	6	7	7	7	6	7	2	7	7	7	7	.5312	.4176
4	1163	5	2	7	6	6	7	7	1	6	7	7	4	.5310	.4159
5	1360	7	5	7	6	6	2	7	2	7	7	7	6	.5307	.4153
6	1538	7	6	7	6	6	2	7	2	7	7	7	7	.5307	.4162
7	648	7	7	6	3	7	4	7	2	6	4	7	2	.5303	.4134
8	432	7	7	7	4	7	5	7	3	7	7	7	7	.5303	.4159
9	634	7	2	7	2	6	2	7	2	7	7	6	6	.5302	.4134
10	1016	1	1	6	7	6	1	1	7	1	1	6	1	.5300	.4170
11	1040	2	1	2	2	2	6	2	6	2	6	1	7	.5300	.4138
12	1528	7	3	7	3	6	1	7	2	5	6	7	3	.5300	.4134
13	1364	1	3	4	2	3	3	2	6	1	2	3	4	.5297	.4154
14	368	7	1	7	7	6	4	7	2	7	6	6	1	.5297	.4143
15	116	2	5	3	7	6	2	3	7	1	5	6	7	.5295	.4140

[a] One factor defined by all items (1–12) and the second (orthogonal) factor by negatively worded items (2,4,6,8,10,12).

[b] Loading of CSES item 8 on the substantive CSE latent factor.

[c] Loading of CSES item 8 on the negative wording method latent factor.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Table 11

Most Influential Cases in the Two Factor Model[a] *(Ranked in Descending Order of LX(10,1)*[b] *and Ascending Order of LX(10,2)*[c]*)*

		CSES Items													
Rank	Case	1	2	3	4	5	6	7	8	9	10	11	12	LX(10,1)	LX(10,2)
1	1046	7	2	7	6	7	4	7	7	7	1	7	4	.5929	.1822
2	277	7	3	7	4	7	6	5	6	7	1	7	7	.5914	.1848
3	1856	2	6	4	4	7	7	2	1	1	7	1	7	.5913	.1826
4	377	7	3	7	2	7	3	7	6	7	2	7	2	.5913	.1808
5	1241	7	6	7	7	7	7	7	7	7	2	7	6	.5913	.1862
6	1040	2	1	2	2	2	6	2	6	2	6	1	7	.5911	.1822
7	32	3	2	4	2	5	2	2	6	3	7	4	2	.5908	.1864
8	850	7	7	7	3	4	1	7	4	7	1	6	4	.5906	.1831
9	227	6	6	7	6	6	6	6	6	7	1	6	7	.5902	.1873
10	1828	7	4	7	6	6	4	6	4	7	2	7	7	.5902	.1849
11	74	7	5	6	3	5	6	6	6	7	1	6	7	.5901	.1861
12	1562	6	2	6	1	4	3	1	3	1	7	4	1	.5901	.1883
13	1782	7	3	7	3	7	6	7	7	5	2	6	3	.5900	.1841
14	211	7	6	6	5	6	1	6	3	6	1	6	5	.5898	.1842
15	1343	6	1	7	1	7	1	5	3	5	1	7	1	.5898	.1795

[a] One factor defined by all items (1–12) and the second (orthogonal) factor by negatively worded items (2,4,6,8,10,12).

[b] Loading of CSES item 10 on the substantive CSE latent factor.

[c] Loading of CSES item 10 on the negative wording method latent factor.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Table 12
Most Influential Cases in the Two Factor Model[a] *(Ranked in Descending Order of LX(12,1)*[b] *and Ascending Order of LX(12,2)*[c]*)*

		CSES Items													
Rank	Case	1	2	3	4	5	6	7	8	9	10	11	12	LX(12,1)	LX(12,2)
1	374	7	1	7	7	7	7	7	7	7	7	7	1	.4344	.6482
2	919	7	3	7	7	7	7	7	7	7	7	7	1	.4341	.6505
3	1040	2	1	2	2	2	6	2	6	2	6	1	7	.4333	.6497
4	1951	6	1	7	7	7	6	7	7	6	4	7	1	.4331	.6486
5	1516	7	7	6	7	6	7	7	7	7	7	7	1	.4329	.6553
6	368	7	1	7	7	6	4	7	2	7	6	6	1	.4327	.6475
7	377	7	3	7	2	7	3	7	6	7	2	7	2	.4326	.6488
8	397	7	4	7	1	5	2	7	5	5	5	7	1	.4321	.6489
9	735	7	2	7	4	5	2	6	4	7	6	7	1	.4320	.6491
10	1856	2	6	4	4	7	7	2	1	1	7	1	7	.4330	.6490
11	1521	7	6	7	6	6	7	7	7	6	7	7	2	.4319	.6522
12	1010	2	1	6	2	5	1	2	4	3	1	3	7	.4317	.6535
13	1999	7	7	7	3	7	7	7	7	7	7	7	3	.4317	.6512
14	204	7	6	7	6	7	2	6	7	7	7	6	2	.4315	.6514
15	648	7	7	6	3	7	4	7	2	6	4	7	2	.4315	.6503

[a] One factor defined by all items (1–12) and the second (orthogonal) factor by negatively worded items (2,4,6,8,10,12).
[b] Loading of CSES item 12 on the substantive CSE latent factor.
[c] Loading of CSES item 12 on the negative wording method latent factor.

While detailed results have been presented for the largest dataset, the same variety of patterns occurred with the other samples. Sample 4 (1,876 Flemish job seekers using a five-point rating scale for the CSES) revealed seven cases with the pattern '5 1 5 1 5 1 5 1 5 1 5 1.' These emerged as the most influential cases in regard to correlation between the positive and negative item factors, and were included among the most influential 15 cases for each item separately. Cases were identified with one inconsistent response (e.g. Case 332: 5 5 5 **1** 5 5 5 5 5 5 5 5), two inconsistent responses (Case 473: 5 5 5 5 5 **1** 5 5 5 **1** 5 5) and so on up to five inconsistent responses (e.g. Case 1556: 5 **1** 5 **1** 5 4 5 **1** 5 **1** 5 **1**). There were also examples of low ratings on positive items appearing with inconsistently high ratings on negative items

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

(e.g. Case 239: 4 5 2 5 4 5 1 5 1 2 1 5). The same variety of patterns was found in Samples 1 and 2.

Discussion

Past simulations of careless responding to negative items have assumed that a careless respondent is careless when answering all negative items. This study evaluated whether this assumption is tenable, using the jackknife method to identify cases with varying levels of inconsistent responding. Ranking jackknife samples according to the correlation between a positive item and negative item factor identified cases with inconsistent responses across many items. Using individual item parameters (with respect to the negative item wording factor) allowed identification of cases with fewer inconsistent responses.

Cases in which respondents give exactly the same answer to all items (e.g. the '5 1 5 1 5 1 5 1 5 1 5 1' patterns noticed in several of the Sample 4 cases) are easy to identify and remove from datasets as being invalid responses. However, it is clear from the influential case analysis in this study that inconsistent responding to negatively-keyed items is not always so easily identified. Given that each table lists less than 1% of cases in the Sample 3 jackknife samples, it is clear that levels of inconsistent responding are highly variable.

Nunnally (1978) notes that inconsistent responding is suggestive of carelessness. Minor inconsistency between item responses is to be expected, especially given the broad nature of the CSE construct and the likely lack of parallelism of items. However, the items have been selected as indicators through a rigorous scale development process, all items are positively correlated with each other, and each loads significantly on a single substantive factor. Cases in which the modal item

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

response is '6' or '7', but one negative item is rated '1' (e.g., Case 197 in Table 7 or Case 1988 in Table 9) show a level of inconsistency which goes beyond what might reasonably be expected in a well-constructed scale.

Given evidence for the existence of carelessness in questionnaire responses (e.g., Nunnally, 1967), the propensity for respondents to answer negatively-worded items in a careless or confused manner (e.g., Barnette, 2000), and the comparatively low test-retest stability of items which include negating linguistic structures (Holden & Fekken, 1990), it seems highly likely that careless responding is an important contributing factor to the more extreme examples of inconsistent response patterns identified in these datasets.

On the basis of the patterns identified, it therefore appears that careless responders may be careless with varying numbers of negatively keyed items. Any simulation of careless responding therefore needs to reflect a much greater diversity of response patterns than past studies, varying the proportion of items answered carelessly, as well as the proportion of careless responders.

Study Two – Monte Carlo Simulation

In this study, careless responding is modelled in a way which reflects the varied levels of carelessness identified in the preceding analysis.

Data sets were created by generating a set of standard normal random variables (items) with specified correlations, converting these to 5 or 7 category ordinal variables with specified levels of skew, before introducing varying levels of simulated careless responding to a subset of these items for a subset of cases. Two techniques were used to identify the number of factors needed to explain the item correlation matrix, both before and after introducing carelessness—Velicer's (1976) minimum

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

average partial correlation (MAP) method and parallel analysis (Hayton et al., 2004; Zwick & Velicer, 1986). Finally, confirmatory factor analysis (CFA) was used to assess the fit of a single-factor model.

The overall simulation was programmed in Python, with data generation carried out in SPSS, and analysis carried out in SPSS, PRELIS and LISREL.

For each level of carelessness, 200 samples of 250 cases were generated. CFA fit indices were averaged over each set of 200 samples; the percentage of samples having one factor (MAP and parallel analysis) and a non-significant chi-square test (CFA) were also recorded.

A detailed description of each of the data generation and analysis steps follows.

Method

Data Sets

The analysis models a twelve item scale, with equal numbers of positively-keyed and negatively-keyed items. The twelve items were all generated in the same way, with no distinction between positively- and negatively-keyed items. This models the situation where recoded negative items are expected to have the same distributional characteristics as positive items. As described below, careless responses were subsequently generated in the data by reverse-coding some of the negative item cases.

The choice of twelve items reflects the need to have an appropriate ratio of indicators to factors, especially in empirical situations where sample sizes may be smaller than desired (Marsh, Kit-Tai, Balla, & Grayson, 1998). The balanced number of positive and negative items is consistent with recommendations aimed at minimising the effect of acquiescent responding (Nunnally, 1978; Spector, 1992). It is also consistent with empirical practice. For example, of the 159 published

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

organisational survey and research scales summarised by the British Telecom Occupational Psychology Unit (1984a, 1984b), 60% included at least one negatively-keyed item. Among this subset of scales, the modal percentage of reversed items was 50%.

Using the approach outlined by Kim (2005) for the calculation of power in structural equation models, a sample size of 250 was decided upon. For a power of .80, using 12 observed variables and an expected RMSEA value of .05, Kim's algorithm suggests a sample of 232 for a one-factor model. The chosen value of 250 is also a realistic sample size for empirical work, and is consistent with common recommendations on minimum samples for structural equation modelling (Kelloway, 1998).

Data sets were created by generating a standard normal random variable for each item in the scale, repeated for the number of cases in the simulated sample. The random numbers were generated using the Mersenne twister algorithm (Matsumoto & Nishimura, 1998) implemented in SPSS. Principal component factor scores were then calculated to ensure independence, and the desired level of correlation between these sets of random numbers was achieved using a matrix decomposition method (Kaiser & Dickman, 1962). Specifically, the random numbers were multiplied by the Cholesky factorisation of the desired correlation matrix, creating multivariate normal score distributions with the desired inter-item correlations. Two different levels of inter-item correlation were used: .3 (to represent a scale covering a relatively broad domain) and .7 (to represent a more tightly focused scale).

The next step involved categorising the variables, and creating realistic levels of skew, consistent with empirically obtained data. Vale and Maurelli (1983) describe a method for creating skewed multivariate data with specified correlations, but their

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

approach is based on generating continuous rather than categorical data. Collapsing continuous variables into a small number of ordered categories significantly reduces the degree of skew and kurtosis. Thus, I created an ordered categorical distribution by dividing the correlated continuous multivariate normal variables into categories using thresholds across the range of the standard normal variables (Bollen & Barb, 1981).

I used skew values for individual positively-worded items in the four different CSES samples (see Chapter 3) to estimate the appropriate range of skew values to model. The samples came from Singapore (2 samples), Belgium and New Zealand, with sample sizes ranging between 178 and 2,088. The mean skew (based on positively worded items) was -.83, and only one item had negative skew exceeding -1.23.

These values are consistent with other values of individual item skew reported in the literature. I obtained individual item skew values for 100 positively and negatively worded items from McPherson and Mohr (2005, Life Orientation Test), Hartshorne (1993, UCLA Loneliness Scale), Dunbar et al. (2000, Rosenberg Self Esteem Scale), and Schriesheim and Kerr (1974, Leadership Behavior Description Questionnaire). The mean skew (absolute values) was .69, with a 90th percentile of 1.15. Hau and Bentler (2004), in their study of non-normal item distributions in structural equation modeling, used comparable values of 0 (normal), 0.5 (slightly non-normal), 1.0 (moderately non-normal) and 1.5 (very non-normal).

In the Monte Carlo studies, target skew levels of 0 (normal), -.83 and -1.2 were used. Desired levels of skew were achieved by selecting appropriate threshold values (Table 13). Applying these thresholds to large samples (N = 500,000) of generated item data reproduced the target levels of skew for items rated on 5- and 7-point response scales.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Table 13
Thresholds used to generate skewed distributions

7-Point Scale:

High Skew (-1.12)	-2.409	-2.054	-1.522	-1.155	-0.305	0.813
Low Skew (-.83)	-2.900	-2.652	1.799	1.433	0.295	0.800
No Skew (0)	-2.1429	-1.2857	-0.4286	0.4286	1.2857	2.1429

5-Point Scale:

High Skew (-1.12)	-2.144	-1.461	-0.745	1.237
Low Skew (-.83)	-2.600	-1.800	-0.940	0.890
No Skew (0)	-1.800	-0.600	0.600	1.800

Once the simulated scale items had been generated, different levels of carelessness were modelled. As demonstrated in Study One above, careless responders do not answer all negatively-keyed items carelessly. Some respondents answer only one item carelessly (and the actual item affected varies across respondents). Similarly, when more than one item is answered carelessly, the actual items vary across cases. Accurate modelling of this process requires consideration of two different values – the percentage of individual respondents who will be careless, and the probability of such respondents answering a particular item in a careless manner. For each careless case, a random uniform variable was used to determine the likelihood of each negatively-keyed item being 'answered' carelessly. Specifically, a uniform random number in the interval [0, 1] was generated for every simulated item rating in the dataset. This value

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

was then compared with the specified probability of answering carelessly to determine if the item would be reverse coded.

The percentage of 'respondents' answering carelessly was varied between 0 and 25% (in steps of .05) and the probability of a reverse-scored item being answered carelessly varied between 0 and 1 (in steps of .25). A value of 0 (for either percentage or probability) corresponds to the baseline case of no careless responding, with combinations of the remaining values providing 20 different levels of carelessness.

Thus, 12 different conditions were used to generate data – 2 levels of response scale (5 and 7 points) X 3 levels of skew (0, -.83, -1.12) X 2 levels of inter-item correlation (.3 and .7). Within each condition, 200 samples of 250 cases each were generated for each of 20 levels of carelessness (plus the no-carelessness condition). This approach provides a much more realistic simulation of carelessness than any previous research, and is consistent with observed patterns in empirical data.

The following analyses were carried out on each sample.

Determining the Number of Factors

As explained on page 64, EFA and CFA methods provide complementary evidence regarding dimensionality, and both approaches were used in the following study.

Velicer's minimum average partial correlation (MAP) (Zwick & Velicer, 1986) method was used to determine the number of factors underlying the item correlation matrix for each sample. This approach is a statistically-based procedure, which sequentially extracts principal components from the correlation matrix until all systematic variance has been accounted for. It is superior to traditional 'rule of thumb'

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

approaches, such as the Kaiser criterion (eigenvalues-greater-than-one) or scree test (Fabrigar et al., 1999; Wood, Tataryn, & Gorsuch, 1996).

Parallel analysis (Hayton et al., 2004; J. L. Horn, 1965) is another statistically-based method for identifying the number of factors to extract. In this approach, random datasets with the same number of variables and observations as the original sample are created (sometimes by permuting the original data). These are used to generate correlation matrices, and eigenvalues are computed. By repeating the process many times, it is possible to identify the probabilities associated with emergence of eigenvalues of given sizes from the random data. If fewer than 5% of randomly generated values for a particular eigenvalue exceed the observed sample value for that eigenvalue, we can be confident that the eigenvalue is unlikely to be mere random noise – it represents a meaningful component.

When used with principal axis factoring, parallel analysis can lead to overextraction, with some negligible factors being identified (Buja & Eyuboglu, 1992). O'Connor (2000) suggests that using both approaches is a good idea, because MAP (if it's wrong) tends to underextract, while PA (if wrong) tends to overextract. I therefore supplemented the use of MAP with parallel analysis based on principal axis factoring.

MAP and parallel analysis were carried out on the datasets using syntax developed by O'Connor (2000), modified to work within the Python programmability extension of SPSS.

CFA Model Fit

Using estimators such as maximum likelihood on ordinal (and skewed) item-level data can result in incorrect parameter estimates, standard errors, and test statistics (T. A. Brown, 2006). Thus, procedures outlined by Jöreskog (2002) for the analysis of

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

ordinal items were followed. Specifically, a polychoric correlation matrix was analyzed, and diagonally weighted least squares (DWLS) estimation was used. This approach has been shown to produce accurate test statistics, parameter estimates and standard errors for ordinal data with sample sizes comparable to that of this study (Flora & Curran, 2004). PRELIS (version 2.72, Jöreskog & Sörbom, 1996) was used to generate the polychoric correlation matrix, and LISREL (version 8.72, Jöreskog & Sörbom, 1999) was used for subsequent analysis. A one-factor CFA model was fitted to both the original data and the data with simulated carelessness. The proportion of samples for which a non-significant Satorra-Bentler chi-square test obtained was recorded for each level of carelessness, and the average values of fit indices (Non-normed Fit Index (NNFI) and standardised root mean residual (SRMR)) across the 200 samples in each condition were calculated. NNFI (also known as the Tucker–Lewis Index) provides a measure of fit compared with the fit of a baseline (null) model, while SRMR is the average discrepancy between the correlations in the sample matrix and the correlations predicted by the model. Simulations by Hu and Bentler (1999) suggest combinations of cutoff values of .95 for NNFI and .06 for SRMR are indicative of good model fit.

Results

The pattern of results for the 5-point and 7-point scales was very similar. Rather than duplicate all the results and figures for both scales, I will present the results for the 7-point scale in detail. Where appropriate, I will comment on any differences in the results for the 5-point scale; tables of MAP results and CFA model fit indices for the 5-point scale simulation are in Appendix B.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Results are presented in both tabular and graphical formats. The tables report accurate values, while the graphs provide an easy way to visualise the trends and patterns across different levels of carelessness, skew and inter-item correlations.

Parallel analysis and MAP gave similar results regarding the emergence of a second factor with increasing carelessness although, as found by Buja and Eyuboglu (1992), parallel analysis was more sensitive. It identified the presence of a second factor at lower levels of carelessness than MAP. For example, with 15% of cases responding carelessly, and a probability of .5 that an item would be answered carelessly, MAP identified only one factor in 55% of samples while PA asserted the existence of two (or more) factors in all 200 samples (7-point scale, moderate skew condition, and inter-item correlation of .3). As the probability of answering carelessly increased to .75, both MAP and PA indicated more than one factor.

PA's sensitivity was also revealed by rare instances when it indicated the existence of three factors in some of the skewed samples. This occurred primarily in conditions where careless respondents had a .5 likelihood of answering negatively-keyed items carelessly. Typically one or two samples out of 200 were indicated as having three factors; the maximum was 11. In 1 sample (out of the 48,000 samples containing some degree of carelessness generated across all conditions), PA indicated the existence of 4 factors.

MAP always indicated either one or two factors – never more. For this reason, a detailed summary of MAP results is presented below. Table 14 summarises the percentage of samples for which Velicer's MAP indicated a single factor underlying the item correlation matrix for the 7-point scale (and Figure 3 presents this information graphically).

Table 14

Percentage[a] *of Samples for which Velicer's MAP Indicated a Single Factor Underlying the Interitem Correlation Matrix (7-Point Response Scale)*

Probability of Carelessly Answering a Negatively-keyed Item	Interitem Correlation = .3						Interitem Correlation = .7						
	Careless Respondents (%)						Careless Respondents (%)						
	0	5	10	15	20	25	0	5	10	15	20	25	
0	100	100	100	100	100	100	100	100	100	100	100	100	
.25	100	100	100	100	100	100	100	100	100	98	99	97	
.50	100	100	98	93	72	**41**	100	70	**14**	**3**	**1**	**0**	Skew = -1.12
.75	100	96	**19**	**2**	**0**	**0**	100	**3**	**0**	**0**	**0**	**0**	
1.00	100	**34**	**0**	**0**	**0**	**0**	100	**0**	**0**	**0**	**0**	**0**	
0	100	100	100	100	100	100	100	100	100	100	100	100	
.25	100	100	100	100	100	100	100	100	97	94	89	86	
.50	100	100	86	55	**26**	**9**	100	53	**5**	**1**	**0**	**0**	Skew = -.83
.75	100	58	**2**	**0**	**0**	**0**	100	**2**	**0**	**0**	**0**	**0**	
1.00	100	**4**	**0**	**0**	**0**	**0**	100	**0**	**0**	**0**	**0**	**0**	
0	100	100	100	100	100	100	100	100	100	100	100	100	
.25	100	100	100	100	100	100	100	100	100	100	100	100	
.50	100	100	100	100	100	100	100	100	96	89	68	**40**	Skew = 0
.75	100	100	100	100	100	100	100	91	**36**	**2**	**0**	**0**	
1.00	100	100	100	99	72	**30**	100	**44**	**2**	**0**	**0**	**0**	

[a]The percentage in each cell is based on 200 samples; values less than 50% shown in bold.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

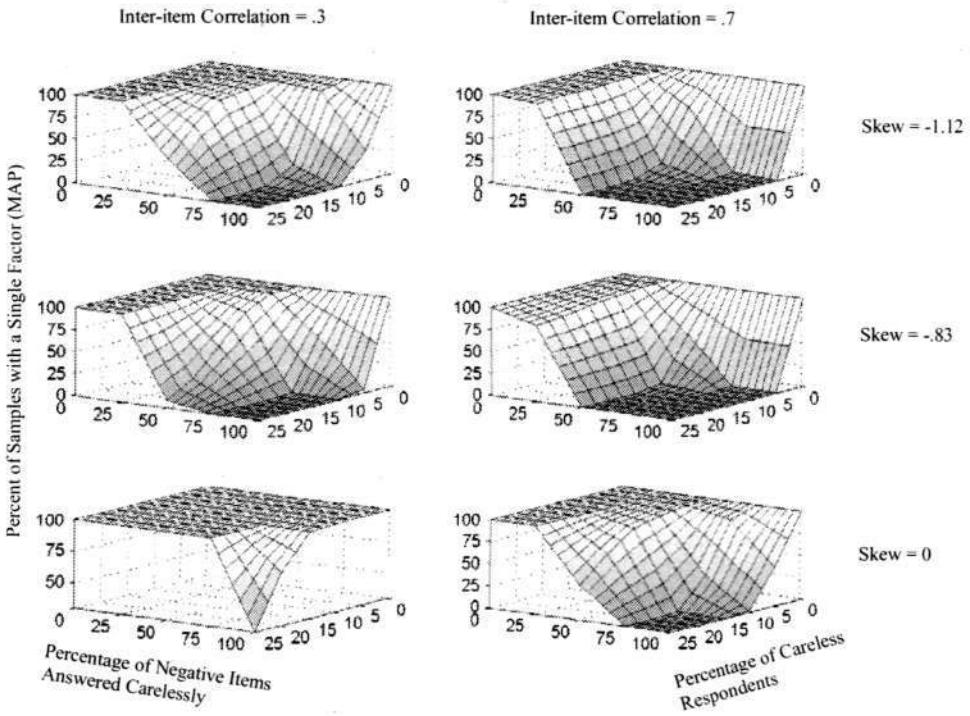


Figure 3. Percentage of samples for which Velicer's MAP indicated a single factor underlying the inter-item correlation matrix (7-point response scale; percentage in each cell based on 200 samples)

A second factor emerged more quickly in the scale with high inter-item correlations (.7). With 5% of respondents having a .75 probability of answering carelessly, only 2% of samples (r = .7, skew = -.83) had a single factor. In comparison, the low inter-item correlation condition (r = .3) still had a majority of unidimensional samples at this level of carelessness.

Results for the 5-point scale were similar, but a second factor was slightly slower to emerge. With 5% of respondents careless, even answering all negative items carelessly still led to MAP indicating only a single factor in 97% of the samples (r =.3, skew = -.83). A second factor emerged in around half of the samples when 15% of respondents had a .75 probability of answering carelessly; at this stage, all samples in

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

the 7-point scale contained two factors. The same trend was evident in the high-skew condition, but the difference between the two scale conditions was less marked.

The no-skew condition showed much less susceptibility to the emergence of a second factor; for $r = .3$, only the highest carelessness condition had a majority of two-factor samples. For $r = .7$, 10% or more of respondents answering with a high probability (.75 or more) was needed to consistently generate a second factor. In the high inter-item correlation condition ($r = .7$), there was no real difference between the 5- and 7-point scales; percentages of samples with one factor were, at most, 3 percentage points higher for the 5-point scale. For $r = .3$, 50% of samples in the highest carelessness condition had one factor (compared with 30% for the 7-point scale).

The Satorra-Bentler chi-square test provides a test of the exact fit of the simulated data to the specified model, with a non-significant ($p < .05$) value indicating model-implied correlations do not differ significantly from sample correlations. Table 15 summarises the percentage of samples within each carelessness condition for which a non-significant chi-square obtained (for the 7-point scale), and Figure 4 displays this graphically. Differences between the two skew conditions are relatively minor. As with the MAP analysis, even quite low levels of carelessness (5% of respondents having a probability of .75 of answering negatively-keyed items carelessly) can lead to rejection of model fit for the majority of samples in the $r = .7$ scale. Although less marked in the low inter-item correlation condition, as few as 10% of respondents carelessly answering an average of 75% negatively-keyed items resulted in a significant chi-square in almost all samples.

As found in the MAP analysis, the 5-point scale simulation showed slightly more resilience to the effect of carelessness. In general, between the extremes of 0%

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

and 100%, the 5-point scale showed slightly higher percentages of samples with non-significant chi-square. However, in almost all conditions, there was agreement across both scale types as to whether the majority of samples fitted the one-factor model.

SRMR (Table 16) is the standardised difference between observed and predicted correlations, with values below .06 indicating good fit. The results show a very close agreement with the factor structure suggested by MAP. Of the 60 cells in Table 14 where MAP indicated low likelihood of a one factor solution (i.e., those cells with fewer than 50% of samples matching a one factor structure), 51 had SRMR values greater than .06. The exceptions occurred primarily in the high skew condition at intermediate levels of carelessness (e.g., 5% or 10% of respondents being careless with half to three-quarters of the items).

NNFI (Table 17, *Figure 5*) measures the extent to which the hypothesised model improves fit, over the null or independence model. Values greater than .95 indicate good fit. NNFI showed greater tolerance of the effects of careless responding than SRMR. SRMR indicated poor fit in 51 cells (i.e., SRMR > .06); only 35 of these cells had NNFI values less than .95.

NNFI and SRMR gave almost identical indications of fit across the 5-point and 7-point scales. The differences tended to occur in the third decimal place – for example, SRMR of .056 in the 7-point scale and a value of .060 in the 5-point scale. Rounding the values to 2 decimal places resulted in near-complete agreement across the two scale types.

Table 15

Percentage[a] *of Non-Significant Satorra-Bentler Chi-square Values for a One-Factor Model Fitted to Samples with Varying Levels of Careless Responding (7-Point Response Scale)*

Probability of Carelessly Answering a Negatively-keyed Item	Interitem Correlation = .3						Interitem Correlation = .7						
	Careless Respondents (%)						Careless Respondents (%)						
	0	5	10	15	20	25	0	5	10	15	20	25	
0	100	100	100	100	100	100	100	100	100	100	100	100	
.25	100	100	100	100	100	100	100	100	99	95	87	77	
.50	100	100	96	71	**35**	**12**	100	74	**12**	**2**	**0**	**0**	Skew = -1.12
.75	100	93	**14**	**0**	**0**	**0**	100	**4**	**0**	**0**	**0**	**0**	
1.00	100	**31**	**0**	**0**	**0**	**0**	100	**0**	**0**	**0**	**0**	**0**	
0	100	100	100	100	100	100	100	100	100	100	100	100	
.25	100	100	100	100	100	100	100	100	96	91	82	69	
.50	100	100	**86**	**46**	**12**	**2**	100	95	**8**	**1**	**0**	**0**	Skew = -.83
.75	100	75	**4**	**0**	**0**	**0**	100	**2**	**0**	**0**	**0**	**0**	
1.00	100	**11**	**0**	**0**	**0**	**0**	100	**0**	**0**	**0**	**0**	**0**	
0	100	100	100	100	100	100	100	100	100	100	100	100	
.25	100	100	100	100	100	100	100	100	100	100	99	98	
.50	100	100	100	100	100	100	100	98	81	**46**	**22**	**3**	Skew = 0
.75	100	100	100	98	76	**44**	100	68	**8**	**1**	**0**	**0**	
1.00	100	100	90	**35**	**7**	**0**	100	**17**	**0**	**0**	**0**	**0**	

[a]The percentage in each cell is based on 200 samples; values less than 50% shown in bold.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

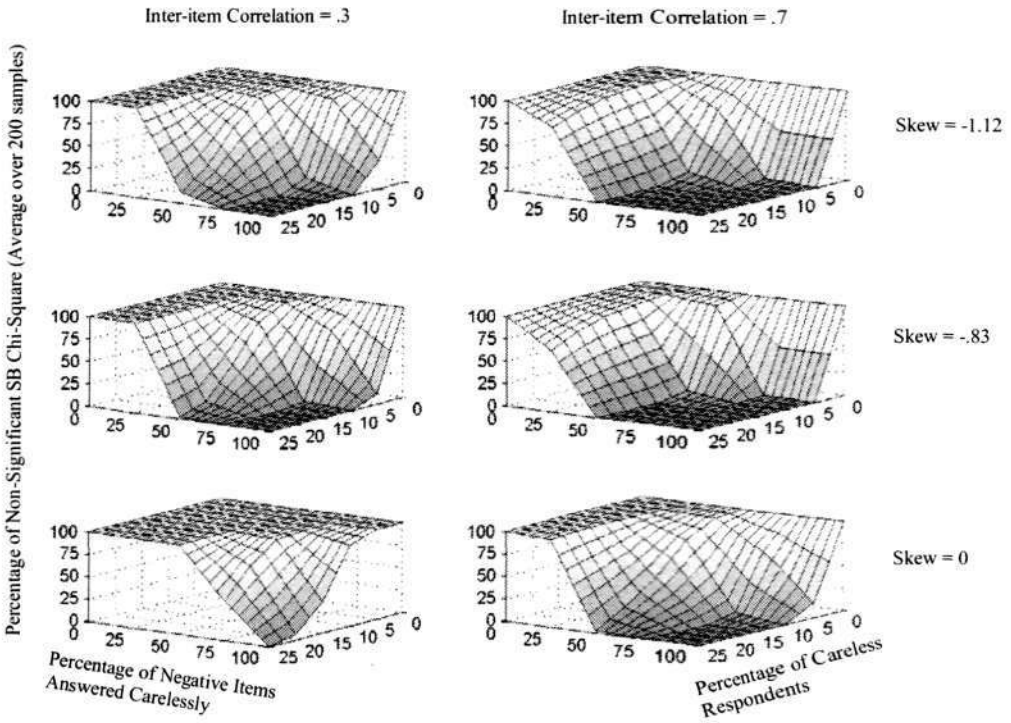


Figure 4. Percentage of samples with non-significant Satorra-Bentler chi-square (7-point response scale; percentage in each cell based on 200 samples)

Table 16

SRMR Values Averaged Over 200 Samples for One-Factor Model Fitted to Samples with Varying Levels of Careless Responding (7-Point Response Scale)[a]

Probability of Carelessly Answering a Negatively-keyed Item)	Interitem Correlation = .3						Interitem Correlation = .7						
	Careless Respondents (%)						Careless Respondents (%)						
	0	5	10	15	20	25	0	5	10	15	20	25	
0	.021	.021	.021	.021	.021	.021	.013	.013	.013	.013	.013	.013	
.25	.021	.025	.028	.032	.034	.036	.013	.019	.023	.026	.030	.032	
.50	.021	.031	.041	.049	.056	**.064**	.013	.029	.044	.056	**.066**	**.076**	Skew = -1.12
.75	.021	.040	**.061**	**.077**	**.091**	**.102**	.013	.046	**.077**	**.104**	**.129**	**.152**	
1.00	.021	.054	**.083**	**.104**	**.118**	**.130**	.013	**.075**	**.131**	**.184**	**.233**	**.275**	
0	.023	.023	.023	.023	.023	.023	.014	.014	.014	.015	.015	.015	
.25	.023	.028	.031	.034	.037	.039	.015	.023	.028	.031	.034	.037	
.50	.023	.036	.046	.055	**.063**	**.069**	.014	.035	.050	**.063**	**.074**	**.082**	Skew = -.83
.75	.023	.047	**.067**	**.083**	**.097**	**.110**	.015	.057	**.090**	**.117**	**.139**	**.161**	
1.00	.023	**.062**	**.092**	**.112**	**.123**	**.133**	.014	**.091**	**.152**	**.204**	**.248**	**.288**	
0	.016	.016	.016	.016	.016	.016	.010	.010	.010	.010	.010	.010	
.25	.016	.018	.021	.023	.025	.027	.010	.015	.018	.020	.023	.024	
.50	.016	.021	.025	.029	.033	.036	.010	.021	.029	.036	.043	.050	Skew = 0
.75	.016	.023	.031	.039	.046	.054	.010	.029	.049	**.069**	**.086**	**.107**	
1.00	.016	.026	.039	.053	**.068**	**.084**	.010	.041	**.079**	**.128**	**.167**	**.215**	

[a] Values > .06 shown in bold

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Table 17

NNFI Values for One-Factor Model Fitted to Samples with Varying Levels of Careless Responding[a]

Probability of Carelessly Answering a Negatively-keyed Item)	Interitem Correlation = .3						Interitem Correlation = .7						
	Careless Respondents (%)						Careless Respondents (%)						
	0	5	10	15	20	25	0	5	10	15	20	25	
.0	1.026	1.027	1.026	1.027	1.027	1.027	1.004	1.004	1.004	1.004	1.004	1.004	
.25	1.026	1.025	1.023	1.020	1.019	1.017	1.004	1.003	1.002	1.001	1.000	0.999	Skew -1.12
.50	1.026	1.019	1.005	0.991	0.974	0.951	1.004	0.999	0.991	0.983	0.973	0.962	
.75	1.026	1.004	0.963	**0.923**	**0.880**	**0.840**	1.004	0.989	0.964	**0.929**	**0.887**	**0.837**	
1.00	1.027	0.978	**0.917**	**0.876**	**0.859**	**0.848**	1.004	0.964	**0.896**	**0.807**	**0.705**	**0.600**	
0	1.025	1.025	1.025	1.025	1.025	1.025	1.003	1.003	1.003	1.003	1.003	1.003	
.25	1.025	1.022	1.020	1.018	1.015	1.013	1.003	1.002	1.001	1.000	0.999	0.998	
.50	1.025	1.014	0.999	0.981	0.959	**0.940**	1.003	0.998	0.989	0.979	0.970	0.960	Skew -.83
.75	1.025	0.994	0.954	**0.914**	**0.874**	**0.832**	1.003	0.985	0.955	**0.916**	**0.875**	**0.824**	
1.00	1.025	0.965	**0.908**	**0.880**	**0.872**	**0.868**	1.003	0.955	**0.876**	**0.778**	**0.683**	**0.577**	
0	1.030	1.030	1.030	1.030	1.030	1.030	1.005	1.005	1.005	1.005	1.005	1.005	
.25	1.030	1.030	1.029	1.028	1.027	1.026	1.005	1.004	1.003	1.002	1.002	1.002	
.50	1.030	1.028	1.026	1.023	1.020	1.016	1.005	1.002	0.999	0.996	0.992	0.987	Skew 0
.75	1.030	1.027	1.019	1.008	0.992	0.970	1.005	0.998	0.988	0.972	0.956	**0.930**	
1.00	1.030	1.023	1.004	0.972	**0.922**	**0.845**	1.005	0.990	0.960	**0.911**	**0.857**	**0.778**	

[a] Values < .95 shown in bold

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

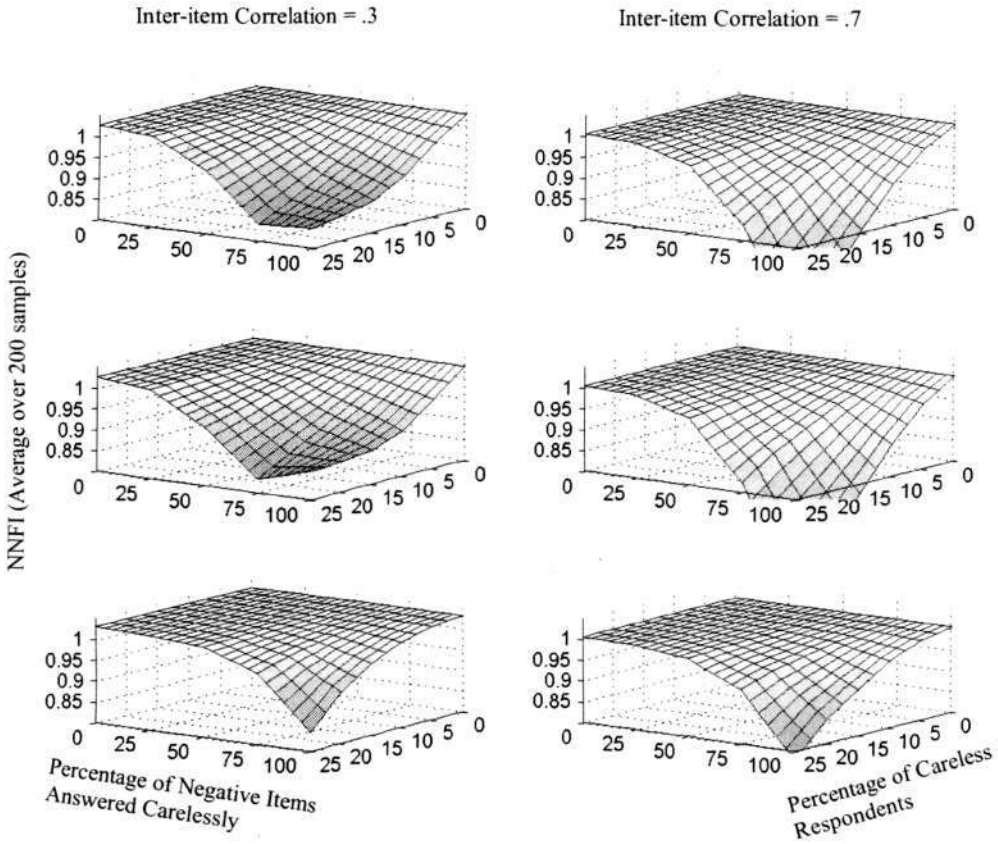


Figure 5. NNFI values for one-factor model averaged over 200 samples, for different levels of careless responding.

Discussion

In this chapter, I have demonstrated that careless responding cannot be realistically modelled by assuming that all careless respondents are careless with every negatively worded item. The variety of patterns of inconsistent responses to positive and negative items argues for a more sophisticated model of carelessness.

The results of a Monte Carlo simulation which realistically models carelessness illustrate the ease with which carelessness can create a spurious factor. The different approaches (parallel analysis, MAP, chi-square, SRMR, and NNFI) show varying

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

sensitivity to the effects of carelessness, but agree at the extremes of both low and high levels of carelessness.

The chi-square values were most sensitive. For example, with an average inter-item correlation of .7, 5% of respondents answering three-quarters of the negative items carelessly resulted in almost all samples failing to achieve a non-significant chi-square for the one-factor model (at both levels of skew). With SRMR, 5% of respondents had to answer all items carelessly before the one-factor model reached unacceptable fit levels, and for NNFI, the one-factor model fitted acceptably even when all items were answered carelessly by 5% of respondents.

The item wording factor was slightly more problematic in the simulation of a scale measuring a narrow construct (i.e., average inter-item correlation of .7). At all levels of skew, MAP and chi-square were consistent in identifying problems with the one-factor model at lower levels of carelessness, compared with the correlation = .3 simulation. While SRMR showed slightly more sensitivity to carelessness in the high inter-item correlation condition, there was virtually no difference for NNFI.

While the five-point scale was less sensitive to the emergence of a wording factor (based on MAP analysis), there was virtually no difference between the seven-point and five-point scales in regard to assessment of one-factor model fit (SRMR, NNFI).

There were very few differences in indications of factor structure between model fit indices for the moderate and high skew conditions.

An important finding from the simulation is that the wording factor also emerged in samples in the no-skew condition. If responses to an item were normally distributed, then the effects of random recoding of a subset of cases could be expected to 'cancel out'. With a symmetric pattern of responses around the scale midpoint, there

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

is an equal chance of items above and below the midpoint being reversed—For every '2' converted to '6' on a 7-point scale, we could expect a '6' to be converted back to a '2.' The net result (given a sufficiently large number of cases) would be no change; the recoded subset would have the same distribution as the non-recoded cases and a wording factor would not emerge. However, this simulation relied on realistically sized samples. While the expected value of skew might have been zero, the actual value varied from sample to sample.

I calculated skew for 24,000 items in the 'no skew' condition (12 items in each of 10 simulation sets, each comprising 200 samples). The mean value was 0 (.00059), with a standard deviation of .13. Values for individual items ranged between a minimum of -.48 and a maximum of .50. At high levels of carelessness, these relatively low levels of skew were clearly enough to create an item wording factor.

The results of the Monte Carlo simulation are consistent with Schmitt and Stults's findings (1985), while extending them considerably. Two of the samples simulated by Schmitt and Stults had average item intercorrelations around .3, and their scales had seven rating points. They concluded that 10% of respondents answering carelessly were sufficient to create a wording factor. In my Monte Carlo simulation, 10% of respondents answering all negative items carelessly was also enough to reduce fit of a single factor model to unacceptable levels (using the Hu and Bentler 1999 criteria of NNFI > .95 and SRMR < .06). This was true of the moderate and high skew conditions for both high and low item intercorrelations. However, it was only true when 10% of respondents answered every negative item carelessly; when fewer items were answered carelessly, the one factor model continued to have adequate fit.

In contrast, use of MAP and the chi-square statistic indicate the emergence of a wording factor at lower levels of carelessness than suggested by Schmitt and Stults. A

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

wording factor was identified by these techniques when as few as 5% of respondents answered all (or most) negatively worded items carelessly.

It is difficult to make a direct comparison with Woods's (2006) findings. She used dichotomous items, and cites evidence that an SRMR value of .08 indicates acceptable fit for categorical outcomes. In her 'no carelessness' condition, SRMR = .08 (with a sample size of 250). This fit deteriorated slightly as carelessness increased to 5% (SRMR = .09, NNFI = .98), and becomes unacceptable at 10% carelessness (SRMR = .12, NNFI = .95). This broad pattern is consistent with the findings of the Monte Carlo simulation described in this chapter.

This simulation has provided a detailed picture of how varying levels of carelessness affect factor structure and CFA model fit. Such information is only valuable, however, if it allows us to improve the quality of empirical data. I address this issue in the next chapter.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

CHAPTER FIVE

IDENTIFYING CARELESS CASES IN EMPIRICAL DATA

In the previous chapter, the jackknife procedure suggested by Rensvold and Cheung (1999) was used to identify cases which exhibited anomalous response patterns in empirical data. Cases identified in this manner displayed inconsistent responses to positively and negatively keyed items. As Nunnally (1978) pointed out, this can be taken as circumstantial evidence of possible careless responding. However, in the absence of any other information on how people responded to the items, it is not possible to conclusively demonstrate the operation of carelessness.

In contrast, 'careless' cases in the Monte Carlo data samples described in the previous chapter can be identified with certainty. By applying the jackknife procedure to the Monte Carlo data, it is possible to assess the extent to which known careless cases are able to be distinguished from non-careless cases. Comparing the distribution of parameter values in the jackknife samples with the distribution in empirical samples will further clarify the extent to which careless responding contributes to the existence of an item wording factor.

If item wording factors result from a small proportion of respondents answering carelessly, then the jackknife procedure may provide a means by which these cases can be identified. By clarifying the extent to which carelessness affects model fit and parameters in the Monte Carlo data, it becomes possible to apply these findings to empirical data. An appropriate cutoff for identifying outlying cases can be established, based on a comparison between simulated and empirical data. Outliers can then be identified, and consideration given to removing them from the sample in order to improve model fit.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

In this chapter, I begin by carrying out a jackknife analysis of datasets generated using the Monte Carlo procedure described in Chapter 4. I then compare the resulting rankings of outlying cases in simulated data with the distribution of cases in Sample 3 to identify the possible extent of carelessness in empirical data. Finally, by removing outlying cases suggested by the jackknife procedure, I investigate the extent to which the item wording factor can be eliminated.

Identifying Careless Cases in the Monte Carlo Data

Instead of using an overall fit index, the correlation between two latent factors (defined by positively and negatively keyed items respectively) will be used. If positively keyed and negatively keyed items are equivalent indicators of a single latent variable, then the correlation between the two factors will approach 1. Increasing levels of careless responding to negative items will reduce this value.

If the existence of an item wording factor in empirical data is solely due to careless responding, then deterioration in the correlation between the positive and negative item factors will provide an indication of the amount of careless responding. Comparison of the correlation coefficient in empirical samples with the various values obtained in the different Monte Carlo samples will provide an indication of the likely amount of careless responding.

The process described above requires estimation of the values of phi at each level of carelessness; details of the distribution of jackknife sample phi values in an empirical dataset; and details of the distribution of jackknife sample phi values for careless and non-careless cases in simulated data (designed to match characteristics of the empirical sample).

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Method

In order to estimate the value of phi at each level of carelessness, 100 samples (each with 500 cases) were generated for each level of carelessness. Interitem correlation was set at .3, and skew at -.8; as described below, these values are consistent with the values in the empirical (Sample 3) dataset. As in Chapter 4, carelessness was varied in terms of both the percentage of careless cases (ranging from 5% to 25% in steps of 5%), and the likelihood of each careless case answering a negatively keyed item carelessly (between .25 and 1 in steps of .25). This resulted in 20 separate levels of carelessness – 5 (careless cases) X 4 (item probabilities). PRELIS was used to generate polychoric correlation and asymptotic covariance matrices. The two factor model (Model 2 in *Figure 1*, p. 67) was fitted in LISREL, using diagonally weighted least squares, and the fit indices and parameter values (including the correlation between factors, phi) for each jackknife sample were collated in SPSS for subsequent analysis.

For each individual sample, the mean value of phi across the 500 jackknife samples was calculated. This mean value was then averaged across the 100 samples in each carelessness condition, to provide an estimate of phi.

In order to identify the extent to which careless cases could be distinguished on the basis of their effect on phi, one sample (of 2000 cases) was generated for each of the 20 levels of carelessness. The jackknife analysis was then carried out on each of these 20 samples in turn, and a histogram plotted to show the distribution of phi values across the 2000 jackknife samples (distinguishing careless and non-careless cases).

To evaluate the efficacy of using the outlying case method to identify carelessness in empirical data, the samples described in Chapter 3 (p. 63) will be used. Detailed results will be provided for Sample 3, as this is the largest sample (N =

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

2,067), and standard techniques for identifying careless responses become more difficult with large samples (Rensvold & Cheung, 1999, p. 305). Summary results will be provided for the other samples.

The average interitem correlation for the twelve CSES items in Sample 3 is .29, and the average skew of the items is -.71. The Monte Carlo simulation with interitem correlation = .3 and item skew -.8 is closest to these values, and was used to generate the simulated data.

Finally, a histogram was generated for Sample 3 data, showing the distribution of phi values across the 2,067 jackknife samples.

Simulated data samples were generated using the same program described in Chapter 4 (p. 109).

Results

The mean correlation between the positive item factor and negative item factor for each level of carelessness is shown in *Figure 6*. The mean ranges from .97 (sd = .02) for the lowest level of carelessness (5% of cases, probability of .25 that a negatively keyed item is answered carelessly) down to .27 (sd = .05) for the highest level of carelessness (25% of cases, all negatively keyed items answered carelessly).

The histograms of phi values for samples at each level of carelessness are included in Appendix C, with two examples (low carelessness (5% careless cases, .25 probability of careless answering) and high carelessness (25% careless cases, all negatively keyed items answered carelessly)) shown in *Figure 7*. The histograms illustrate clearly that careless cases are the most influential in reducing phi. The distributions all demonstrate a degree of positive skew, and the outlying cases on the right of the distribution are all careless cases. These are the cases which, when

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

removed in the jackknife procedure, result in higher values of phi (i.e., result in greater agreement between the positively worded item factor and the negatively worded item factor).

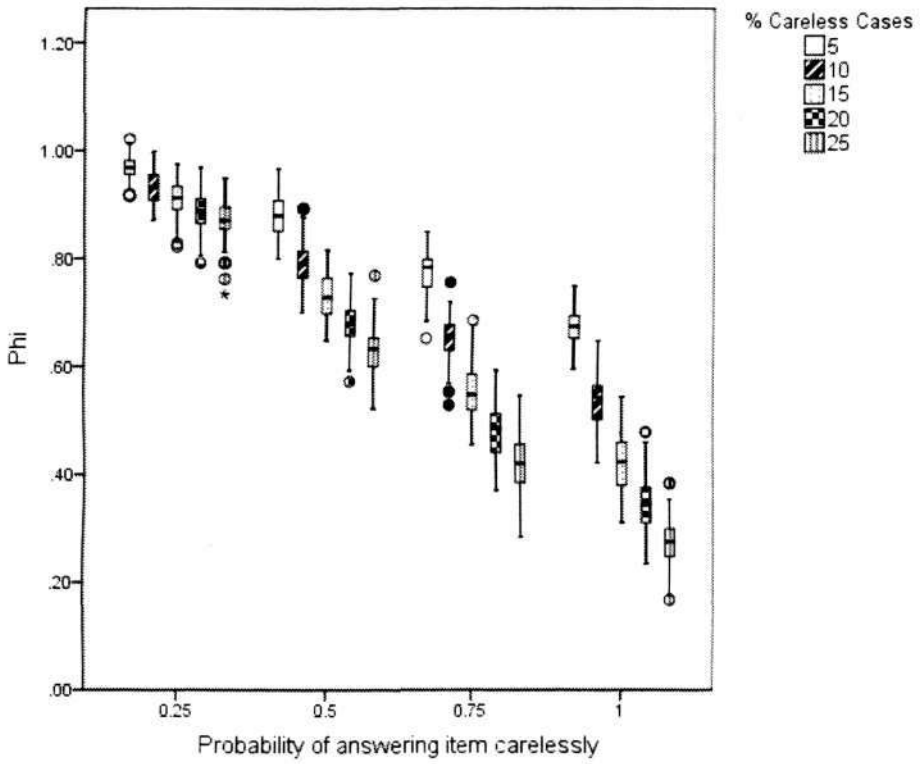
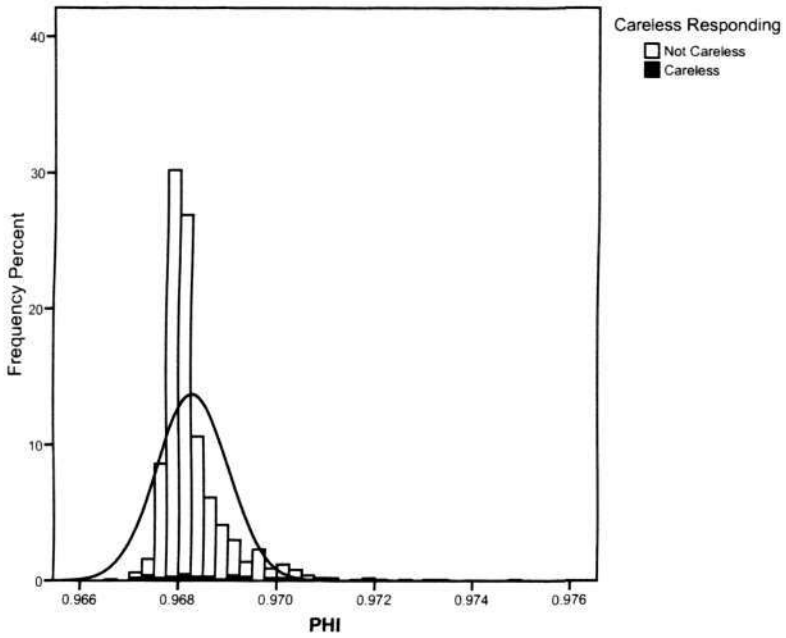
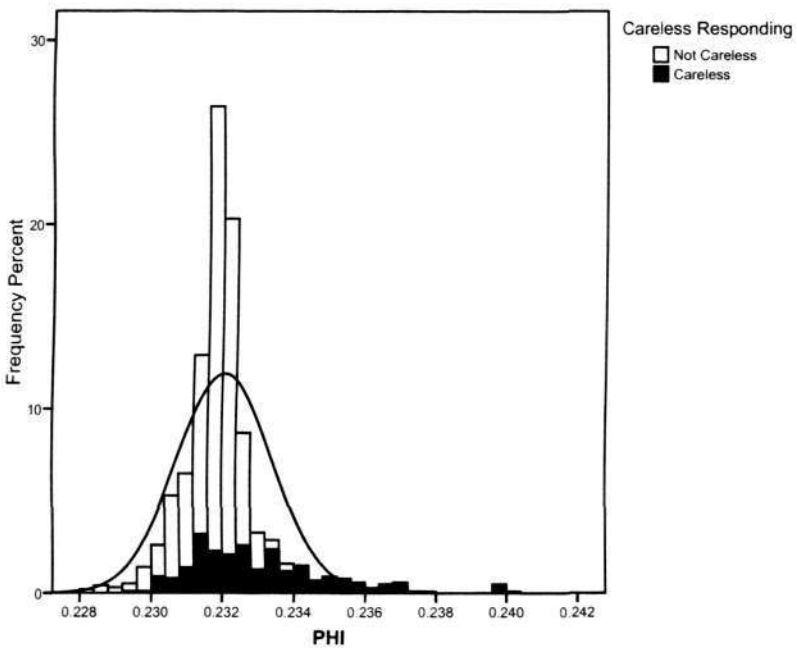


Figure 6. Boxplots of mean correlation (phi) between positive item and negative item scales for each level of carelessness in the simulated data.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library



5% Careless respondents; .25 likelihood of answering negative items carelessly



25% Careless respondents; all negative items answered carelessly

Figure 7. Histograms of mean correlation (phi) between positive item and negative item scales for jackknife samples at two level of carelessness in the simulated data.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

While this is true in all samples, it is seen most clearly in the high carelessness sample (bottom histogram in *Figure 7*). While many of the careless responses fall within the main weight of the distribution, the right tail is comprised exclusively of careless cases.

Figure 8 shows the histogram of phi values for the 2,067 jackknife samples in the Sample 3 data. The distribution is very similar to that for the simulated data; the mean phi value is .68.

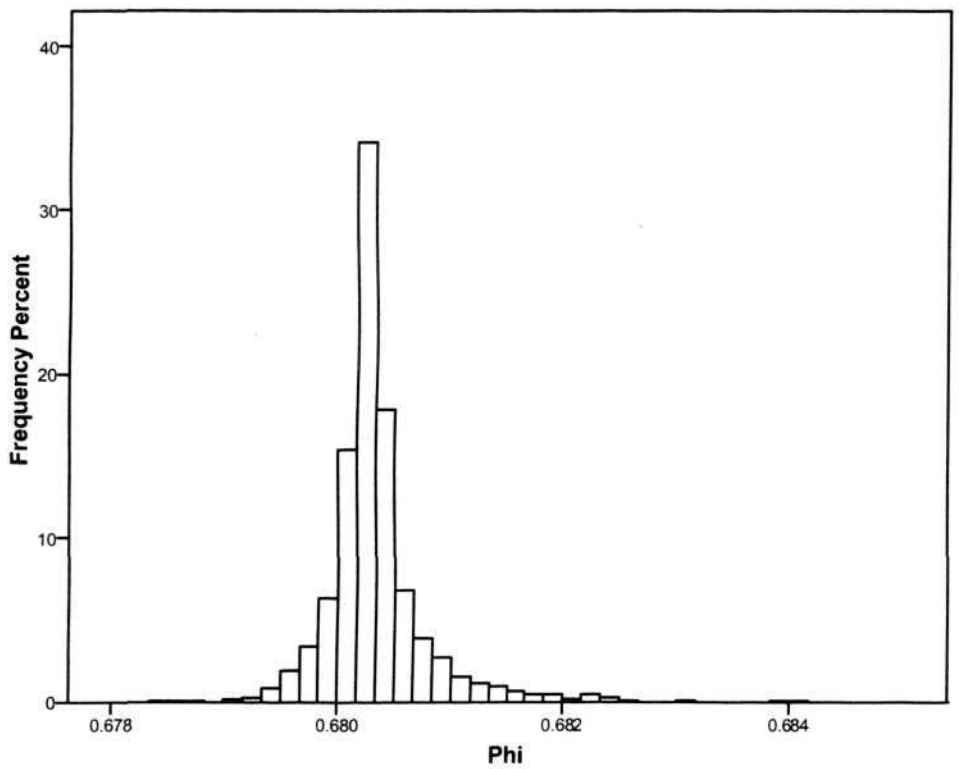


Figure 8. Histogram of mean correlation (phi) between positive item and negative item scales for jackknife samples from Sample 3 data.

Using Simulated Data as a Guide to 'Cleaning' Empirical Data

If the poor fit of Sample 3 data to the one factor model is due solely to carelessness, then the obtained value of phi gives an indication of the likely amount of such carelessness. Based on *Figure 6*, a phi value of .68 falls in the interquartile range

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

of two Monte Carlo distributions – 20% careless cases with a .5 probability of answering negatively keyed items carelessly, and 5% careless cases answering all negatively keyed items carelessly. However, the distribution of phi values in Sample 3 (*Figure 8*) most closely matches the distribution for the 20% careless cases simulation (see Appendix B, p. 207) so this level of carelessness will be used in the following analysis.

By rank ordering cases in this simulated dataset according to descending values of the correlation between positive and negative item scales (phi), it is possible to see how the relative frequency of careless and non-careless cases varies with phi. This is shown in *Figure 9* by means of a stacked histogram. The first column represents the 100 cases which, when left out of the analysis in the jackknife procedure, resulted in the highest values of phi. Of these, 86 are careless cases, and only 14 are non-careless. In fact, 75% of the first 150 cases, and 64% of the first 200 cases were careless (compared with an overall carelessness percentage in the entire sample of only 20%). The ability of phi to accurately classify careless cases was also confirmed using the receiver operating characteristic plot (ROC). With an area under the ROC curve of .5 representing chance classification, phi produced a curve with an area of .67 (95% confidence interval .63-.70, $p < .001$).

Recall that in the simulation, each careless 'respondent' in the simulation had only a .5 likelihood of answering a negative item carelessly. Furthermore, some of these simulated careless item ratings will be identical with original ratings – any rating at the midpoint of the scale will remain unchanged when the item is reverse-coded. Ratings close to the midpoint of the scale will show minor inconsistency – the level of inconsistency that might also be consistent with lack of parallelism among items. However, ratings near the endpoints of item rating scales will show marked

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

inconsistency when reversed (e.g. from '7' to '1'); these are the items that have most impact on definition of the method factor, and which are identified at the left extreme of the distribution in *Figure 8* and *Figure 9*.

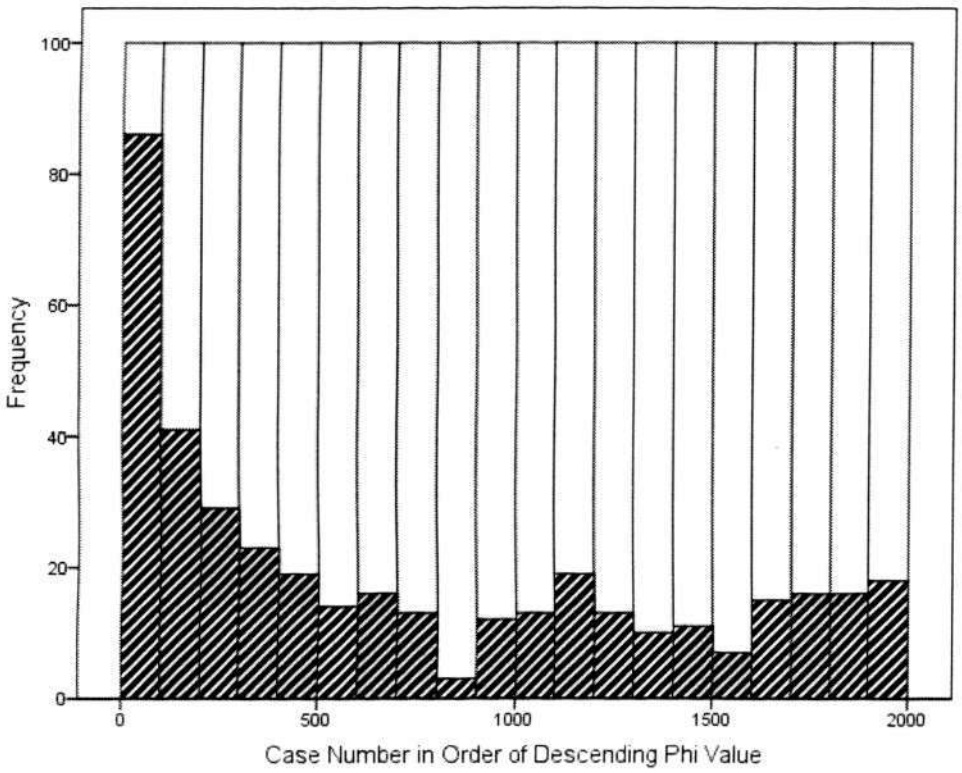


Figure 9. Histogram showing frequency of careless (shaded) and non-careless cases; cases ordered according to decreasing correlation between positive and negative item scales.

If this pattern applies to the empirical data, then it should be possible to remove cases associated with the highest 5% of phi values without discarding many non-careless cases. Furthermore, removal of these cases could be expected to result in improved fit of a one-factor model to the remaining data.

Fit indices are not the only means for assessing the effect of removing outlying cases. As noted in Chapter 3 (p. 74), fitting a one-factor model to Sample 3 resulted in a pattern of residuals indicative of different responses to positively and negatively

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

keyed items. The standardised residuals of the model-implied correlation matrix consistently overestimated the correlations between a positively and a negatively worded item (mean standardised residual of -3.03) and underestimated the correlation between items keyed in the same direction (mean standardised residual of 3.31). If the removal of outlying cases is effective in reducing careless responding, then it should be reflected in convergence of the standardised residuals towards zero.

Finally, removing cases with inconsistent responses to negatively keyed items should result in an increased correlation between the positive item and negative item factors.

Thus, the impact of removing outlying cases on the empirical data can be evaluated by the increase in model fit, the analysis of residuals, and the correlation between positive-item and negative-item scales.

Method

Cases in the Sample 3 data were rank ordered according to descending magnitude of phi when each case was left out of the sample. Different percentages of extreme cases (top 5%, 7.5%, and 10% of phi values) were removed from the data in turn. For each percentage, the number of factors in the remaining data was estimated by use of Velicer's MAP and parallel analysis. Parallel analysis was conducted using raw data permutation, extracting principal components, on 1,000 permuted datasets. The 95th percentile of eigenvalues in the permuted datasets were compared with corresponding sample eigenvalues.

The trimmed samples were also fitted to a one-factor model, and fit indices calculated. Standardised residuals were obtained, and averaged separately for same-

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

direction items (i.e., positive-positive and negative-negative pairs) and for oppositely keyed items (positive-negative pairs).

Finally, a two-factor CFA model (with factors defined by positive and negative items respectively) was fitted to the data, and the correlation (phi) between the two factors calculated.

Results

As noted on p. 74, both parallel analysis and MAP indicated the existence of two factors in the full Sample 3 dataset. After removing the extreme 5% of outlying cases (based on phi), MAP indicated only one factor was necessary to account for the correlation matrix. MAP continued to indicate only one factor in the samples with 7.5% and 10% of outliers removed. In order to ascertain how few cases needed to be removed in order for MAP to indicate only one factor, I iteratively reduced the number of cases removed. Removing as few as 8 cases from the 2,067 cases in Sample 3 (0.4%) was enough for MAP to indicate only one factor underlying the correlation matrix; removing 7 or fewer resulted in two factors being identified.

Parallel analysis indicated the existence of two factors in both the 5% and the 7.5% trimmed samples, but only one factor in the 10% trimmed sample. In the 7.5% trimmed sample, the second eigenvalue in the sample was 1.14, compared with a 95th percentile value of 1.12 in the randomly permuted data. At this level of trimming, the second factor is therefore accounting for less than 10% of variance (compared with 37% by the first factor (eigenvalue = 4.48)).

Fit indices at successive levels of trimming are shown in Table 18. Removal of only 5% of outlying cases reduced the model chi-square by over 40%, and moved fit indices close to acceptable levels. With removal of 7.5% of outliers, NNFI and SRMR

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

attained levels consistent with recommendations for acceptable model fit (Hu & Bentler, 1999), and CFI also indicated good fit, while RMSEA remained somewhat high. There was little change in fit indices as the number of outliers removed increased to 10%.

Given that MAP indicated only one factor with 0.4% of cases removed, I estimated a two factor model (positive items, negative items) on this reduced dataset. The correlation between factors increased from .68 (full sample) to .70 – the 'rule of thumb' level which Nunnally and Bernstein (1994) suggested could be used to indicate that the two sets of items are measuring the same construct.

Table 18
Fit Indices for One-Factor Model Fitted to Sample 3 with Varying Percentages of Outlying Cases Trimmed from Sample

Cases Trimmed	*SB Chi-Square*	*df*	*NNFI*	*RMSEA (90% CI)*	*ECVI (90% CI)*	*CFI*	*GFI*	*SRMR*
0%	1702.03**	54	.88	.12 (.12–.13)	.85 (.78–.91)	.90	.96	.09
5%	959.25**	54	.94	.09 (.09–.10)	.51 (.46–.57)	.95	.98	.06
7.5%	800.00**	54	.95	.08 (.08–.09)	.44 (.40–.49)	.96	.98	.06
10%	709.71**	54	.95	.08 (.08–.09)	.41 (.36–.46)	.96	.99	.06

* $p < .01$

The pattern of average standardised residuals also showed dramatic improvement with the removal of 5% of outlying cases. This is shown graphically in *Figure 10*. The mean standardised residual between pairs of items keyed in the same direction reduced from 3.31 (full sample) to 1.65 (5% outliers removed), and to .82 (10% outliers removed). Similarly, the mean standardised residual for pairs of items keyed in opposite directions increased from -3.03 (full sample) to -1.51 (5% outliers removed), reaching -.8 when 10% of outliers were removed.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Finally, the correlation between positive and negative item factors increased markedly. From .68 (full sample), it increased to .84 (5% trimmed), .88 (7.5% trimmed) and .92 (10% trimmed).

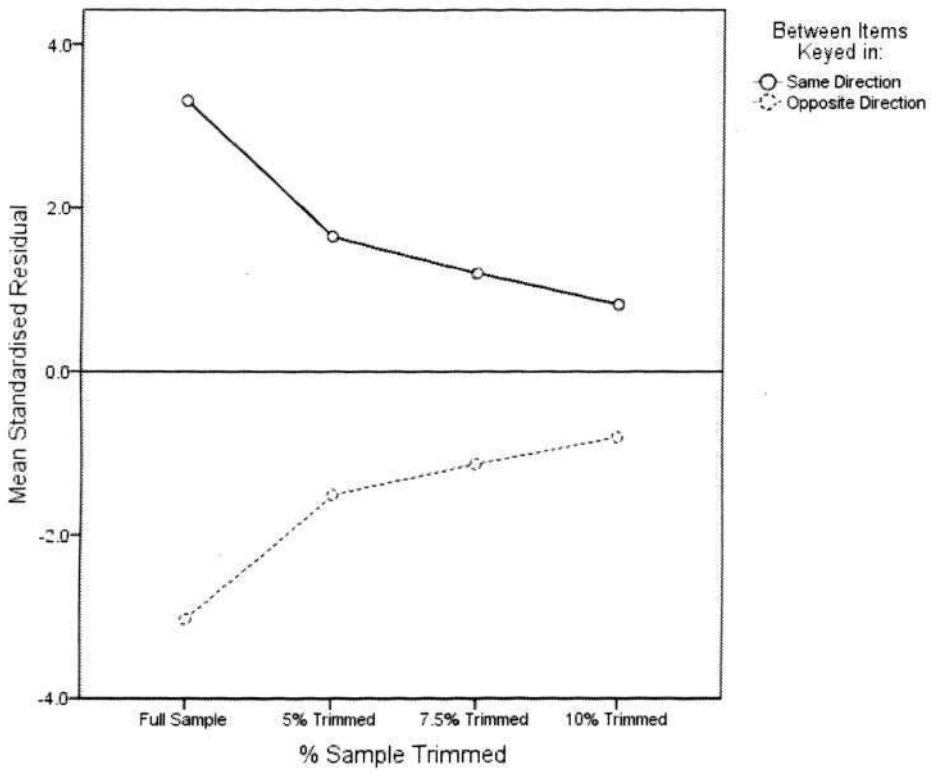


Figure 10. Plot of mean standardised residuals for same-keyed and differently-keyed item pairs (Sample 3 data, with indicated percentages of outliers removed).

While the detailed results presented here focus on Sample 3, I also evaluated the effect of removing varying proportions of outlying cases on the other three samples described in Chapter 3 (see p. 63 for details of sample characteristics). In all cases, marked improvements in fit occurred with deletions of small numbers of outlying (extreme phi value) cases.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

In the other large dataset (Sample 4, N = 1,876) the Satorra Bentler chi square more than halved, from 1724.8 (full sample) to 769.4 with removal of only 5% of cases. NNFI increased from .90 (full sample) to .96 (5% trimmed) and SRMR reduced from .11 (full sample) to .07 (5% trimmed). SRMR reduced further to .06 with 7.5% of outlying cases trimmed. Mean residuals for pairs of similarly keyed items reduced from 3.97 (full sample) to 1.85 (5%) and .98 (10%); between pairs of differently keyed items they moved from -3.51 to -1.62 (5%) and -.91 (10%). Correlation between the positive and negative item factors increased from .58 (full sample) to .80 (5% trimmed), .85 (7.5%) and .89 (10%). MAP indicated a single factor with 18 of 1,876 cases (just under 1%) removed, and with this level of trimming, correlation between the positive and negative item factors increased from .57 to .77.

In Sample 1 (N = 263), Satorra Bentler chi square reduced from 246.0 (full sample) to 125.7 (5% trimmed sample). NNFI increased from .88 (full sample) to .96 (5% trimmed) and SRMR reduced from .10 (full sample) to .07 (5% trimmed). These improvements resulted from removing only 13 cases. SRMR reduced further to .06 with 7.5% of outlying cases trimmed. Mean residuals for pairs of similarly keyed items reduced from 1.52 (full sample) to .80 (5%) and .50 (10%); between pairs of differently keyed items they moved from -1.32 to -.68 (5%) and -.41 (10%). Correlation between the positive and negative item factors increased from .60 (full sample) to .78 (5% trimmed), .83 (7.5%) and .87 (10%). Dropping only one case (.4% of the sample) was enough for MAP to indicate a single factor in the data (and correlation increased from .60 to .63).

In the smallest sample (Sample 2, N = 178) Satorra Bentler chi square reduced from 243.3 (full sample) to 144.8 (5% trimmed sample – nine cases removed). NNFI increased from .86 (full sample) to .94 (5% trimmed) and to .96 with 10% of outlying

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

cases trimmed. SRMR reduced from .13 (full sample) to .09 (5% trimmed); it remained comparatively high, with a value of .08 with 10% of outlying cases trimmed. Mean residuals for pairs of similarly keyed items reduced from 1.47 (full sample) to .83 (5%) and .51 (10%); between pairs of differently keyed items they moved from -1.37 to -.75 (5%) and -.47 (10%). Correlation between the positive and negative item factors increased from .57 (full sample) to .78 (5% trimmed), .82 (7.5%) and .86 (10%). Dropping eight cases (4.5% of the sample) was enough for MAP to indicate a single factor in the data (and correlation increased from .58 to .67).

Thus, for three of the four samples, Velicer's MAP indicated only one factor underlying the correlation matrix with 1% or fewer of influential inconsistent response cases removed (compared with two factors in the full samples – see p. 74). In Sample 2, parallel analysis also indicated only 1 factor with 5% or more of outlying cases removed. In Samples 1 and 4, 10% of outliers had to be removed before parallel analysis indicated only one factor. However, in both samples, the second eigenvalue in the 7.5% trimmed sample data exceeded the 95th percentile value by a very small margin (.004 in Sample 1 and .006 in Sample 4). Rerunning the analyses with 8% of outlying cases removed resulted in parallel analysis indicating only one factor in all three samples.

Discussion

This study set out to determine whether Rensvold and Cheung's (1999) method of identifying outlying cases in structural equation models could be used to identify careless responding in empirical datasets. If so, could the emergence of an item wording factor be ameliorated by removing extreme cases from the sample?

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

The use of simulated data clearly shows that careless responses have the most influence in reducing correlation between positive and negative item factors, thereby reducing fit of the one factor model. At all levels of carelessness, the extreme right tail of the distribution of phi values was dominated by careless cases. As illustrated in *Figure 9* (a sample of 2000, with 20% careless cases), over 80% of the most extreme 100 outlying cases were careless[6].

By varying both the percentage of careless cases and the likelihood of answering a negative item carelessly, a more fine-grained picture of the effects of carelessness is obtained. When there is a low probability of answering items carelessly, then even quite high proportions of careless respondents will have only a minor effect on phi. It is only as the probability increases, that marked drops in the correlation between positive and negative item factors are observed. Thus, having 10% of cases answering all negative items carelessly is not equivalent to 20% of cases answering half of the negative items carelessly.

While it is not possible to know with certainty how many cases in empirical samples answer negatively keyed items carelessly, the value of phi in the positive and negative item factor model can provide some indication. The simulated samples showed a wide range of phi values, from a mean of .27 for the most careless condition, up to .97 for the lowest level of carelessness. If scale items have been carefully designed so that positively and negatively keyed items provide equivalent measures of a single latent variable, then the obtained value of phi provides a clue as to the level of carelessness. In the example analysed in this chapter, the obtained phi value of .68

---

[6] While this sample has a relatively high level of carelessness, the same pattern was also found in other samples. For example, with 10% of careless cases (and a probability of .5 that negative items would be answered carelessly), over 60% of the top 100 outlying cases were careless cases.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

suggested a level of carelessness consistent with 20% of cases answering (on average) half of the negatively keyed items carelessly.

Using the equivalent Monte Carlo sample as a guide, I demonstrated that removing a low number of outlying cases from Sample 3 resulted in a large improvement in the fit of a one-factor model. Dropping between 7.5% and 10% of outliers resulted in multiple indicators (MAP, parallel analysis, standardised residual analysis and model fit indices) indicating a one factor model. Considering that phi suggested as many as 20% of cases exhibited some degree of carelessness, this is an impressive result.

The same pattern applied to the other samples. With the extreme 8% of outlying cases trimmed, these samples all fitted a one factor model.

The dramatic increase in correlation between positive item and negative item factors is also noteworthy. In their discussion of positive and negative wording factors, Nunnally and Bernstein (1994) suggest that parsimony calls for treating the positive and negative items as unidimensional if the separate scales correlate highly (i.e., .7 or higher). None of the CSES samples had correlations this high, but with removal of 5% of outliers, all of them exceeded .7 and, with 7.5% removed, all samples had correlations exceeding .80. This correlation suggests that the two subscales are measuring the same construct; Judge et al. (2003) reported an average reliability of .84 for the scale, and a test-retest reliability of .80.

MAP indicated that removing as few as 1% of influential cases was sufficient to eliminate the wording factor in three of the four samples. This is consistent with O'Connor's (2000) observation that MAP (if it's wrong) tends to underextract, while PA (if wrong) tends to overextract. While CFA model fit indices were still not

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

acceptable with so few cases removed, the finding suggests that use of non-SEM methods such as regression could be considered at this level of sample trimming.

The samples used in the study varied in terms of sample size (178 to 2,067), the number of rating points (5-point and 7-point scales), and geographic source (Singapore, Belgium, New Zealand). This strongly suggests that removal of influential cases with inconsistent responding to negative items is likely to be effective in removing item wording effects in a wide range of scales.

In this analysis, it has been assumed that careless responding to negatively keyed items is the only reason for a reduction in the correlation between positive and negative item factors. While items in the Monte Carlo simulation have the same distributional characteristics, and all inter-item correlations have the same mean value, real items are not so homogenous. In the CSES data, it is unlikely that the set of positive items perfectly matches the set of negative items in acting as indicators of the underlying core self-evaluations latent. Furthermore, the less-than-perfect reliability of the item subsets places an upper bound on the possible correlation between factors. However, the results of this analysis strongly support carelessness as an important determinant, and the value of the jackknife method in reducing its effects.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

CHAPTER SIX

SUMMARY

Summated rating scales (including Likert scales) are undoubtedly the most widely used measurement tool in management, psychology, and other social science disciplines. The scales typically consist of between five and ten statements, each rated on five to seven ordered response categories. It is common practice to include both positive and negative items, so as to encourage thoughtful answers, and reduce the likelihood of acquiescent or stereotyped responding. The popularity of such scales is due to their ease of development, simplicity of use, and flexibility in administration (paper-based, web, or interview). They are also particularly suited for use in structural equation modelling, a technique which allows estimation of latent variables by modelling the error in responses to items acting as indicators of these variables.

One of the most intriguing and frequently observed phenomenon with summated rating scales is the emergence of a two-factor solution in a set of positively and negatively worded items designed to measure a single construct (e.g., Carmines & Zeller, 1979; Gordon et al., 1980; Knight et al., 1988; Kohn & Schooler, 1969). Finding a two-factor solution in a purportedly unidimensional scale can have profound implications for both theory and practice. If our theory is predicated on a single construct (say, 'self esteem') but our attempts to measure it always result in two factors (e.g., Kohn and Schooler's (1969) 'Self-confidence' and 'Self-deprecation') we need to revisit either our theory or our measurement. In practical terms, we are faced with the complication of computing two scale scores instead of one, or on judging when the correlation between factors is high enough to allow for items to be combined into a single score (Nunnally & Bernstein, 1994). The problem is exacerbated in structural

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

equation models, where the specification of a theoretically-implied single factor model for such scales may result in unacceptably bad fit.

Several researchers have sought to explain the emergence of a second factor resulting from the use of both positive and negative items in scales. Some interpret both factors as substantive (e.g., Gordon et al., 1980), while others attribute the second factor to the complexity of negative statements (Benson & Hocevar, 1985), differences in verbal ability of respondents (Marsh, 1986), carelessness (Schmitt & Stults, 1985), method effects (e.g., Idaszak & Drasgow, 1987), or differing item response distributions (Spector et al., 1997).

This dissertation presents the results of a series of empirical and simulation studies suggesting that the two-factor solution, in many cases, is an artefact of careless responding. By careless responding, I mean an aberrant response pattern in which respondents make seemingly contradictory responses to items tapping similar content. For example, a person who strongly agrees with the statement: "I take a positive attitude toward myself" would be expected to disagree with the statement: "All in all, I am inclined to feel that I am a failure." Strong agreement with both statements is an aberrant or inconsistent response pattern, suggestive of careless or confused responding (Nunnally, 1967).

Such carelessness can be intentional (for example, in the case of respondents who have little motivation to complete the scale accurately) or unintentional (for example, where a respondent fails to notice a negative qualifier in an item). Regardless of the reason, the studies reported here indicate that as few as 5–10% of respondents answering carelessly are sufficient to produce an artifactual second factor in a single-construct scale. It is eminently feasible that 5–10% of subjects in samples typically

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

used in research (both students and working adults) could lack the motivation to answer all items in a focused, careful and reflective manner.

The seriousness of the problem, and its likely prevalence, argue for improved means by which to detect and remedy such aberrant responding. The final studies in this dissertation demonstrate (in both simulated and empirical samples) the effectiveness of an approach based on identifying influential cases.

In the following section, I summarise the studies and their conclusions in more detail. I will then discuss alternative means of dealing with the problem when it arises in empirical samples, before suggesting additional research directions implied by my findings.

Summary of Key Findings

In Chapter 3 I demonstrated that the problem of item wording factors associated with negatively-keyed items is still a current issue. Four datasets from different populations revealed the existence of a wording factor in a recently published single-construct scale with six positively and six negatively keyed items (Judge et al.'s (2003) CSES). A single factor CFA had unacceptable fit in all samples. Fit could only be achieved by taking wording effects into account (by allowing uniquenesses to vary between same-keyed items, by modelling a separate wording factor in addition to the substantive core self-evaluations latent, or by having separate positive-item and negative-item factors).

Having identified the presence of an item wording factor in these samples, I used the data to identify patterns of aberrant item responses consistent with carelessness. Carelessness has been identified as an important possible explanation for the emergence of item wording effects, but no studies have systematically looked for

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

varying patterns of carelessness in data. Simulations of careless responding (Schmitt & Stults, 1985; Woods, 2006) have assumed that careless respondents are careless when answering all negatively worded items.

All four samples showed diverse patterns of aberrant responding. Some respondents answered all six negative items in a direction inconsistent with their positive item responses, but many more responded inconsistently to five, four, or fewer. In cases where fewer than six negative items were answered inconsistently, the actual items varied across cases. Clearly the traditional assumption of uniform careless responding to all negatively-keyed items (by careless respondents) is unrealistic.

This analysis enabled me to create a realistic model of careless responding, and to use Monte Carlo simulation to identify the effect of careless responding on factor structure and model fit in a CFA framework. In Chapter 4 I used parallel analysis and MAP to identify the emergence of a wording factor as a function of carelessness, and CFA fit indices to assess the impact on model fit. The factor emerged with relatively low levels of careless responding, and the effect was slightly more marked in the simulated narrow construct scale (with higher average inter-item correlations). It emerged earlier in the 7-point scale than the 5-point scale, but results for fit indices (SRMR, NNFI) were essentially the same across both formats.

An interesting finding was that high levels of carelessness created an item wording factor in items whose distributions were sampled from a normal population. Thus, even if researchers expect an item to have a normal response distribution, sample variation in actual skew levels can result in an item wording factor.

Finally, in Chapter 5 I used Rensvold and Cheung's (1999) jackknife technique (designed to identify influential outliers in latent variable analysis) to demonstrate that extreme careless cases in the Monte Carlo simulation could be reliably identified. By

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

applying this finding to empirical CSES samples, I demonstrate that removal of small numbers of outlying cases can significantly reduce chi-square, and achieve acceptable fit for a single factor CFA model.

Implications

While some researchers have recommended avoiding the use of negative items (Schriesheim et al., 1991) or excluding them from scale scores (Marsh, 1996), this is not always possible. Rorer (1965) pointed out the need for wording to be consistent with content, and that some content required negative wording; linguists have also pointed out the central and indispensable role played by negation in language (e.g., L. R. Horn, 1989).

While this dissertation has focused on negatively keyed items as indicators of an item wording factor, such items are also valid indicators of the substantive latent variable. This is true of all the CSES items (except Item 2) in the four samples analysed in Chapter 3, and has been shown to be true of negatively keyed items in other scales (e.g., Bayazit et al., 2004; Magazine et al., 1996). It is therefore likely that new scales containing negatively-keyed items will continue to be developed, and existing mixed-item scales will continue to be used. The results of this dissertation research have implications for developers and users of such scales.

Whenever a scale uses both positively and negatively keyed items, the existence of an artifactual factor related to item wording should be suspected. If confirmed (by exploratory factor analysis or the comparison of CFA models such as those presented in *Figure 1*, p. 67), it must be addressed. This dissertation discusses two strategies in depth, which may be used to resolve the problem: (1) Identify and remove careless responders; or (2) model the item wording effects explicitly.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Identify and Remove Careless Responders

Application of the jackknife technique to simulated data confirmed its efficacy in identifying inconsistent responders. When applied to empirical data, the jackknife procedure enabled the identification and removal of influential aberrant cases (in terms of inconsistent responding between positive and negative items). When ranked in terms of descending correlation between the positive item and negative item factors, removal of the most extreme 5–10% of the sample was sufficient to remove the item wording factor in all samples.

Clearly, any cases considered for removal from a sample need to be reviewed carefully. In a typical sample of around 300 cases, this may mean reviewing the response patterns of 15-30 respondents, seeking to clarify the process underlying the inconsistency. Obvious explanations (e.g., coding errors) should be investigated. Individuals who have answered all or most of the negative items in a way which is highly inconsistent with their positive item responses (e.g., Cases 1343 and 1435, Table 6) could justifiably be considered to be answering carelessly. Most researchers would agree that it is appropriate to drop a case with responses of '717171717171.' Similarly, a pattern of '617171535171' (Case 1343 in Sample 3) represents an original response pattern of '677777555777' and is also likely to represent careless or non-attentive responding.

Consideration should be given to the possibility that identified aberrant cases are non-representative of the scale's intended population. Respondents whose first language differs from that of the questionnaire, who come from different cultures (Wong, Rindfleisch, & Burroughs, 2003), or who are low in cognitive ability (Cordery & Sevastos, 1993) may well respond in ways that are inconsistent with the intended

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

scale structure. It may be appropriate to exclude such cases in order to obtain fit for a theoretically justified model, designed to describe relationships within a clearly delineated population. However, any decision to drop influential cases from the analysis needs to be reported, along with the justification for doing so (Rensvold & Cheung, 1999).

Given the time and costs involved in gaining adequate sample sizes for quality studies, researchers are understandably reluctant to drop cases unless clearly necessary. However, procedures for identifying and removing traditional outliers are commonly applied to samples of data in order to maintain data integrity, and to avoid making misleading inferences from non-representative cases (Roth & Switzer, 2002). The jackknife procedure identifies influential cases which have highly inconsistent responses across positively-keyed and negatively-keyed items. These cases are responding in a manner which is incompatible with the assumption that both types of items are measuring the same underlying construct. Retaining such anomalous cases in the sample so as to maximise sample size is likely to do more harm than good.

While the jackknife procedure is not yet automated in standard software, it is quite feasible for researchers or practitioners to implement. The Python syntax used in this study (p. 189) appears relatively complex because it is designed to handle different samples, to generate LISREL syntax automatically, and to run several different CFA models. The actual generation of the jackknife samples takes only eight lines. A simple approach is to save the data in a delimited text file, with each case on a separate line. A scripting or programming language can then be used to loop through the data N times (where N is the sample size). On the first loop, all lines (cases) except the first one are written to a separate text file. On the second loop, all lines except the second are appended to this file, and so on until, on the N^{th} loop, all cases except the N^{th} case are

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

appended to the file. This will create a single data file with N * N-1 lines, representing the N different jackknife samples. All major structural equation software programs (e.g., LISREL, Mplus) are able to take such a file as input, analyse all N samples, and report fit indices and parameter values in a form amenable to subsequent sorting and analysis.

Model the Wording Effects Explicitly

In the event that researchers are unwilling to drop aberrant cases, it will be necessary to model the wording effects explicitly. Alternatives (such as allowing error terms of negatively-keyed items to covary; parcelling subsets of items; and use of summed scale scores) are often used in practice, but have their drawbacks.

Allowing error terms to covary can be efficient when there are only two or three negatively-keyed items in a scale, but quickly uses up degrees of freedom as the number of items increases. The use of item parcelling to combine items (e.g., D. J. Brown et al., 2007) obscures the existence of the wording factor, and can result in a well fitting model. However, Bandalos (1993) found that parcelling items from scales which included a main factor (indicated by all items) and a secondary factor (indicated by a subset of items) could lead to biased parameter estimates, and overly optimistic indications of model fit. Thus, item parcelling is not an ideal solution.

Using the summed scale score in a regression model (for recent examples, see Beal et al., 2006; Nikolaou & Judge, 2007; Tsaousis et al., 2007; Wanberg et al., 2005) can be appropriate, but foregoes the benefits of using structural equation modeling. Johnson, Rosen, and Levy (2008) critique this approach in the context of measuring core self-evaluations. Drawing on Edwards (2001), they point out that a summed scale score "disregards measurement error and confounds it with the unique variance

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

associated with each indicator, it fails to capture differences in relationships between constructs and their indicators, and it ignores relationships between indicators and their measures" (p. 402).

As indicated by the analyses in Chapter 3, modelling a negative item method factor in addition to the substantive latent (Model 3 in *Figure 1*, p. 67) will achieve acceptable model fit while staying true to the data. This approach is therefore preferable to use of parcelling or summed scores. As noted by Lance et al. (2002), the method factor will remove item wording variance from the scale, allowing for accurate estimates of relationships between substantive latents. It avoids the risk of overestimating the strength of relationship based on item wording factor variance shared between different substantive latent variables, and gives a more accurate estimate of the contribution of individual items.

Areas for Future Research

The jackknife technique can be used to gain further insights into the nature of respondents who answer survey items inconsistently. While this dissertation explored inconsistent responding in a single measure, researchers often gather information on multiple scales from the same respondents. By identifying inconsistent responders using the jackknife approach separately on each scale, it will be possible to answer a number of questions related to carelessness. For example:

- Do respondents show an increased tendency for inconsistency on scales which are administered near the end of a session (suggesting the possible influence of fatigue)?
- Do some respondents answer all scales inconsistently (suggesting a possible personality or other individual difference cause)?

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

- Are inconsistent responders distinguished by any aspects of the sampling process? For example, are respondents who complete a mail survey after the second or third reminder more likely to answer inconsistently than those who answer immediately? Are students completing a measure for a course requirement less consistent than those who have a more intrinsic motivation for completing it?

Studies aimed at manipulating the care with which respondents complete scales could be used to test hypotheses implied by the above questions. For example, Trott and Jackson (1967) were able to factorially separate positive and negative items by speeding up presentation of items. Competing explanations for the emergence of the wording factor (e.g., low reading ability versus low motivation to respond accurately) could be tested with more precision now that it is possible to identify the most influential inconsistent responders in samples.

As demonstrated in Chapter 4 (p. 99), it is possible to use the jackknife procedure to rank cases according to each negative item's factor loading in turn. This allows the identification of individual items which are most prone to inconsistent responding, and creates the possibility of carrying out studies into characteristics associated with such items. Low's (1996) 'Lexical Invisibility Hypothesis' suggests that questionnaire respondents sometimes overlook words used as modifiers in questionnaire items. Are they more likely to overlook short prefix modifiers used to negate ideas (e.g., 'un' or 'in'), or to miss negating words ('never', 'not')? Comparing the proportions of respondents answering different item formats inconsistently will help answer such questions.

Another avenue of research flows from my demonstration that removing outliers results in high agreement (comparable to overall scale reliability) between

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

positive item and negative item subscales in the CSES. Such a finding strongly supports the case for careless responding as the underlying process creating the two factors. If positive items were measuring a construct which differed substantively from that measured by negative items, removing such small numbers of cases would not change the factor structure. The technique could be used to help resolve debates over whether to make substantive interpretations of scales comprising negative items in other measures (e.g., Gordon et al.'s (1980) union commitment scale, discussed on p. 34).

Given the ubiquity of summated rating scales in organisational research, and the likely continued use of negatively keyed items, research into understanding and controlling item wording effects will have ongoing payoff. More immediately, application of the technique for identifying and removing aberrant cases will help overcome a persistent and troublesome problem associated with use of positive and negative items in scales. The resulting improvement in our measurement quality will lead to a commensurate increase in the quality of our knowledge.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

REFERENCES

Ahlawat, K. S. (1985). On the negative valence items in self-report measures. [Article]. *Journal of General Psychology, 112*(1), 89–99.

Allport, F. H., & Allport, G. W. (1921). Personality traits: Their classification and measurement. *The Journal of Abnormal Psychology and Social Psychology, 16*(1), 6–40.

Aluja, A., Rolland, J.-P., García, L. F., & Rossier, J. (2007). Dimensionality of the Rosenberg Self-Esteem Scale and its relationships with the three-and the five-factor personality models [Brief Report]. *Journal of Personality Assessment, 88*(2), 246–249.

Andrich, D. (1988). The application of an unfolding model of the PIRT type to the measurement of attitude. *Applied Psychological Measurement, 12*(1), 33–51.

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.

Bachman, J. G., & O'Malley, P. M. (1986). Self-concepts, self-esteem, and educational experiences: The frog pond revisited (again). *Journal of Personality and Social Psychology, 50*(1), 35–46.

Bandalos, D. L. (1993). Factors influencing cross-validation of confirmatory factor analysis models. *Multivariate Behavioral Research, 28*(3), 351–374

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling, 9*(1), 78–102.

Barnette, J. J. (1995). *Responses That May Indicate Nonattending Behaviors in Three Self-Administered Educational Surveys*. Paper presented at the Annual Meeting of the Mid-South Educational Research Association (November 9, 1995). from http://eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED393897

Barnette, J. J. (1997). *Effects of Item and Response Set Reversals on Survey Statistics.* Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).

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.

Baumgardner, H. W. (1935). Measuring Negro self-respect. *The Journal of Negro Education, 4*(4), 490-499.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Bayazit, M., Hammer, T. H., & Wazeter, D. L. (2004). Methodological challenges in union commitment studies. *Journal of Applied Psychology, 89*(4), 738–747.

Beal, D. J., Trougakos, J. P., Weiss, H. M., & Green, S. G. (2006). Episodic processes in emotional labor: Perceptions of affective delivery and regulation strategies. *Journal of Applied Psychology, 91*(5), 1053–1065.

Benjamin, W. (2002, April 1–5). *Development and validation of student teaching performance assessment based on Danielson's framework for teaching.* Paper presented at the Annual meeting of the American Educational Research Association, April 1–5, New Orleans, LA.

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

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

Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105*(3), 467–477.

Bird, C. (1940). *Social psychology*. New York: Appleton-Century.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 395–479). Reading, MA: Addison Wesley.

Blau, G. (2007). Partially testing a process model for understanding victim responses to an anticipated worksite closure. *Journal of Vocational Behavior, 71*(3), 401–428.

Block, J. (1971). On further conjectures regarding acquiescence. *Psychological Bulletin, 76*(3), 205–210.

Bollen, K. A., & Barb, K. H. (1981). Pearson's r and coarsely categorized measures. *American Sociological Review, 46*(2), 232–239.

Bono, J. E., & Colbert, A. E. (2005). Understanding responses to multi-source feedback: The role of core self-evaluations. *Personnel Psychology, 58*(1), 171–203.

Bottom, W. P. (2006). *Before the Ford Foundation: Development of the Research Based Model of Business Education.* Paper presented at the Academy of Management Annual Meeting, Atlanta, Ga.

Boyd, S., Ensari, N., Hoffman, C. C., & Newman, D. A. (2007, April 27–29). *Core self-evaluations moderating the job stress—burnout relationship.* Paper presented at the Society for Industrial and Organizational Psychology, New York, NY.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

British Telecom Occupational Psychology Unit. (1984a). *Survey Item Bank: Measures of Organisational Characteristics* (Vol. 2). Bradford, UK: MCB University Press.

British Telecom Occupational Psychology Unit. (1984b). *Survey Item Bank: Measures of Satisfaction* (Vol. 1). Bradford, UK: MCB University Press.

Brown, D. J., Ferris, D. L., Heller, D., & Keeping, L. M. (2007). Antecedents and consequences of the frequency of upward and downward social comparisons at work. *Organizational Behavior and Human Decision Processes, 102*(1), 59–75.

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York: Guilford Press.

Brunborg, G. S. (2008). Core self-evaluations: A predictor variable for job stress. *European Psychologist, 13*(2), 96–102.

Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research, 27*(4), 509–540.

Capon, J., Chernyshenko, O. S., & Stark, S. (2007). Applicability of civilian retention theory in the New Zealand military. *New Zealand Journal of Psychology, 36*(1), 50–56.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and Validity Assessment*. Beverly Hills, CA: Sage Publications.

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. *Journal of Personality and Social Psychology, 67*(2), 319–333.

Chamorro-Premuzic, T., Ahmetoglu, G., & Furnham, A. (2008). Little more than personality: Dispositional determinants of test anxiety (the Big Five, core self-evaluations, and self-assessed intelligence). *Learning and Individual Differences, 18*(2), 258–263.

Chang, L. (1995). Connotatively inconsistent test items. *Applied Measurement in Education, 8*(3), 199–209.

Churchill, G. A., Jr. (1979). A paradigm for developing better measures of marketing constructs. [Article]. *Journal of Marketing Research, 16*(1), 64–73.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309–319.

Cliff, N., Collins, L. M., Zatkin, J., Gallipeau, D., & McCormick, D. J. (1988). An ordinal scaling method for questionnaire and other ordinal data. *Applied Psychological Measurement, 12*(1), 83–97.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Cordery, J. L., & Sevastos, P. P. (1993). Responses to the original and revised Job Diagnostic Survey : Is education a factor in responses to negatively worded items? *Journal of Applied Psychology, 78*(1), 141–143.

Corwyn, R. F. (2000). The factor structure of global self-esteem among adolescents and adults. *Journal of Research in Personality, 34*(4), 357–379.

Cronbach, L. J. (1941). An experimental comparison of the multiple true-false and multiple multiple-choice tests. *Journal of Educational Psychology, 32*(7), 533–543.

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

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6*, 475–494.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*(4), 349–354.

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

Dunbar, M., Ford, G., Hunt, K., & Der, G. (2000). Question wording effects in the assessment of global self-esteem. *European Journal of Psychological Assessment, 16*(1), 13–19.

Edwards, A. L., & Kenney, K. C. (1946). A comparison of the Thurstone and Likert techniques of attitude scale construction. *Journal of Applied Psychology, 30*(1), 72–-83.

Edwards, J. R. (1994). The study of congruence in organizational behavior research: Critique and a proposed alternative. *Organizational Behavior and Human Decision Processes, 58*(1), 51–100.

Edwards, J. R. (2001). Multidimensional constructs in organizational behavior research: An integrative analytical framework. *Organizational Research Methods, 4*(2), 144–192.

Erez, A., & Judge, T. A. (2001). Relationship of core self-evaluations to goal setting, motivation, and performance. *Journal of Applied Psychology, 86*(6), 1270–1279.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272–299.

Fenigstein, A., Scheier, M. F., & Buss, A. H. (1975). Public and private self-consciousness: Assessment and theory. *Journal of Consulting and Clinical Psychology, 43*(4), 522–527.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

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.

Fornell, C. (1983). Issues in the application of covariance structure analysis: A comment. *Journal of Consumer Research, 9*(4), 443–448.

Fritz, M. F. (1927). Guessing in a true-false test. *Journal of Educational Psychology, 18*(8), 558–561.

Fuchs, A. H., & Milar, K. S. (2002). Psychology as a science. In D. K. Freedheim & I. B. Weiner (Eds.), *Handbook of Psychology* (Vol. 1: History of Psychology, pp. 1–26).

Gana, K., Alaphilippe, D., & Bailly, N. (2005). Factorial structure of the French version of the Rosenberg Self-Esteem Scale among the elderly. *International Journal of Testing, 5*(2), 169–176.

Gerbing, D. W., & Anderson, J. C. (1984). On the meaning of within-factor correlated measurement errors. [Article]. *Journal of Consumer Research, 11*(1), 572–578.

Gerbing, D. W., & Anderson, J. C. (1988). An dpdated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research, 25*(2), 186–192.

Gerbing, D. W., & Anderson, J. C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 40–65). Newbury Park: Sage Publications.

Gilley, W. F., & Uhlig, G. E. (1993). Factor analysis and ordinal data. *Education, 114*(2), 258–264.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt & F. Ostendorf (Eds.), *Personality Psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

Goldsmith, R. E. (1986). Dimensionality of the Rosenberg Self-Esteem Scale. *Journal of Social Behavior and Personality, 1*(2), 253–264.

Gordon, M. E., Philpot, J. W., Burt, R. E., Thompson, C. A., & Spiller, W. E. (1980). Commitment to the union: Development of a measure and an examination of its correlates. *Journal of Applied Psychology, 65*(4), 479–499.

Gorsuch, R. L. (1990). Common factor analysis versus component analysis: Some well and little known facts. *Multivariate Behavioral Research, 25*(1), 33–39.

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.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Greenwald, H. J., & Satow, Y. (1970). A short social desirability scale. *Psychological Reports, 27*(1), 131–135.

Hackman, J. R., & Oldham, G. R. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology, 60*(2), 159–170.

Haig, B. D., & Borsboom, D. (2008). On the conceptual foundations of psychological measurement. *Measurement: Interdisciplinary Research & Perspective, 6*(1), 1–6.

Hart, E., Leary, M. R., & Rejeski, W. J. (1989). The measurement of social physique anxiety. *Journal of Sport and Exercise Psychology, 11*(1), 94–104.

Hartshorne, T. S. (1993). Psychometric properties and confirmatory factor analysis of the UCLA Loneliness Scale. *Journal of Personality Assessment, 61*(1), 182–195.

Hau, K.-T., & Marsh, H. W. (2004). The use of item parcels in structural equation modelling: Non-normal data and small sample sizes. *British Journal of Mathematical & Statistical Psychology, 57*(2), 327–351.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*(2), 191–205.

Henmon, V. A. C. (1919). Air service tests of aptitude for flying. *Journal of Applied Psychology, 3*(2), 103–109.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1*(1), 104–121.

Holden, R. R., & Fekken, G. C. (1990). Structured psychopathological test item characteristics and validity. *Psychological Assessment, 2*(1), 35–40.

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.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185.

Horn, L. R. (1989). *A Natural History of Negation*. Chicago: University of Chicago Press.

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

Hunt, E. (1978). Mechanics of verbal ability. *Psychological Review, 85*(2), 109–130.

Hurley, A. E., Scandura, T. A., Schriesheim, C. A., Brannick, M. T., Seers, A., Vandenberg, R. J., et al. (1997). Exploratory and confirmatory factor analysis:

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Guidelines, issues, and alternatives. *Journal of Organizational Behavior, 18*(6), 667–683.

Idaszak, J. R., & Drasgow, F. (1987). A revision of the Job Diagnostic Survey: Elimination of a measurement artifact. *Journal of Applied Psychology, 72*(1), 69–74.

Jackson, D. N., & Messick, S. J. (1957). A note on 'ethnocentrism' and acquiescent response sets. *The Journal of Abnormal and Social Psychology, 54*(1), 132-134.

Jackson, D. N., & Messick, S. J. (1958). Content and style in personality assessment. *Psychological Bulletin, 55*(4), 243–252.

Johnson, E. C., Kristof-Brown, A. L., Van Vianen, A. E. M., De Pater, I. E., & Klein, M. R. (2003). Expatriate social ties: Personality antecedents and consequences for adjustment. *International Journal of Selection and Assessment, 11*(4), 277–288.

Johnson, R. D., Marakas, G. M., & Palmer, J. W. (2006). Differential social attributions toward computing technology: An empirical investigation. *International Journal of Human-Computer Studies, 64*(5), 446–460.

Johnson, R. E., Rosen, C. C., & Levy, P. E. (2008). Getting to the core of core self-evaluation: A review and recommendations. *Journal of Organizational Behavior, 29*(3), 391–413.

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 294–316). Newbury Park: Sage Publications.

Jöreskog, K. G. (2002). *Structural Equation Modeling with Ordinal Variables using LISREL*. Lincolnwood, IL: Scientific Software International, Inc.

Jöreskog, K. G., & Sörbom, D. (1996). *PRELIS 2: A User's Reference Guide*. Chicago: Scientific Software International, Inc.

Jöreskog, K. G., & Sörbom, D. (1999). *LISREL 8: User's Reference Guide*. Chicago: Scientific Software International.

Judge, T. A., & Bono, J. E. (2001). Relationship of core self-evaluations traits—self-esteem, generalized self-efficacy, locus of control, and emotional stability—with job satisfaction and job performance: A meta-analysis. *Journal of Applied Psychology, 86*(1), 80–92.

Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology, 87*(4), 765–780.

Judge, T. A., Bono, J. E., & Locke, E. A. (2000). Personality and job satisfaction: The mediating role of job characteristics. *Journal of Applied Psychology, 85*(2), 237–249.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2002). Are measures of self-esteem, neuroticism, locus of control, and generalized self-efficacy indicators of a common core construct? *Journal of Personality and Social Psychology, 83*(3), 693–710.

Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2003). The core self-evaluations scale: Development of a measure. *Personnel Psychology, 56*(2), 303–331.

Judge, T. A., Locke, E. A., & Durham, C. C. (1997). The dispositional causes of job satisfaction: A core evaluations approach. *Research in Organizational Behavior, 19*, 151–188.

Judge, T. A., Locke, E. A., Durham, C. C., & Kluger, A. N. (1998). Dispositional effects on job and life satisfaction: The role of core evaluations. *Journal of Applied Psychology, 83*(1), 17–34.

Judge, T. A., Van Vianen, A. E. M., & De Pater, I. E. (2004). Emotional stability, core self-evaluations, and job outcomes: A review of the evidence and an agenda for future research. *Human Performance, 17*(3), 325–346.

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika, 27*(2), 179–182.

Kamerlingh Onnes, H. (1983). The significance of quantitative research in physics: Inaugural address at the University of Leiden (1882). In H. Casimar (Ed.), *Haphazard Reality: Half a Century of Science* (pp. 160–161).

Kaufman, P., Rasinski, K. A., Lee, R., & West, J. (1991). *National Education Longitudinal Study of 1988. Quality of the responses of eighth-grade students in NELS88*. Washington, DC: U.S. Department of Education.

Kaup, B., Zwaan, R. A., & Lüdtke, J. (2007). The experiential view of language comprehension: How is negation represented? In F. Schmalhofer & C. A. Perfetti (Eds.), *Higher Level Language Processes in the Brain: Inference and Comprehension Processes* (pp. 255–288). Mahwah, NJ: Lawrence Erlbaum Associates.

Kelloway, E. K. (1998). *Using LISREL for Structural Equation Modeling: A Researcher's Guide*. Thousand Oaks, CA: Sage Publications Inc.

Kelloway, E. K., Catano, V. M., & Southwell, R. R. (1992). The construct validity of union commitment: Development and dimensionality of shorter scale. *Journal of Occupational & Organizational Psychology, 65*(3), 197–211.

Kevles, D. J. (1968). Testing the army's Intelligence: Psychologists and the military in World War I. *The Journal of American History, 55*(3), 565–581.

Kim, K. H. (2005). The relation among fit indexes, power, and sample size in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 12*(3), 368–390.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Kluemper, D. H. (2008). Trait emotional intelligence: The impact of core-self evaluations and social desirability. *Personality and Individual Differences, 44*(6), 1402–1412.

Knight, R. G., Chisholm, B. J., Marsh, N. V., & Godfrey, H. P. D. (1988). Some normative, reliability, and factor analytic data for the revised UCLA Loneliness Scale. *Journal of Clinical Psychology, 44*(2), 203–206.

Kohn, M. L., & Schooler, C. (1969). Class, occupation, and orientation. *American Sociological Review, 34*(5), 659–678.

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.

Leary, M. R. (1983). A brief version of the Fear of Negative Evaluation Scale. *Personality And Social Psychology Bulletin, 9*(3), 371–375.

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology, 140*, 1–55.

Loehlin, J. C. (1990). Component analysis versus common factor analysis: A case of disputed authorship. *Multivariate Behavioral Research, 25*(1), 29–31.

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51*(5), 493–504.

Lorge, I. (1937). Gen-like: Halo or reality [Conference paper abstract]. *Psychological Bulletin, 34*(8).

Low, G. (1996). Intensifiers and hedges in questionnaire items and the lexical invisibility hypothesis. *Applied Linguistics, 17*(1), 1–37.

Macintyre, S. (1987). *West of Scotland Twenty-07: Health in the Community. The Survey's Background and Rationale*. Glasgow: MRC Medical Sociology Unit.

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.

Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology, 22*(1), 37–49.

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70*(4), 810–819.

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* (pp. 177–198). Thousand Oaks, CA: Sage.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Marsh, H. W., Kit-Tai, H., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33*(2), 181–220.

Marshall, G. N., Wortman, C. B., Kusulas, J. W., & Hervig, L. K. (1992). Distinguishing optimism from pessimism: Relations to fundamental dimensions of mood and personality. *Journal of Personality and Social Psychology, 62*(6), 1067–1074.

Masin, S. C., & Agostini, A. (1991). Attentional scanning and space errors. *Perception & Psychophysics, 50*(3), 285–289.

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation, 8*(1), 3–30.

Mayer, L. S. (1971). A note on treating ordinal data as interval data. *American Sociological Review, 36*(3), 519–520.

McGee, G. W., & Ford, R. C. (1987). Two (or more?) dimensions of organizational commitment: Reexamination of the affective and continuance commitment scales. *Journal of Applied Psychology, 72*, 638–642.

McPherson, J., & Mohr, P. (2005). The role of item extremity in the emergence of keying-related factors: An exploration with the Life Orientation Test. *Psychological Methods, 10*(1), 120–131.

Meyer, J. P., & Allen, N. J. (1984). Testing the 'side-bet theory' of organizational commitment: Some methodological considerations. *Journal of Applied Psychology, 69*, 372–378.

Meyer, J. P., Allen, N. J., & Gellatly, I. R. (1990). Affective and continuance commitment to the organization: Evaluation measures and analysis of concurrent and time-lagged relations. *Journal of Applied Psychology, 75*(6), 710–720.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156–166.

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., et al. (2000). *Manual for the Patterns of Adaptive Learning Scales (PALS)*. Ann Arbor, MI: University of Michigan.

Miller, T. R., & Cleary, T. A. (1993). Direction of wording effects in balanced scales. *Educational and Psychological Measurement, 53*(1), 51–60.

Moser, C. A., & Kalton, G. (1971). *Survey Methods in Social Investigation* (2nd ed.). London: Heinemann Educational.

Nikolaou, I., & Judge, T. A. (2007). Fairness reactions to personnel selection techniques in Greece: The role of core self-evaluations. *International Journal of Selection and Assessment, 15*(2), 206–219.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Nunnally, J. C. (1967). *Psychometric Theory*. New York: McGraw-Hill.

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

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). New York: McGraw-Hill.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers, 32*(3), 396–402.

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.

Pajo, K., Guenole, N., Mallon, M., & Ward, R. (2005). *The impact of core self-evaluations on perceptions of support, commitment, and role-conflict.* Paper presented at the 65th Annual Academy of Management Conference, Hawaii, August 2005.

Peabody, D. (1961). Attitude content and agreement set in scales of authoritarianism, dogmatism, anti-Semitism, and economic conservatism. *The Journal of Abnormal and Social Psychology, 63*(1), 1–11.

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.

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics, 2*(1), 13–43.

Pressey, S. L. (1921). A group scale for investigating the emotions. *The Journal of Abnormal Psychology and Social Psychology, 16*(1), 55–64.

Quilty, L. C., Oakman, J. M., & Risko, E. (2006). Correlates of the Rosenberg Self-Esteem scale method effects. *Structural Equation Modeling, 13*(1), 99–117.

Rensvold, R. B., & Cheung, G. W. (1999). Identification of influential cases in structural equation models using the jackknife method. *Organizational Research Methods, 2*(3), 293–308.

Rorer, L. G. (1965). The great response-style myth. *Psychological Bulletin, 63*(3), 129–156.

Rosenberg, M. (1962). The association between self-esteem and anxiety. *Journal of Psychiatric Research, 1*(2), 135–152.

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Rosenberg, M. (1965). *Society and the Adolescent Self-Image*. Princeton, NJ: Princeton University Press.

Roth, P. L., & Switzer, F. S., III. (2002). Outliers and influential cases: Handling those discordant contaminated maverick rogues. In S. G. Rogelberg (Ed.), *Handbook of Research Methods in Industrial and Organizational Psychology* (pp. 297–309). Malden, MA: Blackwell Publishers.

Russell, D. W. (1996). UCLA Loneliness Scale (Version 3): Reliability, validity, and factor structure. *Journal of Personality Assessment, 66*(1), 20–40.

Russell, D. W., Peplau, L. A., & Cutrona, C. E. (1980). The revised UCLA Loneliness Scale: Concurrent and discriminant validity evidence. *Journal of Personality and Social Psychology, 39*(3), 472–480.

Russell, D. W., Peplau, L. A., & Ferguson, M. L. (1978). Developing a measure of loneliness. *Journal of Personality Assessment, 42*(3), 290–294.

Sager, J. K., Strutton, H. D., & Johnson, D. A. (2006). Core self-evaluations and salespeople. *Psychology and Marketing, 23*(2), 95–113.

Salvaggio, A. N., Schneider, B., Nishii, L. H., Mayer, D. M., Ramesh, A., & Lyon, J. S. (2007). Manager personality, manager service quality orientation, and service climate: Test of a model. *Journal of Applied Psychology, 92*(6), 1741–1750.

Samelson, F., & Yates, J. F. (1967). Acquiescence and the F Scale. *Psychological Bulletin, 68*(2), 91–103.

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

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.

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

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

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Schriesheim, C. A., & Kerr, S. (1974). Psychometric properties of the Ohio State leadership scales. *Psychological Bulletin, 81*(11), 756–765.

Simms, L. J. (2008). Classical and modern methods of psychological scale construction. *Social and Personality Psychology Compass, 2*(1), 414–433.

Snook, S. C., & Gorsuch, R. L. (1989). Component analysis versus common factor analysis: A Monte Carlo study. *Psychological Bulletin, 106*(1), 148–154.

Spector, P. E. (1992). *Summated Rating Scale Construction: An Introduction.* Newbury Park, CA: Sage Publications.

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

Spence Laschinger, H. K., & Finegan, J. (2008). Situational and dispositional predictors of nurse manager burnout: A time-lagged analysis. *Journal of Nursing Management, 16*(5), 601–607.

Spielberger, C. D., Vagg, P. R., Barker, L. R., Donham, G. W., & Westberry, L. G. (1980). The factor structure of the State-Trait Anxiety Inventory. In I. G. Sarason & C. D. Spielberger (Eds.), *Stress and Anxiety* (Vol. 7). New York: Hemisphere.

Starch, D. (1922). The use and limitations of psychological tests. *Harvard Business Review, 1*(1), 71–80.

Stogdill, R. M. (1963). *Manual for the Leader Behavior Description Questionnaire—Form XII.* Columbus: Bureau of Business Research, Ohio State University.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 10–39). Newbury Park: Sage Publications.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*(2), 193–210.

Thurstone, L. L. (1928). Attitudes can be measured. *The American Journal of Sociology, 33*(4), 529–554.

Thurstone, L. L., & Chave, E. J. (1929). *The Measurement of Attitude.* Chicago, Ill.,: The University of Chicago press.

Tomás, J. M., & Oliver, A. (1999). Rosenberg's self-esteem scale: Two factors or method effects. *Structural Equation Modeling, 6*(1), 84–98.

Trott, D. M., & Jackson, D. N. (1967). An experimental analysis of acquiescence. *Journal of Experimental Research in Personality, 2*, 278–288.

Tsaousis, I., Nikolaou, I., Serdaris, N., & Judge, T. A. (2007). Do the core self-evaluations moderate the relationship between subjective well-being and

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

physical and psychological health? *Personality and Individual Differences, 42*(8), 1441–1452.

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika, 48*(3), 465–471.

Vallières, F., & Vallerand, J. (1990). Traduction et validation canadienne-française de l'échelle de l'estime de soi de Rosenberg. *International Journal of Psychology, 25*, 305–316.

Van Hoye, G. (2006). *Social influences on organizational attractiveness: Word-of-mouth communication as a recruitment source.* Unpublished doctoral dissertation, Ghent University, Belgium.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*(3), 321–327.

Wanberg, C. R., Glomb, T. M., Song, Z., & Sorenson, S. (2005). Job-search persistence during unemployment: A 10-wave longitudinal study. *Journal of Applied Psychology, 90*(3), 411–430.

Weems, G. H., Onwuegbuzie, A. J., & Collins, K. M. T. (2006). The role of reading comprehension in responses to positively and negatively worded items on rating scales. *Evaluation and Research in Education, 19*(1), 3–20.

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

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and overextraction on principal axis factor analysis with varimax rotation. *Psychological Methods, 1*(4), 354–365.

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.

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34*(6), 806–838.

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

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Appendix A

Python Program Syntax Written for this Research

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Program for Generating Datasets

```
*       nbsamp   Number of random samples to generate and analyse
*       smpsize  The size of each sample (N = 200, 250 etc)
*       nbitems  Number of items in the scale (e.g. 12, 6)
*       nbrev    Number of items to reverse code - i.e. subject to carelessness (e.g. 6
for CSE model)
*       nbscal   Number of scale points (e.g. 5 or 7 rating points)
*       itemcor  The desired value for inter-item correlations (Could use Miller's
randgen to estimate RhoController)
*       caserev  The proportion of cases to have responses reversed
*       itemrev  The probability of each item being reversed, within a case selected in
(7) above
*   skew     Values = 2 (hi skew), 1 (lo skew), 0 (normal)
*   path        Location of input data files DEFAULT ('C:\temp\').
SET RNG=MT/MTINDEX=RANDOM .
BEGIN PROGRAM python.
import spss
nbsamp=200
smpsize=250
nbitems=12
nbrev=6
nbscal=5
# Value of .319 for itemcor gives r=.3 for 7-point scales; .743 gives r=.7 for non-
skewed 7-point scales, N=100000
# I think earlier progs were run using slightly higher value of .337
# Value of .338 for itemcor gives r=.3 for 5-point scales; .785 gives r=.7 for non-
skewed 5-point scales, N=100000
itemcor=.785
caserev=.25
itemrev=.0
skew = 1
path='C:/Documents and Settings/ajeffrey/My Documents/'
################################################################################
# This program is designed to generate                                         #
# skewed standard normal rv, with specified intercorrelations. These rv        #
# will then be categorised into ordered categorical (ordinal) items by         #
# using thresholds designed to create the target levels of skew and            #
# kurtosis at the item level. The following lines enter thresholds as          #
# tuples for 5 and 7 point scales, followed (after other variables) by         #
# lines creating string to use in later RECODE statements. Means only          #
# need to change thresholds in tuples, not in rest of code.                    #
################################################################################
T7hi = (-2.409, -2.054, -1.522, -1.155, -0.305, 0.813) # Estimated skew = -1.2
T7lo = (-2.9, -2.652, -1.799, -1.433, -0.295, 0.800)   # Estimated skew = -.83
T7norm = (-2.1429, -1.2857, -0.4286, 0.4286, 1.2857, 2.1429) # Equal intervals, zero
skew
T5hi = (-2.144, -1.461, -0.745, 1.237)                  # Estimated skew = -1.23
T5lo = (-2.6,  -1.8,  -0.94,   0.89)                   # Estimated skew -.83
# NOTE: Correction to values for T5Lo made July 19 2007.
T5norm = (-1.8000, -0.6000, 0.6000, 1.8000)             # Equal intervals, zero skew
recode7hi = "(LO THRU "+repr(T7hi[0])+" = 1) (LO THRU "+repr(T7hi[1])+" = 2) \
(LO THRU "+repr(T7hi[2]       )+" = 3) (LO THRU "+repr(T7hi[3])+" = 4) (LO THRU
"+repr(T7hi[4])+" \
 = 5) (LO THRU "+repr(T7hi[5])+" = 6) (LO THRU HI=7) "
recode7lo = "(LO THRU "+repr(T7lo[0])+" = 1) (LO THRU "+repr(T7lo[1])+" = 2) (LO THRU "
\
+repr(T7lo[2])+" = 3) (LO THRU "+repr(T7lo[3])+" = 4) (LO THRU "+repr(T7lo[4]) \
+" = 5) (LO THRU "+repr(T7lo[5])+" = 6) (LO THRU HI=7) "
recode7norm = "(LO THRU "+repr(T7norm[0])+" = 1) (LO THRU "+repr(T7norm[1])+" = 2) (LO
THRU " \
+repr(T7norm[2])+" = 3) (LO THRU "+repr(T7norm[3])+" = 4) (LO THRU "+repr(T7norm[4]) \
+" = 5) (LO THRU "+repr(T7norm[5])+" = 6) (LO THRU HI = 7) "
recode5hi = "(LO THRU "+repr(T5hi[0])+" = 1) (LO THRU "+repr(T5hi[1])+" = 2) (LO THRU "
\
+repr(T5hi[2])+" = 3) (LO THRU "+repr(T5hi[3])+" = 4) (LO THRU HI=5) "
recode5lo = "(LO THRU "+repr(T5lo[0])+" = 1) (LO THRU "+repr(T5lo[1])+" = 2) (LO THRU "
\
+repr(T5lo[2])+" = 3) (LO THRU "+repr(T5lo[3])+" = 4) (LO THRU HI=5) "
recode5norm = "(LO THRU "+repr(T5norm[0])+" = 1) (LO THRU "+repr(T5norm[1])+" = 2) (LO
THRU " \
+repr(T5norm[2])+" = 3) (LO THRU "+repr(T5norm[3])+" = 4) (LO THRU HI=5) "
varList=[]                                              # list 'rl' to 'rn' where
n=nbitems
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

```
for i in range(nbitems):
        varList.append('r'+repr(i+1))
rLast=varList[nbitems-1]                                    # r12 for 12 items, r6 for 6
items etc
prLast='p'+rLast
nrLast='n'+rLast
scleList=[]                                                 # to create numbers for variable
names
for i in range(nbscal):
    scleList.append(repr(i+1))
nLast=scleList[nbscal-1]                                    # e.g. 7 for 7-point scale, 5 for
5-point scale
nFirst=scleList[0]
BFirst='B'+nLast+'r1'                                       # B7r1 for 7-point, B5r1 for 5-
point scale
BLast='B'+nLast+rLast                                       #B7r12 for 12-item scale, using
7-point scale
revFirst='r_'+BFirst
revLast='r_'+'B'+repr(nbscal)+'r'+repr(nbrev)
BRevLast='B'+repr(nbscal)+'r'+repr(nbrev)                   #First nbrev items are neg worded
unirvLast='unirand'+repr(nbrev)                             #uniform rv for selecting
careless responses
# create varlists and file lists using 'for i in range(%(nbitems)s)' and
varlist.append. Can then use this list to
# generate %var values to use instead of the concatenated names (e.g. using index
nbitem-1 to get last var in list
for i in range(nbsamp):
        spss.Submit("""
INPUT PROGRAM .
VECTOR r(%(nbitems)s).
LOOP #i = 1 TO %(smpsize)s .
LOOP #j = 1 TO %(nbitems)s .
COMPUTE r(#j) = normal(1) .
END LOOP .
END CASE .
END LOOP .
END FILE .
END INPUT PROGRAM .
* Factor procedure computes n=nbitems principal component factor scores, which are
standard MVN .
* Default method of principal components is used. Resulting factor scores are
independent and MVN.
FACTOR variables = r1 to %(rLast)s /criteria = factors(%(nbitems)s) /save=reg (all,pr)
.
* use matrix to set corr matrix.
* x is a smpsize by nbitems matrix of independent standard normals .
* cor is the target covariance matrix .
* cho is the Cholesky factor of cor .
* newx is the smpsize by nbitems data matrix which has target covariance matrix .
MATRIX .
GET x / variables=pr1 to %(prLast)s.
COMPUTE cor=MAKE(%(nbitems)s,%(nbitems)s,%(itemcor)s) .
CALL SETDIAG(cor,1) .
COMPUTE cho=chol(cor).
COMPUTE newx=x*cho.
SAVE newx /outfile=* /variables= nr1 to %(nrLast)s .
END MATRIX.
DO IF %(skew)s EQ 2 .
 DO IF %(nbscal)s EQ 7 .
  RECODE nr1 TO %(nrLast)s %(recode7hi)s INTO %(BFirst)s TO %(BLast)s .
 ELSE IF %(nbscal)s EQ 5 .
  RECODE nr1 TO %(nrLast)s %(recode5hi)s  INTO %(BFirst)s TO %(BLast)s.
 END IF .
ELSE IF %(skew)s EQ 1 .
 DO IF %(nbscal)s EQ 7 .
  RECODE nr1 TO %(nrLast)s %(recode7lo)s INTO %(BFirst)s TO %(BLast)s .
 ELSE IF %(nbscal)s EQ 5 .
  RECODE nr1 TO %(nrLast)s %(recode5lo)s INTO %(BFirst)s TO %(BLast)s.
 END IF .
ELSE IF %(skew)s EQ 0 .
 DO IF %(nbscal)s EQ 7 .
  RECODE nr1 TO %(nrLast)s %(recode7norm)s  INTO %(BFirst)s TO %(BLast)s .
 ELSE IF %(nbscal)s EQ 5 .
  RECODE nr1 TO %(nrLast)s %(recode5norm)s INTO %(BFirst)s TO %(BLast)s.
  END IF.
END IF .
VARIABLE LEVEL %(BFirst)s TO %(BLast)s  (ORDINAL).
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

```
FORMATS %(BFirst)s TO %(BLast)s (F3.0).

EXECUTE .
* Create a set of nbrev variables = to the original negatively worded items .
* Later change the reversed items to new value so can keep track.
DO REPEAT temp1=%(revFirst)s to %(revLast)s
/temp2=%(BFirst)s to %(BRevLast)s .
COMPUTE temp1 = temp2 .
END REPEAT .
VARIABLE LEVEL %(revFirst)s TO %(revLast)s  (ORDINAL).
FORMATS %(revFirst)s TO %(revLast)s (F3.0).

* Select proportion of cases who are careless (caserev) and reverse randomly selected
proportion (itemrev) of item scores .
DO REPEAT temp4=unirand1 TO %(unirvLast)s  /score=%(BFirst)s TO %(BRevLast)s
/score1=%(revFirst)s TO %(revLast)s .
COMPUTE temp4 = uniform(1) .
* Use casenum and smpsize to select the cases who are careless responders .
DO IF $casenum < %(caserev)s * %(smpsize)s .
* Use itemrev and uniform random number to select items for reversal in careless cases
.
DO IF temp4 < %(itemrev)s .
* Reverse the value of the chosen score in the the r_B items .
COMPUTE score1 = %(nbscal)s+1-score .
END IF .
END IF .
END REPEAT .
        """%locals())
        posList=[]
        negList=[]
        crlsList=[]
        strtNeg=nbitems                         # assumes 12 rv N(0,1) appear in dataset
before scale items
        strtPos=nbitems + nbrev
        strtCrls=nbitems + (2*nbrev)
        endCrls=nbitems + 3*(nbrev)
        for j in range(strtNeg,strtPos):
                varName=spss.GetVariableName(j)
                spss.Submit(""" VARIABLE LABEL %s 'negItem' .""" %(varName))
                negList.append(varName)
        for j in range(strtPos,strtCrls):
                varName=spss.GetVariableName(j)
                spss.Submit(""" VARIABLE LABEL %s 'posItem' .""" %(varName))
                posList.append(varName)
        for j in range(strtCrls,endCrls):
                varName=spss.GetVariableName(j)
                spss.Submit(""" VARIABLE LABEL %s 'crlsItem' .""" %(varName))
                crlsList.append(varName)
        spss.Submit("COMPUTE meanPos = MEAN(" + ", ".join(posList)+")" +".")
        spss.Submit("""
* Write the values of run parameters to each file so they can be used when files
processed by other programs .
COMPUTE nbsamp=RND(%(nbsamp)s) .
COMPUTE smpsize=RND(%(smpsize)s) .
COMPUTE nbitems=RND(%(nbitems)s) .
COMPUTE nbscal=RND(%(nbscal)s) .
COMPUTE nbrev=RND(%(nbrev)s) .
COMPUTE itemcor=%(itemcor)s .
COMPUTE caserev=%(caserev)s .
COMPUTE itemrev=%(itemrev)s .
COMPUTE skew=RND(%(skew)s) .
EXECUTE .
VARIABLE LEVEL nbsamp TO nbrev, skew  (ORDINAL).
FORMATS nbsamp TO nbrev, skew (F5.0) .
FORMATS itemcor (F5.3) .
       """%locals())
       fileN=repr(i+1)
       spss.Submit("SAVE OUTFILE='"+path+fileN+".sav' .")
END PROGRAM.
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Program for Conducting Parallel Analysis

```
* Parallel Analysis Program For Raw Data and Data Permutations.
* Uses Brian O'Connor code http://flash.lakeheadu.ca/~boconno2/nfactors/nfactors.html .

* !RawPar nbsamp=5  nbitems=  nbscal=  ndatsets=  kind=  randtype=  percent=  path=   .
* nbsamp       Number of sample datasets to be read into the parallel analysis macro
DEFAULT (100)
* smpsize      Number of cases in sample (read from file)
* nbitems      Number of items in the scale DEFAULT (12)
* nbscal       Number of scale points (e.g. 5 or 7 rating points) DEFAULT (7)
* ndatsets     Desired number of parallel data sets  DEFAULT (100)
* kind         1 for principal components analysis
*                     2 for principal axis/common factor analysis DEFAULT (2)
* randtype     1 for normally distributed random data generation
*                     2 for permutations of the raw data set DEFAULT (2)
* percent      Desired percentile DEFAULT (95)
* path         Location of input data files DEFAULT ('C:\temp\').


BEGIN PROGRAM python.
############################################################################
# Output shows 1.sav as the data file name for each of the matrix runs,  #
# but printing out interim results (e.g. the a1, a2 matrices) shows that #
# it is analysing the required datasets.                                 #
############################################################################
import spss, os
ndatsets=100
kind=2
randtype=1
percent=95
path='C:/Documents and Settings/ajeffrey/My Documents/'
path2='C:\\DOCUME~1\\ajeffrey\\MYDOCU~1\\' # Used for sending commands to DOS with
os.system()
sav='.sav'          # Use for file naming e.g. in GET raw / FILE = command
txt='.txt'
pa='pa.sav'         # Use for saving results of parallel analysis
pa_sum='pa.txt'     # Use for accumulating N of factors identified across files
pa_all='pa_all'     # Name of file with all pa results (for importing into spss)
varList=[]          # Used to hold list of variables to carry out analysis on
############################################################################
# Following lines read the values of nbsamp, smpsize, nbitems and nbscal #
# from the first data set (1.sav). If you want to run the program on     #
# less than nbsamp samples, add the line nbsamp=n (where n is less than  #
# nbsamp) on the line following <dataCursor.close()>.                    #
############################################################################
spss.Submit("GET FILE='"+path+"1.sav' .")
dataCursor=spss.Cursor()
firstRow=dataCursor.fetchone()
for i in range(spss.GetVariableCount()):
    if spss.GetVariableName(i) == 'nbsamp':
        nbsamp=firstRow[i]
    if spss.GetVariableName(i) == 'smpsize':
        smpsize=firstRow[i]
    elif spss.GetVariableName(i) == 'nbitems':
        nbitems=firstRow[i]
    elif spss.GetVariableName(i) == 'nbscal':
        nbscal=firstRow[i]
dataCursor.close()
# print "nbsamp = ", nbsamp
# print "smpsize = ", smpsize
# print "nbitems = ", nbitems
# print "nbscal = ", nbscal
for j in range(spss.GetVariableCount()):
    if spss.GetVariableLabel(j)== 'posItem' or spss.GetVariableLabel(j)=='crlsItem':
        varList.append(spss.GetVariableName(j))
varList=" ".join(varList)
# print varList
spss.Submit("SET mxloops=18000 printback=off width=80 seed = 1953125 .")
# Begin loop for sequentially analysing the multiple data samples.
for cnt in range(nbsamp):
    indx=cnt+1
    fileN=repr(cnt+1)
    spss.Submit("""
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

```
MATRIX .
GET raw / FILE ='%(path)s%(fileN)s%(sav)s' / VAR = %(varList)s .
* COMPUTE ncases    = nrow(raw).
COMPUTE nvars     = ncol(raw).

* principal components analysis & random normal data generation.
DO IF (%(kind)s = 1 and %(randtype)s = 1).
COMPUTE nm1 = 1 / (%(smpsize)s-1).
COMPUTE vcv = nm1 * (sscp(raw) - ((t(csum(raw))*csum(raw))/%(smpsize)s)).
COMPUTE d = inv(mdiag(sqrt(diag(vcv)))).
COMPUTE realeval = eval(d * vcv * d).
COMPUTE evals = make(nvars,%(ndatsets)s,-9999).
LOOP #nds = 1 to %(ndatsets)s.
COMPUTE x = sqrt(2 * (ln(uniform(%(smpsize)s,nvars)) * -1) ) &*
            cos(6.283185 * uniform(%(smpsize)s,nvars) ).
COMPUTE vcv = nm1 * (sscp(x) - ((t(csum(x))*csum(x))/%(smpsize)s)).
COMPUTE d = inv(mdiag(sqrt(diag(vcv)))).
COMPUTE evals(:,#nds) = eval(d * vcv * d).
END LOOP.
END IF.

* principal components analysis & raw data permutation.
DO IF (%(kind)s = 1 and %(randtype)s = 2).
COMPUTE nm1 = 1 / (%(smpsize)s-1).
COMPUTE vcv = nm1 * (sscp(raw) - ((t(csum(raw))*csum(raw))/%(smpsize)s)).
COMPUTE d = inv(mdiag(sqrt(diag(vcv)))).
COMPUTE realeval = eval(d * vcv * d).
COMPUTE evals = make(nvars,%(ndatsets)s,-9999).
LOOP #nds = 1 to %(ndatsets)s.
COMPUTE x = raw.
LOOP #c = 1 to nvars.
LOOP #r = 1 to (%(smpsize)s -1).
COMPUTE k = trunc( (%(smpsize)s - #r + 1) * uniform(1,1) + 1 )  + #r - 1.
COMPUTE d = x(#r,#c).
COMPUTE x(#r,#c) = x(k,#c).
COMPUTE x(k,#c) = d.
END LOOP.
END LOOP.
COMPUTE vcv = nm1 * (sscp(x) - ((t(csum(x))*csum(x))/%(smpsize)s)).
COMPUTE d = inv(mdiag(sqrt(diag(vcv)))).
COMPUTE evals(:,#nds) = eval(d * vcv * d).
END LOOP.
END IF.

* PAF/common factor analysis & random normal data generation.
DO IF (%(kind)s = 2 and %(randtype)s = 1).
COMPUTE nm1 = 1 / (%(smpsize)s-1).
COMPUTE vcv = nm1 * (sscp(raw) - ((t(csum(raw))*csum(raw))/%(smpsize)s)).
COMPUTE d = inv(mdiag(sqrt(diag(vcv)))).
COMPUTE cr = (d * vcv * d).
COMPUTE smc = 1 - (1 &/ diag(inv(cr)) ).
CALL setdiag(cr,smc).
COMPUTE realeval = eval(cr).
COMPUTE evals = make(nvars,%(ndatsets)s,-9999).
COMPUTE nm1 = 1 / (%(smpsize)s-1).
LOOP #nds = 1 to %(ndatsets)s.
COMPUTE x = sqrt(2 * (ln(uniform(%(smpsize)s,nvars)) * -1) ) &*
            cos(6.283185 * uniform(%(smpsize)s,nvars) ).
COMPUTE vcv = nm1 * (sscp(x) - ((t(csum(x))*csum(x))/%(smpsize)s)).
COMPUTE d = inv(mdiag(sqrt(diag(vcv)))).
COMPUTE r = d * vcv * d.
COMPUTE smc = 1 - (1 &/ diag(inv(r)) ).
CALL setdiag(r,smc).
COMPUTE evals(:,#nds) = eval(r).
END LOOP.
END IF.

* PAF/common factor analysis & raw data permutation.
DO IF (%(kind)s = 2 and %(randtype)s = 2).
COMPUTE nm1 = 1 / (%(smpsize)s-1).
COMPUTE vcv = nm1 * (sscp(raw) - ((t(csum(raw))*csum(raw))/%(smpsize)s)).
COMPUTE d = inv(mdiag(sqrt(diag(vcv)))).
COMPUTE cr = (d * vcv * d).
COMPUTE smc = 1 - (1 &/ diag(inv(cr)) ).
CALL setdiag(cr,smc).
COMPUTE realeval = eval(cr).
COMPUTE evals = make(nvars,%(ndatsets)s,-9999).
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

```
COMPUTE nm1 = 1 / (%(smpsize)s-1).
LOOP #nds = 1 to %(ndatsets)s.
COMPUTE x = raw.
LOOP #c = 1 to nvars.
LOOP #r = 1 to (%(smpsize)s -1).
COMPUTE k = trunc( (%(smpsize)s - #r + 1) * uniform(1,1) + 1 )  + #r - 1.
COMPUTE d = x(#r,#c).
COMPUTE x(#r,#c) = x(k,#c).
COMPUTE x(k,#c) = d.
END LOOP.
END LOOP.
COMPUTE vcv = nm1 * (sscp(x) - ((t(csum(x))*csum(x))/%(smpsize)s)).
COMPUTE d = inv(mdiag(sqrt(diag(vcv)))).
COMPUTE r = d * vcv * d.
COMPUTE smc = 1 - (1 &/ diag(inv(r)) ).
CALL setdiag(r,smc).
COMPUTE evals(:,#nds) = eval(r).
END LOOP.
END IF.

* identifying the eigenvalues corresponding to the desired percentile.
COMPUTE num = rnd((%(percent)s*%(ndatsets)s)/100).
COMPUTE results = { t(1:nvars), realeval, t(1:nvars), t(1:nvars) }.
LOOP #root = 1 to nvars.
COMPUTE ranks = rnkorder(evals(#root,:)).
LOOP #col = 1 to %(ndatsets)s.
do if (ranks(1,#col) = num).
COMPUTE results(#root,4) = evals(#root,#col).
BREAK.
END IF.
END LOOP.
END LOOP.
COMPUTE results(:,3) = rsum(evals) / %(ndatsets)s.

* Uncomment next line and END IF at end of block to print results once every 20
samples.
DO IF (%(cnt)s+1 = 1 OR mod(%(cnt)s+1,20) = 0.0) .
PRINT /title="PARALLEL ANALYSIS:".
DO IF (%(kind)s = 1 and %(randtype)s = 1).
PRINT /title="Principal Components & Random Normal Data Generation".
ELSE IF (%(kind)s = 1 and %(randtype)s = 2).
PRINT /title="Principal Components & Raw Data Permutation".
ELSE IF (%(kind)s = 2 and %(randtype)s = 1).
PRINT /title="PAF/Common Factor Analysis & Random Normal Data Generation".
ELSE IF (%(kind)s = 2 and %(randtype)s = 2).
PRINT /title="PAF/Common Factor Analysis & Raw Data Permutation".
END IF.
COMPUTE specifs = {%(smpsize)s; nvars; %(ndatsets)s; %(percent)s}.
PRINT specifs /title="Specifications for this Run:"
 /rlabels="Ncases" "Nvars" "Ndatsets" "Percent".
PRINT results
 /title="Raw Data Eigenvalues, & Mean & Percentile Random Data Eigenvalues"
 /clabels="Root" "Raw Data" "Means" "Prcntyle"  /format "f12.6".
END IF .

COMPUTE root = results(:,1).
COMPUTE rawdata = results(:,2).
COMPUTE percntyl = results(:,4).
COMPUTE a1 = results > 0.0 .
COMPUTE a2 = a1 &* results .
COMPUTE a3 = a2(:,2) - a2(:,4) .
COMPUTE a4 ={%(indx)s,csum(a3 > 0)} .
PRINT a4 .
WRITE a4 /OUTFILE = '%(path)s%(indx)s%(pa_sum)s' /FIELD = 1 TO 10 .
* DO IF (%(cnt)s=0) .
* COMPUTE summary = a4 .
* ELSE .
* COMPUTE summary = {summary; a4} .
* END IF .
* DO IF (%(indx)s = %(nbsamp)s) .
* WRITE a4 /OUTFILE = '%(path)s%(pa_sum)s' /FIELD = 1 TO 10 BY 5.
* END IF .

* Uncomment following lines to see output of matrix caculations.
* PRINT a1 /title "a1 = results > 0"/ format=f5.2 .
* PRINT a2 /title "a2 = a1 &* results"/ format=f5.2 .
* PRINT a3 /title "raw eigen (a2) - percentyl eigen (a2)"/ format=f5.2 .
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

```
* PRINT a4 /title "sample number (%(cnt)s+1) and N of factors (csum(a3 > 0)"/ .
* PRINT pa_sum /title "summary of pa results)"/ format=f5.2 .

* SAVE pa_sum /outfile='%(path)s%(pa_sum)s'.
* SAVE results /outfile=%(path)s%(pa)s'
       / var=root rawdata means percntyl.
END MATRIX.
    """%locals())
# Call DOS command to combine all the txt file outputs, then delete them.
os.system('copy '+path2+'*pa.txt '+path2+pa_all+txt)
os.system('del '+path2+'*pa.txt ')
# Call spss to import the data file and save it as spss.sav file.
spss.Submit("""
GET DATA  /TYPE = TXT
 /FILE = '%(path)s%(pa_all)s%(txt)s'
 /DELCASE = LINE
 /DELIMITERS = " "
 /ARRANGEMENT = DELIMITED
 /FIRSTCASE = 1
 /IMPORTCASE = ALL
 /VARIABLES =
 Sample F3.0
 PA_N F1.0 .
SORT CASES BY
  Sample (A) .
SAVE OUTFILE='%(path)s%(pa_all)s%(sav)s'
 /COMPRESSED.
"""%locals())
END PROGRAM.
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Program for Conducting Velicer's MAP

```
*  Velicer's Minimum Average Partial (MAP) Test.
* Uses Brian O'Connor code http://flash.lakeheadu.ca/~boconno2/nfactors/nfactors.html .

* nbsamp       Number of sample datasets to be read into the parallel analysis macro
DEFAULT (100)
* nbitems      Number of items in the scale DEFAULT (12)
* nbscal       Number of scale points (e.g. 5 or 7 rating points) DEFAULT (7)
* percent      Desired percentile DEFAULT (95)
* path         Location of input data files DEFAULT ('C:\temp\')

BEGIN PROGRAM python.
import spss, os
percent=95                  #Desired percentile DEFAULT (95)
path='C:/Documents and Settings/ajeffrey/My Documents/'  #Location of input data files
DEFAULT ('C:\temp\')
path2='C:\\DOCUME~1\\ajeffrey\\MYDOCU~1\\' #For DOS commands using My Docs
sav='.sav'              # Use for file naming e.g. in GET raw / FILE = command
txt='.txt'
varList=[]              # Used to hold list of variables to carry out analysis on
###########################################################################
# Following lines read the values of nbsamp, smpsize, nbitems and nbscal #
# from the first data set (1.sav). If you want to run the program on     #
# less than nbsamp samples, add the line nbsamp=n (where n is less than  #
# nbsamp) on the line following <dataCursor.close()>.                    #
###########################################################################
spss.Submit("GET FILE='"+path+"1.sav' .")
dataCursor=spss.Cursor()
firstRow=dataCursor.fetchone()
for i in range(spss.GetVariableCount()):
    if spss.GetVariableName(i) == 'nbsamp':
        nbsamp=firstRow[i]
    if spss.GetVariableName(i) == 'smpsize':
        smpsize=firstRow[i]
    elif spss.GetVariableName(i) == 'nbitems':
        nbitems=firstRow[i]
    elif spss.GetVariableName(i) == 'nbscal':
        nbscal=firstRow[i]
dataCursor.close()
# nbsamp = 3    ############### COMMENT OUT THIS LINE TO RUN ALL SAMPLES #########
# print "nbsamp = ", nbsamp
# print "smpsize = ", smpsize
# print "nbitems = ", nbitems
# print "nbscal = ", nbscal

for j in range(spss.GetVariableCount()):
    if spss.GetVariableLabel(j)== 'posItem' or spss.GetVariableLabel(j)=='crlsItem':
        varList.append(spss.GetVariableName(j))
varList=" ".join(varList)
# print varList
spss.Submit("SET mxloops=18000 printback=off width=80 seed = 1953125 .")
# Begin loop for sequentially analysing the multiple data samples.
for cnt in range(nbsamp):
    indx=cnt+1
    fileN=repr(cnt+1)
    spss.Submit("""
GET FILE ='%(path)s%(fileN)s%(sav)s' .
CORRELATION  VARIABLES = %(varList)s / MATRIX OUT ('%(path)scorrdata.sav') .
MATRIX.
MGET /type= corr /file='%(path)scorrdata.sav' .

CALL eigen (cr,eigvect,eigval).
COMPUTE loadings = eigvect * sqrt(mdiag(eigval)).
COMPUTE fm = make(nrow(cr),2,-9999).
COMPUTE fm(1,2) = (mssq(cr)-ncol(cr))/(ncol(cr)*(ncol(cr)-1)).
LOOP #m = 1 to ncol(cr) - 1.
COMPUTE a = loadings(:,1:#m).
COMPUTE partcov = cr - (a * t(a)).
COMPUTE d = mdiag( 1 / (sqrt(diag(partcov))) ).
COMPUTE pr = d * partcov * d.
COMPUTE fm(#m+1,2) = (mssq(pr)-ncol(cr))/(ncol(cr)*(ncol(cr)-1)).
END LOOP.
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

```
* identifying the smallest fm value & its location (= # factors).
COMPUTE minfm = fm(1,2).
COMPUTE nfactors = 0.
LOOP #s = 1 to nrow(fm).
COMPUTE fm(#s,1) = #s -1.
DO IF ( fm(#s,2) < minfm ).
COMPUTE minfm = fm(#s,2).
COMPUTE nfactors = #s - 1.
END IF.
END LOOP.

PRINT /title="Velicer's Minimum Average Partial (MAP) Test:".
PRINT eigval  /title="Eigenvalues" /format "f12.6".
PRINT fm /title="Velicer's Average Squared Correlations"/format "f12.6".
PRINT minfm/title="The smallest average squared correlation is"/format "f12.6".
PRINT nfactors /title="The number of components is".
COMPUTE a5 = {%(indx)s,nfactors} .

WRITE a5 /OUTFILE = '%(path)s%(indx)smap.txt' /FIELD = 1 TO 10 BY 5 .
END MATRIX.
    """%locals())

# Call DOS command to combine all the txt file outputs, then delete them.
os.system('copy '+path2+'*map.txt '+path2+'map_all.txt')
os.system('del '+path2+'*map.txt ')
# Call spss to import the data file and save it as spss.sav file.
spss.Submit("""
GET DATA  /TYPE = TXT
 /FILE = '%(path)smap_all.txt'
 /DELCASE = LINE
 /DELIMITERS = " "
 /ARRANGEMENT = DELIMITED
 /FIRSTCASE = 1
 /IMPORTCASE = ALL
 /VARIABLES =
 Sample F3.0
 MAP_N F1.0 .
SORT CASES BY
  Sample (A) .
SAVE OUTFILE='%(path)smap_all.sav'
 /COMPRESSED.
"""%locals())

END PROGRAM.
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Program for Running PRELIS and LISREL Analyses

```
BEGIN PROGRAM python.
import spss, os

path='C:/Documents and Settings/ajeffrey/My Documents/'
LisrelPath='C:\\Program Files\\lisrel88\\'
path2='C:\\DOCUME~1\\ajeffrey\\MYDOCU~1\\' # Used for sending commands to DOS with
os.system()
sav='.sav'          # Use for file naming e.g. in GET raw / FILE = command
dat='.dat'          # Use for file naming e.g. in WRITE OUTFILE / FILE = command
txt='.txt'

varList=[]          # Used to hold list of variables to carry out analysis on
#############################################################################
# Following lines read the values of nbsamp, smpsize, nbitems and nbscal    #
# from the first data set (1.sav). To run the program on less than          #
# nbsamp samples, add the line nbsamp=n (where n is less than               #
# nbsamp) on the line following <dataCursor.close()>.                       #
#############################################################################
spss.Submit("GET FILE='"+path+"1.sav' .")
dataCursor=spss.Cursor()
firstRow=dataCursor.fetchone()
for i in range(spss.GetVariableCount()):
    if spss.GetVariableName(i) == 'nbsamp':
        nbsamp=firstRow[i]
    if spss.GetVariableName(i) == 'smpsize':
        smpsize=firstRow[i]
    elif spss.GetVariableName(i) == 'nbitems':
        nbitems=firstRow[i]
    elif spss.GetVariableName(i) == 'nbscal':
        nbscal=firstRow[i]
    elif spss.GetVariableName(i) == 'nbrev':
        nbrev=firstRow[i]
dataCursor.close()
# nbsamp=5
# print "nbsamp = ", nbsamp
# print "smpsize = ", smpsize
# print "nbitems = ", nbitems
# print "nbscal = ", nbscal
for j in range(spss.GetVariableCount()):
    if spss.GetVariableLabel(j)== 'posItem' or spss.GetVariableLabel(j)=='crlsItem':
        varList.append(spss.GetVariableName(j))
varList=" ".join(varList)
f=open(path+'labels.txt', 'a')
for i in range(nbsamp):
    f.write(varList+'\n')
f.close()
# print varList
spss.Submit("SET mxloops=18000 printback=off width=80 .")
#############################################################################
# Begin loop which opens each n.sav file, and saves an ascii text file n.dat  #
# keeping only the scale items to be used in LISREL analysis. WRITE creates   #
# ascii data files for use in PRELIS. Text files then copied into a single    #
# file, and original files deleted. Note: labels.txt file created above       #
# contains nbsamp lines, with each line containing the varialbe names.        #
#############################################################################
for cnt in range(nbsamp):
    indx=cnt+1
    fileN=repr(cnt+1)
    fileS='LIS'+repr(cnt+1)
    spss.Submit("""
GET FILE='%(path)s%(fileN)s%(sav)s' .
WRITE OUTFILE='%(path)s%(fileS)s%(dat)s' /%(varList)s .
EXECUTE .
    """%locals())
os.system('copy '+path2+'LIS*.dat '+path2+'All.dat')
os.system('del '+path2+'LIS*.*')

# Create the PRELIS syntax file, for generating polychoric correlation matrices
f=open(path+'Polychoric.pr2', 'a')
f.write('CREATE POLYCHORIC CORRELATION AND ASYMPTOTIC COVARIANCE MATRIX FROM DATA\n')
f.write('DA NI='+repr(int(nbitems))+' NO='+repr(int(smpsize))+'
RP='+repr(int(nbsamp))+'\n')
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

```
f.write('LA='+path2+'labels.txt\nRA='+path2+'ALL.dat\nOR ALL\n')
f.write('OU MA=PM PM='+path2+'ALL.COR SA='+path2+'All.ACM\n')
f.close()

# Run the PRELIS program
os.chdir(LisrelPath)
os.system('Prelis28.exe '+path2+'Polychoric.pr2 '+path2+'Polychoric.out')

# Create the LISREL syntax file for fitting one-factor model
f=open(path+'Lis1factor.ls8', 'a')
f.write('TI CFA USING POLYCHORIC MATRIX AND DWLS - 1 FACTOR MODEL\n')
f.write('DA NI='+repr(int(nbitems))+' NO='+repr(int(smpsize))+' MA=PM
RP='+repr(int(nbsamp))+'\n')
f.write('LA='+path2+'labels.txt\nPM='+path2+'ALL.COR\nAC='+path2+'ALL.ACM\n')
f.write('MO NX='+repr(int(nbitems))+' NK=1 PH=ST LX=FU,FR\n')
f.write('LK\n    Latent\nOU ME=DWLS GF='+path2+'oneFac.GF PV='+path2+'oneFac.PV
SV='+path2+'oneFac.SV\n')
f.close()

# Run the LISREL one-factor program
os.chdir(LisrelPath)
os.system('lisrel88.exe '+path2+'Lis1factor.ls8 '+path2+'Lis1factor.out')


# Create the LISREL syntax file for fitting two-factor model
# Start by creating pattern matrix for Lambda X loadings
LXpos="1 0 "
LXcrls="1 1 "
LX1=int(nbitems-nbrev)*LXpos
LX2=int(nbrev)*LXcrls
f=open(path+'Lis2factor.ls8', 'a')
f.write('TI CFA USING POLYCHORIC MATRIX AND DWLS - 2 FACTOR MODEL\n')
f.write('DA NI='+repr(int(nbitems))+' NO='+repr(int(smpsize))+' MA=PM
RP='+repr(int(nbsamp))+'\n')
f.write('LA='+path2+'labels.txt\nPM='+path2+'ALL.COR\nAC='+path2+'ALL.ACM\n')
f.write('MO NX='+repr(int(nbitems))+' NK=2 PH=DI\n')
f.write('LK\n    Latent Method\n')
f.write('PA LX\n')
f.write(LX1+LX2)
f.write('\nOU ME=DWLS GF='+path2+'twoFac.GF PV='+path2+'twoFac.PV
SV='+path2+'twoFac.SV\n')
f.close()

# Run the LISREL two-factor program
os.chdir(LisrelPath)
os.system('lisrel88.exe '+path2+'Lis2factor.ls8 '+path2+'Lis2factor.out')

END PROGRAM.
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Program for Collating Output of Alpha, EFA and CFA Analyses

```
* Program collates parallel analysis, MAP and alpha calculations into one data file
* and calculates summary stats/frequencies. It then reads LISREL fit indices file (GF)
* and provides summary stats on fit indices for one and two factor models            .
SET RESULTS=LISTING .
BEGIN PROGRAM python.
import spss
path='C:/Documents and Settings/ajeffrey/My Documents/'
spss.Submit("GET FILE='"+path+"1.sav' .")
spss.Submit("LIST VARIABLES=nbsamp TO skew /CASES=1 .")
dataCursor=spss.Cursor()
firstRow=dataCursor.fetchone()
for i in range(spss.GetVariableCount()):
    if spss.GetVariableName(i) == 'nbsamp':
        nbsamp=firstRow[i]
    if spss.GetVariableName(i) == 'smpsize':
        smpsize=firstRow[i]
    elif spss.GetVariableName(i) == 'nbitems':
        nbitems=firstRow[i]
    elif spss.GetVariableName(i) == 'nbscal':
        nbscal=firstRow[i]
    elif spss.GetVariableName(i) == 'nbrev':
        nbrev=firstRow[i]
    elif spss.GetVariableName(i) == 'itemcor':
        itemcor=firstRow[i]
    elif spss.GetVariableName(i) == 'caserev':
        caserev=firstRow[i]
    elif spss.GetVariableName(i) == 'itemrev':
        itemrev=firstRow[i]
    elif spss.GetVariableName(i) == 'skew':
        skew=firstRow[i]
dataCursor.close()
###############################################################################################
###############
# FOLLOWING BLOCK USES OUTPUT COMMAND, WHICH IS ONLY AVAILABLE FROM SPSS 15.0 ON. LEAVE
COMMENTED OUT#
# Following block starts by closing spss output without saving. It then creates a
string called      #
# outputName for use in naming the new output document. The conditionals on the next
few lines are   #
# just to create tidy file names (e.g. all digits taking up same number of spaces) so
they will sort #
# in proper order in file directories (when I use the word macro to concatenate all
word files,      #
# having the files in the proper sequence makes the resulting document easier to
navigate).          #
###############################################################################################
###############
# spss.Submit("OUTPUT CLOSE NAME=ALL .")
# if itemrev < .01: ir='000'
# elif itemrev == 1: ir='100'
# else:  ir='0'+repr(100*itemrev)[0:2]
# if caserev < .01: cr='00'
# elif caserev == .05: cr='05'
# else: cr=repr(100*caserev)[0:2]
# if skew == 2:
#     sk='Hi'
# elif skew == 1:
#     sk='Lo'
# elif skew == 0:
#     sk='Nrml'
#
outputName='i'+repr(nbitems)+'r'+repr(nbrev)+'s'+repr(nbscal)+'c'+repr(10*(round(10*ite
mcor,0)))[0:2] \
#     +'crv'+cr+'irv'+ir+'S'+sk
# spss.Submit("OUTPUT NEW NAME=%(outputName)s ."%locals())
print "%i item scale with %i rating points and %i reverse-coded items." % (nbitems,
nbscal, nbrev)
print "Proportion of careless cases = %g, each careless with %g of reverse-coded
items." % (caserev, itemrev)
print "Target inter-item correlation = %g. %i samples, each with %i cases." % (itemcor,
nbsamp, smpsize)
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

```
print "Target skew = %g (2 = High, 1 = Low, 0 = Normal)." % (skew)
# print "nbsamp = ", nbsamp
# print "smpsize = ", smpsize
# print "nbitems = ", nbitems
# print "nbscal = ", nbscal

spss.Submit("""
GET FILE='%(path)spa_all.sav'.
MATCH FILES /FILE=*
    /FILE = '%(path)smap_all.sav'
        /FILE = '%(path)sorigAlpha.sav'
        /RENAME=(CronbachsAlpha = origAlpha)
        /FILE = '%(path)scrlsAlpha.sav'
        /RENAME=(CronbachsAlpha = crlsAlpha)
        /KEEP=Sample, PA_N, MAP_N, origAlpha, crlsAlpha, NofItems .
SAVE OUTFILE='%(path)ssummary.sav' .
FREQUENCIES
  VARIABLES=PA_N MAP_N
  /ORDER=  ANALYSIS .
DESCRIPTIVES
  VARIABLES=origAlpha crlsAlpha
  /STATISTICS=MEAN STDDEV MIN MAX .
* Calculates min, max, mean, sd of key indices - SB Chi Square, p value,     .
* RMSEA, SRMR, and NNFI. Could look at calculating proportion of cases in    .
* which SB chi sq was non-significant. Or could concatenate files, use       .
* python to group (first 200, second 200) and print output by group.

data list free file = '%(path)soneFac.gf'
/Sample Converg        Proper df ChiSqMF pChiSqMF ChiSqNT pChiSqNT ChiSqSB pChiSqSB
  ChiSqNN pChiSqNN NCP NCPLo NCPHi MinFitFn PopF0 F0Lo F0Hi RMSEA RMSEALo
  RMSEAHi pRMSEA05 ECVI ECVILo ECVIHi ECVISat ECVIInd ChiSqInd AICInd AICMod
  AICSat CAICInd CAICMod CAICSat RMR SRMR GFI AGFI PGFI NFI NNFI PNFI CFI IFI
  RFI CritN .
"""%locals())
END PROGRAM .
FORMATS Sample TO df (F5.0) /ChiSqMF TO CritN (F8.3).
FILE LABEL One Factor CFA of Data.
RECODE  pChiSqSB  (LO THRU .049999=0)  (LO THRU Highest=1)  INTO  p_sig_SB_1F
   / SRMR (LO THRU .059999=0) (LO THRU HI=1) INTO SRMR_06_1F
   / NNFI (LO THRU .949999=0) (LO THRU HI=1) INTO NNFI_95_1F.
VARIABLE LABELS p_sig_SB_1F 'Sig of SB Chi Sq'
   /SRMR_06_1F 'SRMR > .06' /NNFI_95_1F 'NNFI > .95'.
VALUE LABELS Converg 0 'Converged .0005<ChiSq<.9995' 1 'Failed to Converge' 2
'Converged chiSq<.0005 or >.9995 No CI calculated'
  /p_sig_SB_1F 0 'sig p < .05' 1 'Nonsig p > .05'/ SRMR_06_1F 0 '< .06' 1 '>= .06'
  /NNFI_95_1F 0 '< .95' 1 '>= .95'.
VARIABLE LABELS
Sample 'Sample'
Converg 'Convergence (0=OK)'
Proper
df 'Degrees of Freedom'
ChiSqMF 'Minimum Fit Function Chi-Square'
pChiSqMF 'p value (Minimum Fit Function Chi-Square)'
ChiSqNT 'Normal Theory Weighted Least Squares Chi-Square'
pChiSqNT 'p value (Normal Theory WLS Chi-Square)'
ChiSqSB 'Satorra-Bentler Scaled Chi-Square'
pChiSqSB 'p value (SB Scaled Chi-Square)'
ChiSqNN 'Chi-Square Corrected for Non-Normality'
pChiSqNN 'p value (Chi-Sqaure corrected for Non-Normality)'
NCP     'Estimated Non-centrality Parameter (NCP)'
NCPLo '90% Confidence Interval for NCP - Lower bound'
NCPHi '90% Confidence Interval for NCP - Upper bound'
MinFitFn 'Minimum Fit Function Value'
PopF0 'Population Discrepancy Function Value (F0)'
F0Lo '90% Confidence Interval for F0 - Lower bound'
F0Hi '90% Confidence Interval for F0 - Upper bound'
RMSEA 'Root Mean Square Error of Approximation (RMSEA)'
RMSEALo '90% Confidence Interval for RMSEA - Lower bound'
RMSEAHi '90% Confidence Interval for RMSEA - Lower bound'
pRMSEA05 'P-Value for Test of Close Fit (RMSEA < 0.05)'
ECVI 'Expected Cross-Validation Index (ECVI)'
ECVILo '90% Confidence Interval for ECVI - Lower bound'
ECVIHi '90% Confidence Interval for ECVI - Upper bound'
ECVISat 'ECVI for Saturated Model'
ECVIInd 'ECVI for Independence Model'
ChiSqInd 'Chi-Square for Independence Model'
AICInd 'Independence AIC'
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

```
AICMod 'Model AIC'
AICSat 'Saturated AIC'
CAICInd 'Independence CAIC'
CAICMod 'Model CAIC'
CAICSat 'Saturated CAIC'
RMR 'Root Mean Square Residual (RMR)'
SRMR 'Standardized RMR'
GFI 'Goodness of Fit Index (GFI)'
AGFI 'Adjusted Goodness of Fit Index (AGFI)'
PGFI 'Parsimony Goodness of Fit Index (PGFI)'
NFI 'Normed Fit Index (NFI)'
NNFI 'Non-Normed Fit Index (NNFI)'
PNFI 'Parsimony Normed Fit Index (PNFI)'
CFI 'Comparative Fit Index (CFI)'
IFI 'Incremental Fit Index (IFI)'
RFI 'Relative Fit Index (RFI)'
CritN 'Critical N (CN)' .
DESCRIPTIVES
  VARIABLES=ChiSqSB df pChiSqSB RMSEA SRMR NNFI
  /STATISTICS=MEAN STDDEV MIN MAX .
FREQUENCIES
  VARIABLES=p_sig_SB_1F SRMR_06_1F NNFI_95_1F
  /ORDER=  ANALYSIS .
EXECUTE .
BEGIN PROGRAM .
spss.Submit("""
data list free file = '%(path)stwoFac.gf'
/Sample Converg        Proper df ChiSqMF pChiSqMF ChiSqNT pChiSqNT ChiSqSB pChiSqSB
  ChiSqNN pChiSqNN NCP NCPLo NCPHi MinFitFn PopF0 F0Lo F0Hi RMSEA RMSEALo
  RMSEAHi pRMSEA05 ECVI ECVILo ECVIHi ECVISat ECVIInd ChiSqInd AICInd AICMod
  AICSat CAICInd CAICMod CAICSat RMR SRMR GFI AGFI PGFI NFI NNFI PNFI CFI IFI
  RFI CritN .
"""%locals())
END PROGRAM .
FORMATS Sample TO df (F5.0) /ChiSqMF TO CritN (F8.3).
FILE LABEL Two Factor CFA of Data.
RECODE  pChiSqSB  (LO THRU .049999=0)  (LO THRU Highest=1)  INTO  p_sig_SB_2F
   / SRMR (LO THRU .059999=0) (LO THRU HI=1) INTO SRMR_06_2F
   / NNFI (LO THRU .949999=0) (LO THRU HI=1) INTO NNFI_95_2F .
VARIABLE LABELS p_sig_SB_2F 'Sig of SB Chi Sq'
   /SRMR_06_2F 'SRMR > .06' / NNFI_95_2F 'NNFI > .95'.
VALUE LABELS Converg 0 'Converged .0005<ChiSq<.9995' 1 'Failed to Converge' 2
'Converged chiSq<.0005 or >.9995 No CI calculated'
 /p_sig_SB_2F 0 'sig p < .05' 1 'Nonsig p > .05' / SRMR_06_2F 0 '< .06' 1 '>= .06'
  /NNFI_95_2F 0 '< .95' 1 '>= .95'.
VARIABLE LABELS
Sample 'Sample'
Converg 'Convergence (0=OK)'
Proper
df 'Degrees of Freedom'
ChiSqMF 'Minimum Fit Function Chi-Square'
pChiSqMF 'p value (Minimum Fit Function Chi-Square)'
ChiSqNT 'Normal Theory Weighted Least Squares Chi-Square'
pChiSqNT 'p value (Normal Theory WLS Chi-Square)'
ChiSqSB 'Satorra-Bentler Scaled Chi-Square'
pChiSqSB 'p value (SB Scaled Chi-Square)'
ChiSqNN 'Chi-Square Corrected for Non-Normality'
pChiSqNN 'p value (Chi-Sqaure corrected for Non-Normality)'
NCP     'Estimated Non-centrality Parameter (NCP)'
NCPLo '90% Confidence Interval for NCP - Lower bound'
NCPHi '90% Confidence Interval for NCP - Upper bound'
MinFitFn 'Minimum Fit Function Value'
PopF0 'Population Discrepancy Function Value (F0)'
F0Lo '90% Confidence Interval for F0 - Lower bound'
F0Hi '90% Confidence Interval for F0 - Upper bound'
RMSEA 'Root Mean Square Error of Approximation (RMSEA)'
RMSEALo '90% Confidence Interval for RMSEA - Lower bound'
RMSEAHi '90% Confidence Interval for RMSEA - Lower bound'
pRMSEA05 'P-Value for Test of Close Fit (RMSEA < 0.05)'
ECVI 'Expected Cross-Validation Index (ECVI)'
ECVILo '90% Confidence Interval for ECVI - Lower bound'
ECVIHi '90% Confidence Interval for ECVI - Upper bound'
ECVISat 'ECVI for Saturated Model'
ECVIInd 'ECVI for Independence Model'
ChiSqInd 'Chi-Square for Independence Model'
AICInd 'Independence AIC'
AICMod 'Model AIC'
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

```
AICSat 'Saturated AIC'
CAICInd 'Independence CAIC'
CAICMod 'Model CAIC'
CAICSat 'Saturated CAIC'
RMR 'Root Mean Square Residual (RMR)'
SRMR 'Standardized RMR'
GFI 'Goodness of Fit Index (GFI)'
AGFI 'Adjusted Goodness of Fit Index (AGFI)'
PGFI 'Parsimony Goodness of Fit Index (PGFI)'
NFI 'Normed Fit Index (NFI)'
NNFI 'Non-Normed Fit Index (NNFI)'
PNFI 'Parsimony Normed Fit Index (PNFI)'
CFI 'Comparative Fit Index (CFI)'
IFI 'Incremental Fit Index (IFI)'
RFI 'Relative Fit Index (RFI)'
CritN 'Critical N (CN)' .
DESCRIPTIVES
  VARIABLES=ChiSqSB df pChiSqSB RMSEA SRMR NNFI
  /STATISTICS=MEAN STDDEV MIN MAX .
FREQUENCIES
  VARIABLES=p_sig_SB_2F SRMR_06_2F NNFI_95_2F
  /ORDER=  ANALYSIS .
EXECUTE .
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Program for Generating Rensvold and Cheung Jackknife Samples

```
BEGIN PROGRAM .
import spss, os, fileinput

file='Pajo_CSE.sav'     # Insert the name of the CSE data file to be analysed
# varList will select all variables beginning with 'cse' (upper or lower case' and use
# all of these in analysis - make sure only the 12 CSES variables begin with these
letters.
# The LISREL pattern matrix assumes items are entered in sequential order
# CSE01, CSE02, CSE03 etc. If not, then should use spss command to reorder; e.g."
# SAVE OUTFILE='C:/temp/cse.sav'
# /KEEP= CSE01 CSE02 CSE03 etc.
path='C:/Documents and Settings/ajeffrey/My Documents/'
LisrelPath='C:\\Program Files\\lisrel88\\'
path2='C:\\DOCUME~1\\ajeffrey\\MYDOCU~1\\' # Used for sending commands to DOS with
os.system()
sav='.sav'          # Use for file naming e.g. in GET raw / FILE = command
dat='.dat'          # Use for file naming e.g. in WRITE OUTFILE / FILE = command
txt='.txt'

varList=[]          # Used to hold list of variables to carry out analysis on
##############################################################################
# Read file into spss and create variable list                               #
##############################################################################
spss.Submit("GET FILE='"+path+file+"' .")
for j in range(spss.GetVariableCount()):
    if spss.GetVariableName(j)[0:3].lower() == 'cse':
        varList.append(spss.GetVariableName(j))
varList=" ".join(varList)
nbitems=len(varList.split())
if nbitems != 12:
    print "WARNING!! varList contains %s vars; it should have exactly 12 vars" %
nbitems
# The number of jackknife samples to be created equals the number of cases in the
original data:
nsamples=spss.GetCaseCount()
# The sample size for each jackknife samples is 1 less than original sample size:
smpsize=nsamples-1
print "This dataset contains %s cases" % smpsize
# Create file for labels, 1 for each of the jackknife samples.
# Already have model fit for original data, so this anlaysis will only do the
# jackknife samples. (Tricky to automate complete and jackknife samples together
# as the complete sample has one more case than the jackknife samples).
f=open(path+'labels.txt', 'w')
for i in range(spss.GetCaseCount()):
    f.write(varList+'\n')
f.close()
# print varList
spss.Submit("SET mxloops=18000 printback=off width=80 .")
##############################################################################
# First save the complete dataset as ascii text file COMPLETE.dat , keeping only    #
# CSES item variables.                                                       #
##############################################################################
spss.Submit("""
GET FILE='%(path)s%(file)s' .
WRITE OUTFILE='%(path)sCOMPLETE%(dat)s' /%(varList)s .
EXECUTE .
"""%locals())
##############################################################################
# Then create a file jackknife.dat which has all the jackknife samples, concatenated #
# For N cases in original file, this file should have N*(N-1) lines                  #
# The first N-1 lines are the data with case 1 excluded; next N-1 lines              #
# are data for case 2 excluded, etc.                                                 #
##############################################################################

filePath = path+"COMPLETE"+dat
lineCount = len(open(filePath, 'rU').readlines())
print "File %s has %d lines." % (filePath, lineCount)

f=open(path+'jackknife.dat', 'w')
for cnt in range(lineCount):
    for index, case in enumerate(fileinput.input(filePath)):
        if index != cnt:
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

```
            f.write(case)
        if cnt != lineCount-1:
            f.write('\n')
f.close()

# Create the PRELIS syntax file, for generating polychoric correlation matrices
f=open(path+'Polychoric.pr2', 'w')
f.write('CREATE POLYCHORIC CORRELN AND ACMs FROM JACKKNIFE DATA\n')
# Note in following line that the number of samples (the RP= value) is 1 more than
# the number of cases (smpsize).
f.write('DA NI='+repr(int(nbitems))+' NO='+repr(int(smpsize))+'
RP='+repr(int(nsamples))+'\n')
f.write('LA='+path2+'labels.txt\nRA='+path2+'jackknife.dat\nOR ALL\n')
f.write('OU MA=PM PM='+path2+'ALL.COR SA='+path2+'All.ACM\n')
f.close()

# Run the PRELIS program
os.chdir(LisrelPath)
os.system('Prelis28.exe '+path2+'Polychoric.pr2 '+path2+'Polychoric.out')

# Create the LISREL syntax file for fitting one-factor model
f=open(path+'Lis1factor.ls8', 'w')
f.write('TI JACKKNIFE CFAs USING POLYCHORIC MATRIX AND DWLS - 1 FACTOR MODEL\n')
# Note that number of repetitions (nsamples) is 1 more than the number of cases
(smpsize)
f.write('DA NI='+repr(int(nbitems))+' NO='+repr(int(smpsize))+' MA=PM
RP='+repr(int(nsamples))+'\n')
f.write('LA='+path2+'labels.txt\nPM='+path2+'ALL.COR\nAC='+path2+'ALL.ACM\n')
f.write('MO NX='+repr(int(nbitems))+' NK=1 PH=ST LX=FU,FR\n')
f.write('LK\n   Latent\nOU ME=DWLS GF='+path2+'oneFac.GF PV='+path2+'oneFac.PV
SV='+path2+'oneFac.SV\n')
f.close()

# Run the LISREL one-factor program
os.chdir(LisrelPath)
os.system('lisrel88.exe '+path2+'Lis1factor.ls8 '+path2+'Lis1factor.out')


# Create the LISREL syntax file for fitting two-factor model, using a positive item
# factor and a negative item factor (NOT a wording factor plus substantive).
f=open(path+'Lis2factor.ls8', 'w')
f.write('TI CFA USING POLYCHORIC MATRIX AND DWLS - 2 FACTOR MODEL\n')
f.write('DA NI='+repr(int(nbitems))+' NO='+repr(int(smpsize))+' MA=PM
RP='+repr(int(nsamples))+'\n')
f.write('LA='+path2+'labels.txt\nPM='+path2+'ALL.COR\nAC='+path2+'ALL.ACM\n')
f.write('MO NX='+repr(int(nbitems))+' NK=2 PH=ST\n')
f.write('LK\n  PosItems NegItems\n')
f.write('PA LX\n')
# Following pattern matrix assumes items are entered in sequential order
# CSE01, CSE02, CSE03 etc. If not, then should use spss command to reorder; e.g."
# SAVE OUTFILE='C:/temp/cse.sav'
# /KEEP= CSE01 CSE02 CSE03 etc.
# Or could use SE command in LISREL to get them in right order.
f.write('6(1 0 0 1)')
f.write('\nOU ME=DWLS GF='+path2+'twoFac.GF PV='+path2+'twoFac.PV
SV='+path2+'twoFac.SV\n')
f.close()

# Run the LISREL two-factor program
os.chdir(LisrelPath)
os.system('lisrel88.exe '+path2+'Lis2factor.ls8 '+path2+'Lis2factor.out')

# But, a substantive factor (12 items) plus method factor (neg items) might be more
useful.
# By separating out the substantive variance from the method variance, we could look at
effect
# of cases on the method factor (lambda X values for method latent). Rank by these to
see which
# cases most influence the method factor loadings, item by item.

f=open(path+'LisSMfactor.ls8', 'w')
f.write('TI CFA USING POLY MATRIX AND DWLS - SUBST + Method FACTOR MODEL\n')
f.write('DA NI='+repr(int(nbitems))+' NO='+repr(int(smpsize))+' MA=PM
RP='+repr(int(nsamples))+'\n')
f.write('LA='+path2+'labels.txt\nPM='+path2+'ALL.COR\nAC='+path2+'ALL.ACM\n')
f.write('MO NX='+repr(int(nbitems))+' NK=2 PH=DI\n')
f.write('LK\n  CSE Method\n')
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

```
f.write('PA LX\n')
# Following pattern matrix assumes items are entered in sequential order
# CSE01, CSE02, CSE03 etc. If not, then should use spss command to reorder; e.g."
# SAVE OUTFILE='C:/temp/cse.sav'
# /KEEP= CSE01 CSE02 CSE03 etc.
# Or could use SE command in LISREL to get them in right order.
f.write('6(1 0 1 1)')
f.write('\nOU ME=DWLS GF='+path2+'SubMetFac.GF PV='+path2+'SubMetFac.PV
SV='+path2+'SubMetFac.SV\n')
f.close()

# Run the LISREL two-factor program
os.chdir(LisrelPath)
os.system('lisrel88.exe '+path2+'LisSMfactor.ls8 '+path2+'LisSMfactor.out')

END PROGRAM .
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Program for Collating Output of Jackknife Analyses

```
* Program collates fit indices and parameter values for jackknife samples into one data
file
* and calculates summary stats/frequencies.
SET RESULTS=LISTING .
BEGIN PROGRAM .
import spss
path='C:/Documents and Settings/ajeffrey/My Documents/' # Designed to be run under
desktop terminal

spss.Submit("""
* Calculates min, max, mean, sd of key indices - SB Chi Square, p value,     .
* RMSEA, SRMR, and NNFI. Could look at calculating proportion of cases in    .
* which SB chi sq was non-significant. Or could concatenate files, use       .
* python to group (first 200, second 200) and print output by group.

data list free file = '%(path)soneFac.gf'
/Sample Converg        Proper df ChiSqMF pChiSqMF ChiSqNT pChiSqNT ChiSqSB pChiSqSB
  ChiSqNN pChiSqNN NCP NCPLo NCPHi MinFitFn PopF0 F0Lo F0Hi RMSEA RMSEALo
  RMSEAHi pRMSEA05 ECVI ECVILo ECVIHi ECVISat ECVIInd ChiSqInd AICInd AICMod
  AICSat CAICInd CAICMod CAICSat RMR SRMR GFI AGFI PGFI NFI NNFI PNFI CFI IFI
  RFI CritN .
"""%locals())
END PROGRAM .
FORMATS Sample TO df (F5.0) /ChiSqMF TO CritN (F8.3).
FILE LABEL One Factor CFA of Data.
RECODE  pChiSqSB  (LO THRU .049999=0)  (LO THRU Highest=1)  INTO  p_sig_SB_1F
   / SRMR (LO THRU .059999=0) (LO THRU HI=1) INTO SRMR_06_1F
   / NNFI (LO THRU .949999=0) (LO THRU HI=1) INTO NNFI_95_1F.
VARIABLE LABELS p_sig_SB_1F 'Sig of SB Chi Sq'
   /SRMR_06_1F 'SRMR > .06' /NNFI_95_1F 'NNFI > .95'.
VALUE LABELS Converg 0 'Converged .0005<ChiSq<.9995' 1 'Failed to Converge' 2
'Converged chiSq<.0005 or >.9995 No CI calculated'
  /p_sig_SB_1F 0 'sig p < .05' 1 'Nonsig p > .05'/ SRMR_06_1F 0 '< .06' 1 '>= .06'
  /NNFI_95_1F 0 '< .95' 1 '>= .95'.
VARIABLE LABELS
Sample 'Sample'
Converg 'Convergence (0=OK)'
Proper
df 'Degrees of Freedom'
ChiSqMF 'Minimum Fit Function Chi-Square'
pChiSqMF 'p value (Minimum Fit Function Chi-Square)'
ChiSqNT 'Normal Theory Weighted Least Squares Chi-Square'
pChiSqNT 'p value (Normal Theory WLS Chi-Square)'
ChiSqSB 'Satorra-Bentler Scaled Chi-Square'
pChiSqSB 'p value (SB Scaled Chi-Square)'
ChiSqNN 'Chi-Square Corrected for Non-Normality'
pChiSqNN 'p value (Chi-Sqaure corrected for Non-Normality)'
NCP     'Estimated Non-centrality Parameter (NCP)'
NCPLo '90% Confidence Interval for NCP - Lower bound'
NCPHi '90% Confidence Interval for NCP - Upper bound'
MinFitFn 'Minimum Fit Function Value'
PopF0 'Population Discrepancy Function Value (F0)'
F0Lo '90% Confidence Interval for F0 - Lower bound'
F0Hi '90% Confidence Interval for F0 - Upper bound'
RMSEA 'Root Mean Square Error of Approximation (RMSEA)'
RMSEALo '90% Confidence Interval for RMSEA - Lower bound'
RMSEAHi '90% Confidence Interval for RMSEA - Lower bound'
pRMSEA05 'P-Value for Test of Close Fit (RMSEA < 0.05)'
ECVI 'Expected Cross-Validation Index (ECVI)'
ECVILo '90% Confidence Interval for ECVI - Lower bound'
ECVIHi '90% Confidence Interval for ECVI - Upper bound'
ECVISat 'ECVI for Saturated Model'
ECVIInd 'ECVI for Independence Model'
ChiSqInd 'Chi-Square for Independence Model'
AICInd 'Independence AIC'
AICMod 'Model AIC'
AICSat 'Saturated AIC'
CAICInd 'Independence CAIC'
CAICMod 'Model CAIC'
CAICSat 'Saturated CAIC'
RMR 'Root Mean Square Residual (RMR)'
SRMR 'Standardized RMR'
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

```
GFI 'Goodness of Fit Index (GFI)'
AGFI 'Adjusted Goodness of Fit Index (AGFI)'
PGFI 'Parsimony Goodness of Fit Index (PGFI)'
NFI 'Normed Fit Index (NFI)'
NNFI 'Non-Normed Fit Index (NNFI)'
PNFI 'Parsimony Normed Fit Index (PNFI)'
CFI 'Comparative Fit Index (CFI)'
IFI 'Incremental Fit Index (IFI)'
RFI 'Relative Fit Index (RFI)'
CritN 'Critical N (CN)' .
*DESCRIPTIVES
* VARIABLES=ChiSqSB df pChiSqSB RMSEA SRMR NNFI
*  /STATISTICS=MEAN STDDEV MIN MAX .
*FREQUENCIES
*  VARIABLES=p_sig_SB_1F SRMR_06_1F NNFI_95_1F
*  /ORDER=  ANALYSIS .
* EXECUTE .
SAVE OUTFILE='C:/Documents and Settings/ajeffrey/My Documents/OneFactor summary.sav'.
BEGIN PROGRAM .

# Reads in the output file of model parameters for two-factor model
spss.Submit("""
data list free file = '%(path)stwoFac.PV'
/Sample Converg        Proper LX_CSE01 LX_CSE02 LX_CSE03 LX_CSE04 LX_CSE05 LX_CSE06
  LX_CSE07 LX_CSE08 LX_CSE09 LX_CSE10 LX_CSE11 LX_CSE12 PHI TD_CSE01 TD_CSE02
  TD_CSE03 TD_CSE04 TD_CSE05 TD_CSE06 TD_CSE07 TD_CSE08 TD_CSE09 TD_CSE10
  TD_CSE11 TD_CSE12 .
  """%locals())
END PROGRAM .
FORMATS Sample TO Proper (F5.0) /LX_CSE01 TO TD_CSE12 (F8.3).
FILE LABEL Two Factor CFA of Data.
EXECUTE .
SAVE OUTFILE='C:/Documents and Settings/ajeffrey/My Documents/TwoFactor summary.sav'.

BEGIN PROGRAM .
# Reads in the parameter file of model parameters for two-factor (subst and method)
model.
spss.Submit("""
data list free file = '%(path)sSubMetFac.PV'
/Sample Converg Proper
  LXcse_01 LXcse_02 LXmet_02 LXcse_03 LXcse_04 LXmet_04 LXcse_05 LXcse_06 LXmet_06
  LXcse_07 LXcse_08 LXmet_08 LXcse_09 LXcse_10 LXmet_10 LXcse_11 LXcse_12 LXmet_12
  TD_CSE01 TD_CSE02 TD_CSE03 TD_CSE04 TD_CSE05 TD_CSE06 TD_CSE07 TD_CSE08
  TD_CSE09 TD_CSE10 TD_CSE11 TD_CSE12 .
  """%locals())
END PROGRAM .
FORMATS Sample TO Proper (F5.0) /LXcse_01 TO TD_CSE12 (F8.3).
FILE LABEL Sub and Meth Factor CFA of Data.
EXECUTE .
SAVE OUTFILE='C:/Documents and Settings/ajeffrey/My Documents/SubMetFactor
summary.sav'.
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Program to Combine the SPSS Data Files Containing Jackknife Output

```
BEGIN PROGRAM .
import spss
# Use 'copy file name' (right click) to get path to insert in
following 'GET FILE' commands
# if you run from a different directory (following paths are designed
for terminal server).
# Also need to change the filename/path in the last 'SAVE' command .

file='Pajo_CSE'       # Change for different analyses [JK_CSE,
Pajo_CSE, Greet_CSE, CT_CSE]

# use path names in following lines if files are in the terminal
server My Docs directory .
path='C:/Documents and Settings/ajeffrey/My Documents/'
spss.Submit("""
GET FILE='%(path)s%(file)s.sav'.
SORT CASES BY CaseNumber.
DATASET NAME CSEdata.
GET FILE='%(path)sOneFactor summary.sav'.
SORT CASES BY SAMPLE.
RENAME VARIABLES (SAMPLE=CaseNumber).
DATASET NAME OneFactorModel.
GET FILE='%(path)sTwoFactor summary.sav'.
SORT CASES BY SAMPLE.
RENAME VARIABLES (SAMPLE=CaseNumber).
DATASET NAME TwoFactorModel.
GET FILE='%(path)sSubMetFactor summary.sav'.
SORT CASES BY SAMPLE.
RENAME VARIABLES (SAMPLE=CaseNumber).
DATASET NAME SubMetModel.


MATCH FILES
 /FILE='CSEdata'
 /FILE='OneFactorModel'
 /RENAME=(Converg Proper=Converg_1F Proper_1F)
 /FILE='TwoFactorModel'
 /RENAME=(Converg Proper=Converg_2F Proper_2F)
 /FILE='SubMetModel'
 /RENAME=(Converg Proper=Converg_SM Proper_SM)
 /BY CaseNumber.
EXECUTE.
SAVE OUTFILE='%(path)s%(file)s_IC_Analysis.sav'.
"""%locals())
END PROGRAM .
```

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Appendix B

MAP and One-Factor Model Fit Indices for 5-Point Scale

Table B1

Percentage of Samples for which Velicer's MAP indicated a Single Factor Underlying the Interitem Correlation Matrix, for varying levels of Interitem Correlation, Skew, and Levels of Careless Responding (5-Point Response Scale, 12 Items with 6 Reverse-Scored; percentage in each cell based on 200 samples)

Probability of Carelessly Answering a Negatively-keyed Item	Interitem Correlation = .3 Careless Respondents (%)						Interitem Correlation = .7 Careless Respondents (%)						
	0	5	10	15	20	25	0	5	10	15	20	25	
0	100	100	100	100	100	100	100	100	100	100	100	100	
.25	100	100	100	100	100	100	100	100	100	100	99	100	
.50	100	100	100	100	99	97	100	90	59	**17**	**3**	**2**	Skew = -1.12
.75	100	100	96	53	**21**	**5**	100	**30**	**0**	**0**	**0**	**0**	
1.00	100	95	**22**	**1**	**0**	**0**	100	**0**	**0**	**0**	**0**	**0**	
0	100	100	100	100	100	100	100	100	100	100	100	100	
.25	100	100	100	100	100	100	100	100	100	100	96	98	
.50	100	100	100	100	100	99	100	78	**30**	**4**	**0**	**0**	Skew = -.83
.75	100	100	97	57	**25**	**5**	100	**6**	**0**	**0**	**0**	**0**	
1.00	100	97	**22**	**2**	**1**	**0**	100	**0**	**0**	**0**	**0**	**0**	
0	100	100	100	100	100	100	100	100	100	100	100	100	
.25	100	100	100	100	100	100	100	100	100	100	100	100	
.50	100	100	100	100	100	100	100	100	96	88	68	**43**	Skew = 0
.75	100	100	100	100	100	100	100	88	**35**	**5**	**0**	**0**	
1.00	100	100	100	98	88	**50**	100	**46**	**4**	**0**	**0**	**0**	

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Table B2

Percentage of Non-Significant Satorra-Bentler Chi-square Values for One-Factor Model Fitted to Samples with Varying Interitem Correlation, Skew, and Levels of Careless Responding (5-Point Response Scale, 12 Items with 6 Reverse-Scored; percentage in each cell based on 200 samples)

Probability of Carelessly Answering a Negatively-keyed Item	Interitem Correlation = .3 Careless Respondents (%)						Interitem Correlation = .7 Careless Respondents (%)						
	0	5	10	15	20	25	0	5	10	15	20	25	
0	100	100	100	100	100	100	100	100	100	100	100	100	
.25	100	100	100	100	100	100	100	100	97	98	94	95	
.50	100	100	99	86	60	**34**	100	81	**40**	**10**	**2**	**1**	Skew = -1.12
.75	100	97	**45**	**5**	**0**	**0**	100	**16**	**1**	**0**	**0**	**0**	
1.00	100	60	**2**	**0**	**0**	**0**	100	**0**	**0**	**0**	**0**	**0**	
0	100	100	100	100	100	100	100	100	100	100	100	100	
.25	100	100	100	100	100	100	100	100	97	95	86	86	
.50	100	100	99	87	67	**41**	100	67	**21**	**3**	**0**	**0**	Skew = -.83
.75	100	97	51	**9**	**1**	**0**	100	**4**	**0**	**0**	**0**	**0**	
1.00	100	62	**1**	**0**	**0**	**0**	100	**0**	**0**	**0**	**0**	**0**	
0	100	100	100	100	100	100	100	100	100	100	100	100	
.25	100	100	100	100	100	100	100	100	100	99	98	97	
.50	100	100	100	100	100	100	100	98	74	**36**	**15**	**6**	Skew = 0
.75	100	100	100	95	77	**49**	100	53	**3**	**1**	**0**	**0**	
1.00	100	100	92	**40**	**6**	**0**	100	**8**	**0**	**0**	**0**	**0**	

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Table B3

SRMR Values for One-Factor Model Fitted to Samples with Varying Interitem Correlation, Skew, and Levels of Careless Responding (5-Point Response Scale, 12 Items with 6 Reverse-Scored)[a]

Probability of Carelessly Answering a Negatively-keyed Item)	Interitem Correlation = .3 Careless Respondents (%)						Interitem Correlation = .7 Careless Respondents (%)						
	0	5	10	15	20	25	0	5	10	15	20	25	
0	.035	.035	.035	.035	.035	.035	.019	.019	.018	.018	.018	.019	
.25	.035	.037	.039	.042	.043	.045	.019	.027	.033	.034	.039	.039	
.50	.035	.041	.047	.054	**.060**	**.066**	.018	.039	.051	**.063**	**.072**	**.082**	Skew = -1.12
.75	.035	.047	**.061**	**.077**	**.089**	**.103**	.019	.054	**.086**	**.115**	**.139**	**.161**	
1.00	.035	.057	**.084**	**.105**	**.120**	**.132**	.019	**.085**	**.145**	**.205**	**.254**	**.301**	
0	.034	.034	.034	.034	.034	.034	.020	.019	.020	.019	.019	.019	
.25	.034	.036	.038	.040	.042	.043	.020	.030	.035	.039	.042	.044	
.50	.034	.040	.045	.053	.058	**.063**	.019	.044	.058	**.070**	**.083**	**.091**	Skew = -.83
.75	.034	.044	.059	**.074**	**.086**	**.099**	.019	**.071**	**.102**	**.130**	**.151**	**.174**	
1.00	.034	.055	**.081**	**.101**	**.115**	**.131**	.019	**.103**	**.178**	**.235**	**.274**	**.319**	
0	.022	.022	.022	.022	.022	.022	.014	.014	.014	.014	.014	.014	
.25	.022	.024	.026	.028	.029	.031	.014	.018	.02	.023	.024	.027	
.50	.023	.026	.029	.033	.037	.040	.014	.023	.032	.040	.048	.055	Skew = 0
.75	.023	.028	.035	.042	.050	.057	.014	.033	.054	**.077**	**.095**	**.115**	
1.00	.022	.030	.041	.057	**.069**	**.084**	.014	.049	**.091**	**.142**	**.188**	**.232**	

[a] Values > .06 shown in bold

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

Table B4

NNFI Values for One-Factor Model Fitted to Samples with Varying Interitem Correlation, Skew, and Levels of Careless Responding (5-Point Response Scale, 12 Items with 6 Reverse-Scored)[a]

Interitem Correlation = .3

Probability of Carelessly Answering a Negatively-keyed Item	Careless Respondents (%)					
	0	5	10	15	20	25
.0	1.017	1.017	1.017	1.018	1.017	1.017
.25	1.017	1.016	1.015	1.014	1.013	1.013
.50	1.017	1.013	1.007	0.997	0.985	0.971
.75	1.017	1.006	0.983	0.951	**0.919**	**0.878**
1.00	1.017	0.989	**0.944**	**0.906**	**0.886**	**0.873**
0	1.018	1.018	1.018	1.018	1.018	1.018
.25	1.018	1.017	1.017	1.015	1.015	1.014
.50	1.018	1.014	1.008	0.996	0.987	0.976
.75	1.018	1.007	0.984	0.953	**0.920**	**0.878**
1.00	1.018	0.990	**0.943**	**0.906**	**0.880**	**0.859**
0	1.025	1.025	1.025	1.025	1.026	1.025
.25	1.025	1.025	1.025	1.024	1.023	1.022
.50	1.025	1.024	1.022	1.020	1.016	1.013
.75	1.025	1.022	1.016	1.007	0.992	0.973
1.00	1.026	1.019	1.004	0.971	**0.933**	**0.874**

Interitem Correlation = .7

Careless Respondents (%)							
0	5	10	15	20	25		
1.004	1.002	1.002	1.002	1.002	1.002		
1.001	1.001	1.001	1.001	1.000	1.000	Skew	
1.002	0.999	0.996	0.992	0.987	0.981	-1.12	
1.002	0.994	0.980	0.960	**0.936**	**0.909**		
1.002	0.979	**0.940**	**0.882**	**0.818**	**0.741**		
1.003	1.003	1.003	1.003	1.003	1.003		
1.003	1.002	1.001	1.000	0.999	0.998	Skew	
1.003	0.998	0.989	0.979	0.970	0.960	-.83	
1.003	0.985	0.955	**0.916**	**0.875**	**0.824**		
1.003	0.955	**0.876**	**0.778**	**0.683**	**0.577**		
1.003	1.002	1.003	1.002	1.002	1.003		
1.002	1.002	1.002	1.001	1.001	1.001	Skew	
1.002	1.001	0.998	0.996	0.992	0.989	0	
1.002	0.997	0.988	0.974	0.959	**0.939**		
1.002	0.989	0.962	**0.920**	**0.867**	**0.801**		

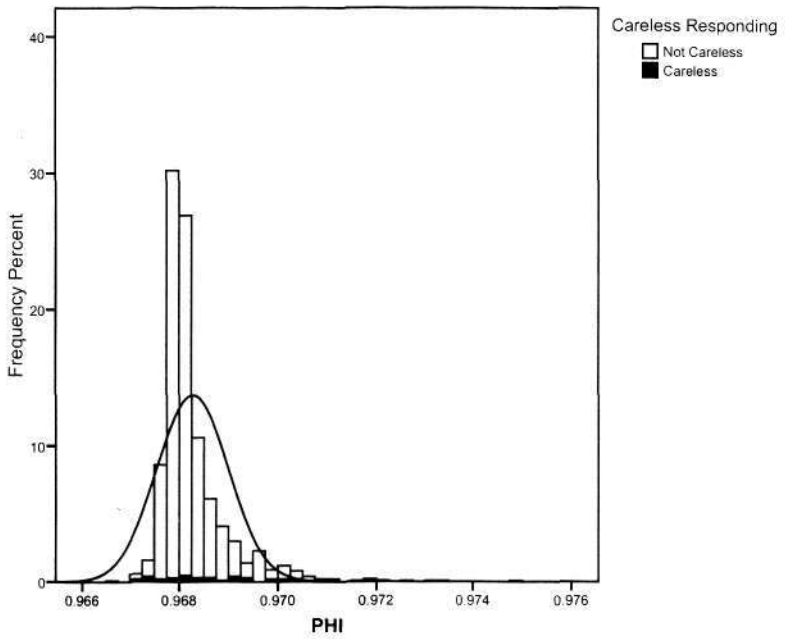
[a] Values < .95 shown in bold

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

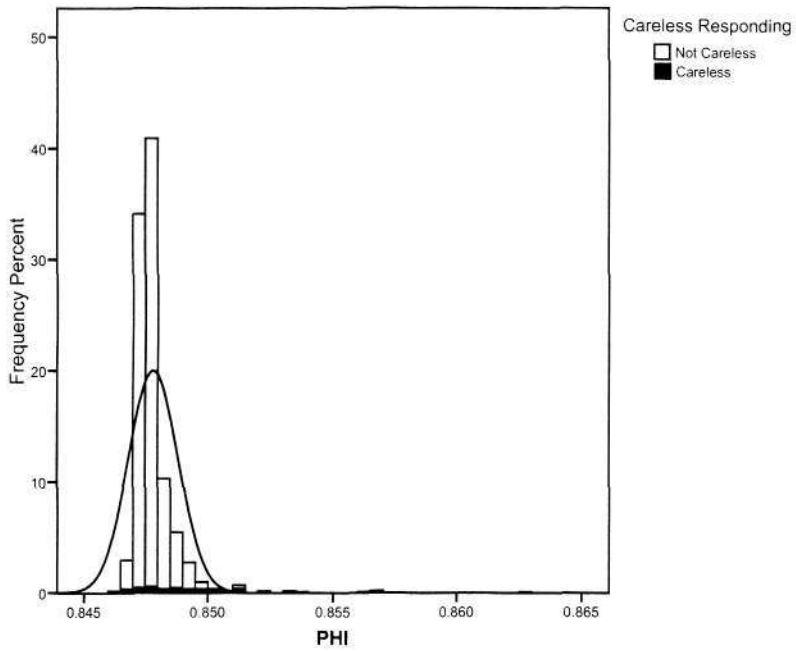
Appendix C

Phi Values for Samples at Each Level of Carelessness

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

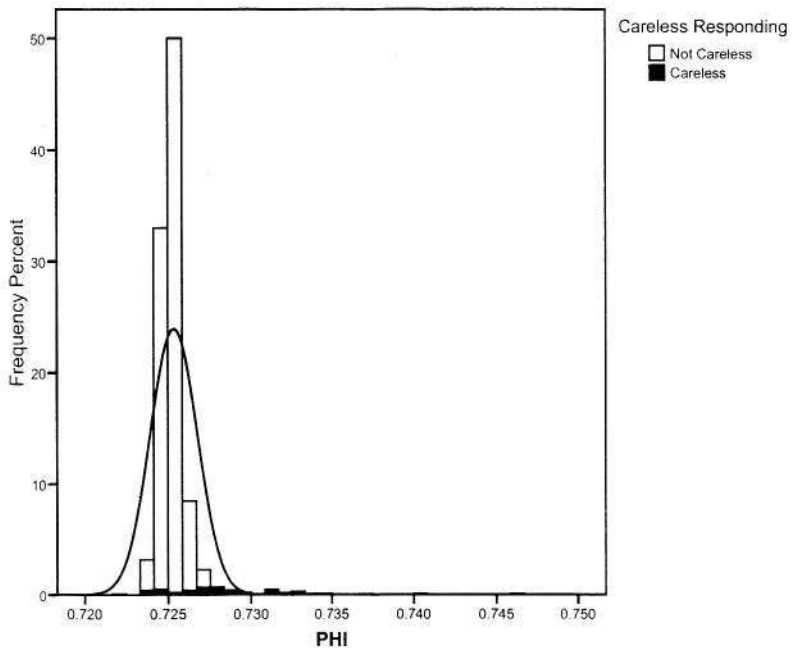


5% Careless respondents; .25 likelihood of answering negative items carelessly

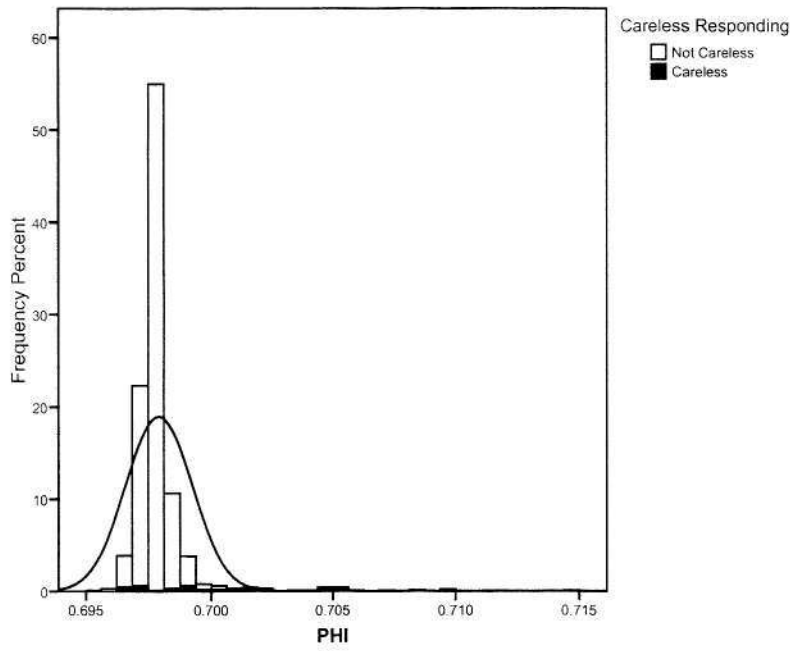


5% Careless respondents; .5 likelihood of answering negative items carelessly

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

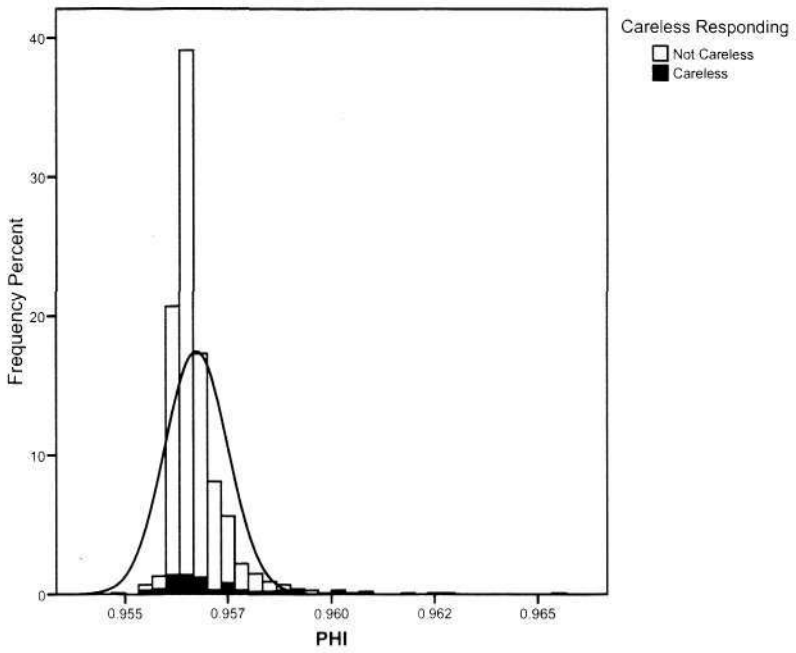


5% Careless respondents; .75 likelihood of answering negative items carelessly

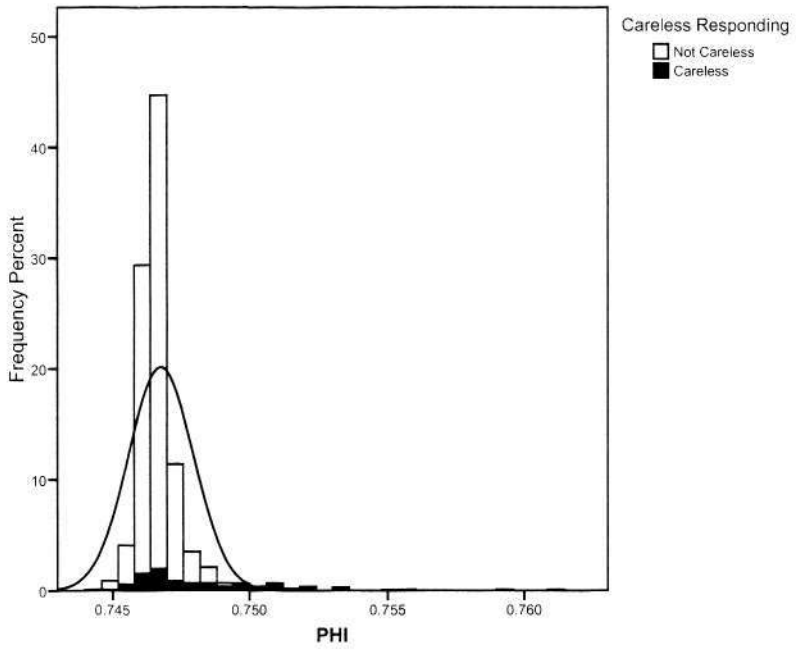


5% Careless respondents; .all negative items answered carelessly

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

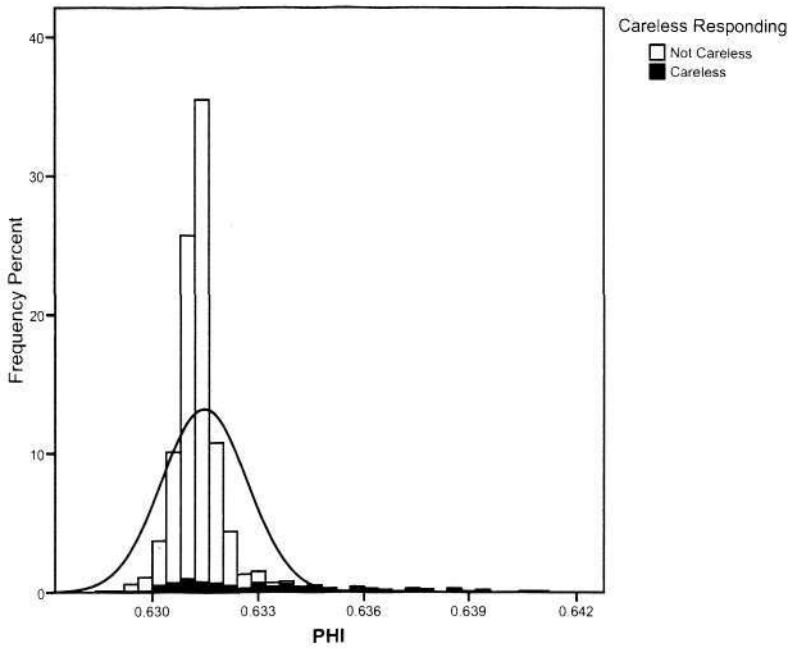


10% Careless respondents; .25 likelihood of answering negative items carelessly

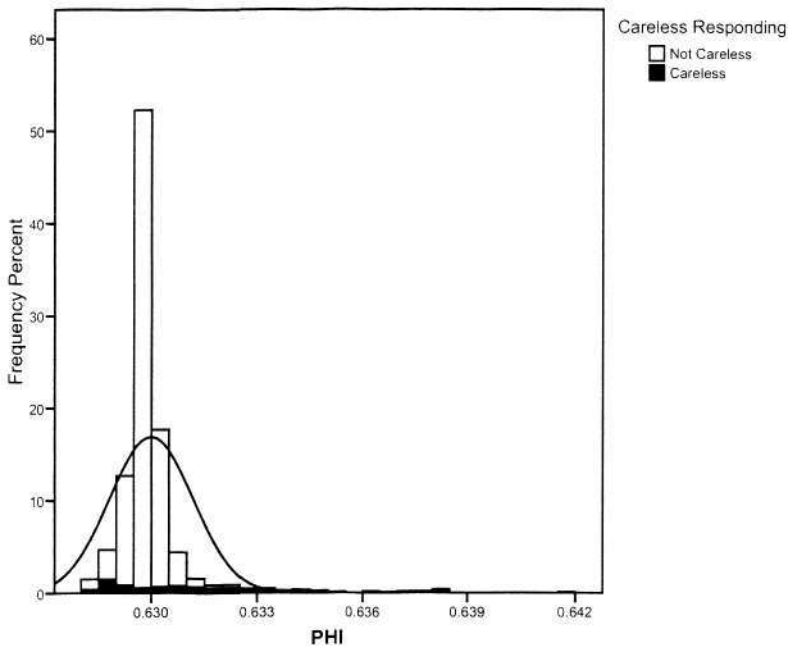


10% Careless respondents; .5 likelihood of answering negative items carelessly

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

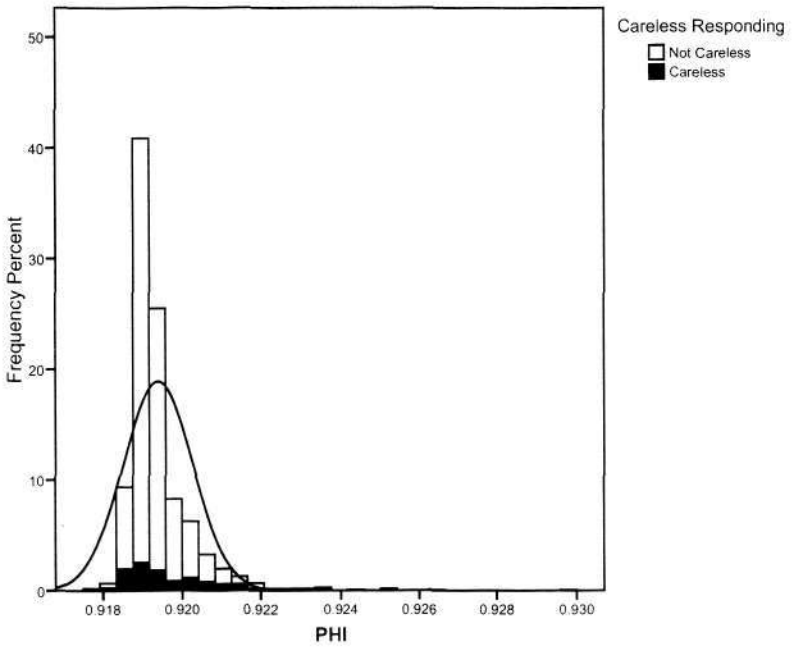


10% Careless respondents; .75 likelihood of answering negative items carelessly

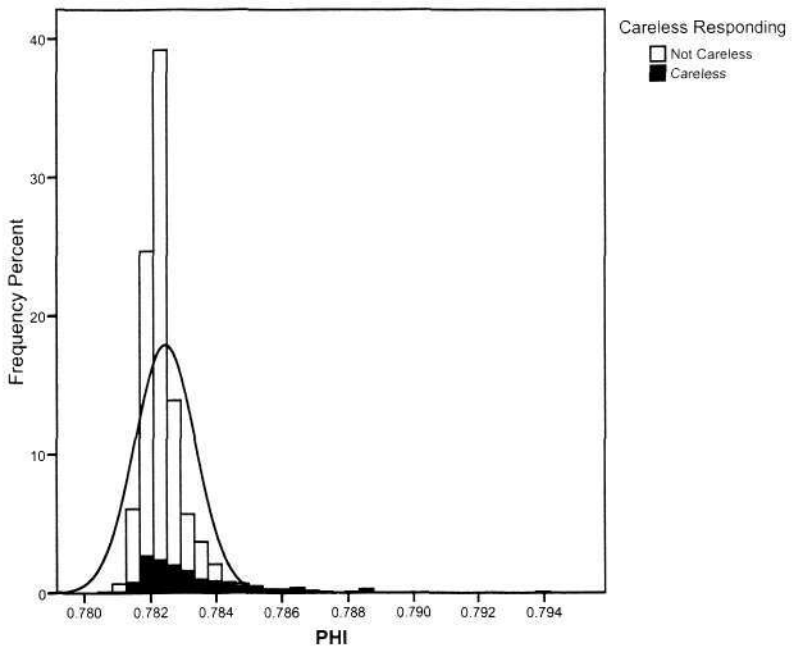


10 Careless respondents; all negative items answered carelessly

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

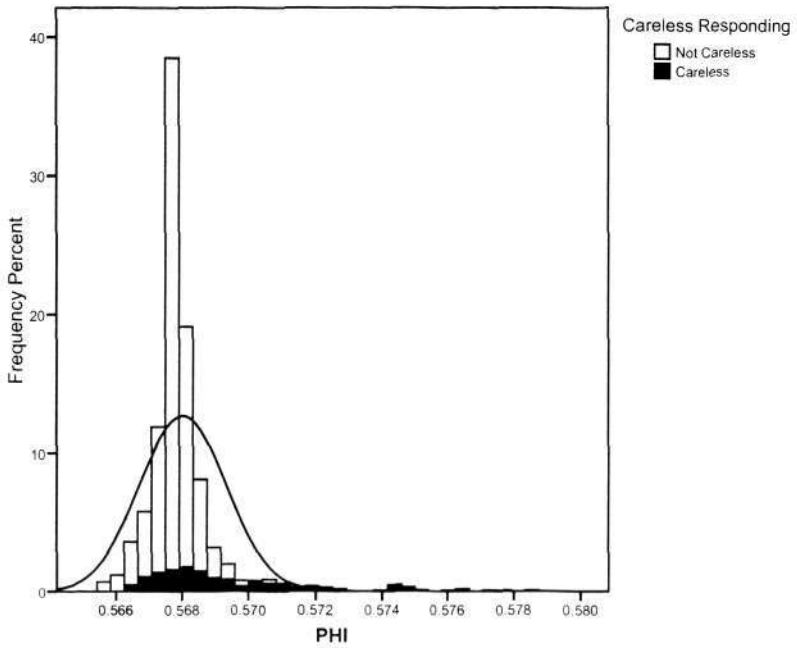


15% Careless respondents; .25 likelihood of answering negative items carelessly

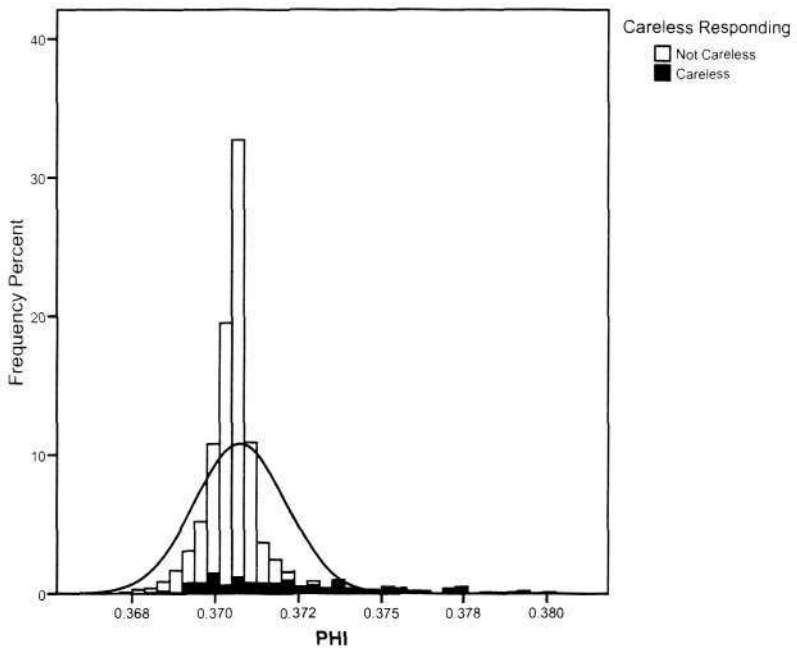


15% Careless respondents; .5 likelihood of answering negative items carelessly

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

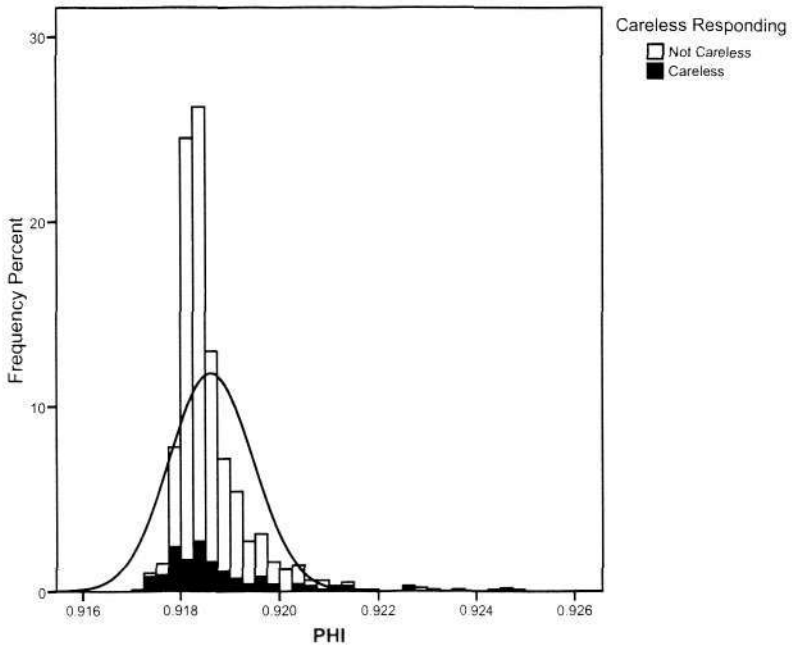


15% Careless respondents; .75 likelihood of answering negative items carelessly

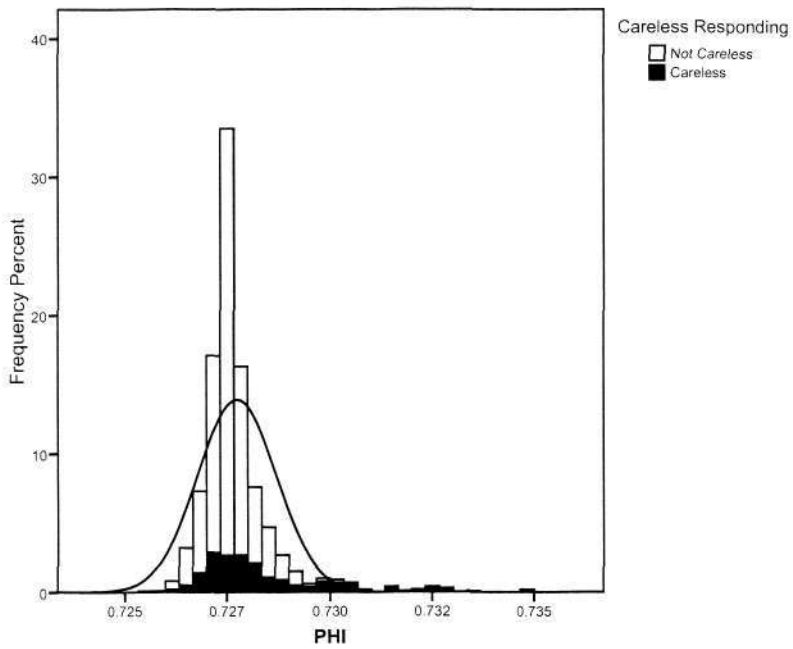


15% Careless respondents; all negative items answered carelessly

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

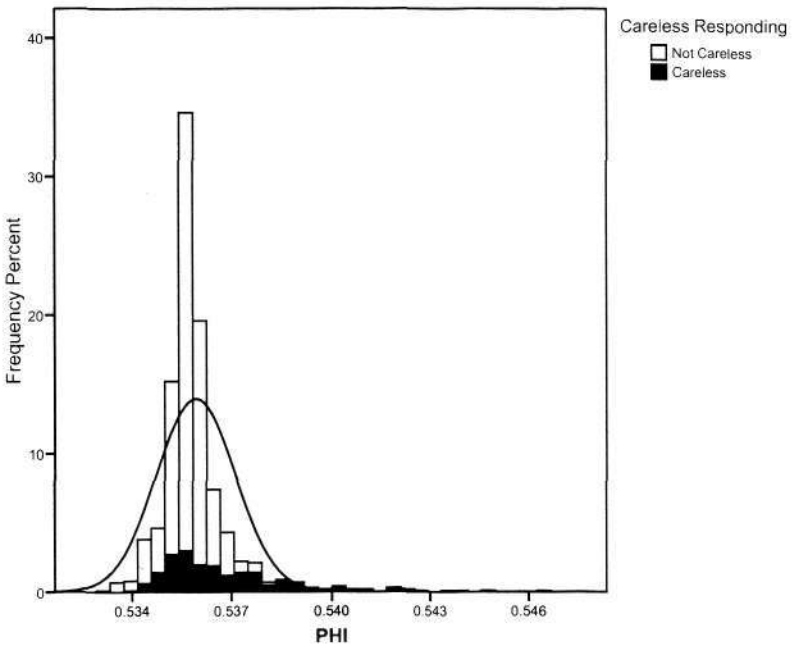


20% Careless respondents; .25 likelihood of answering negative items carelessly

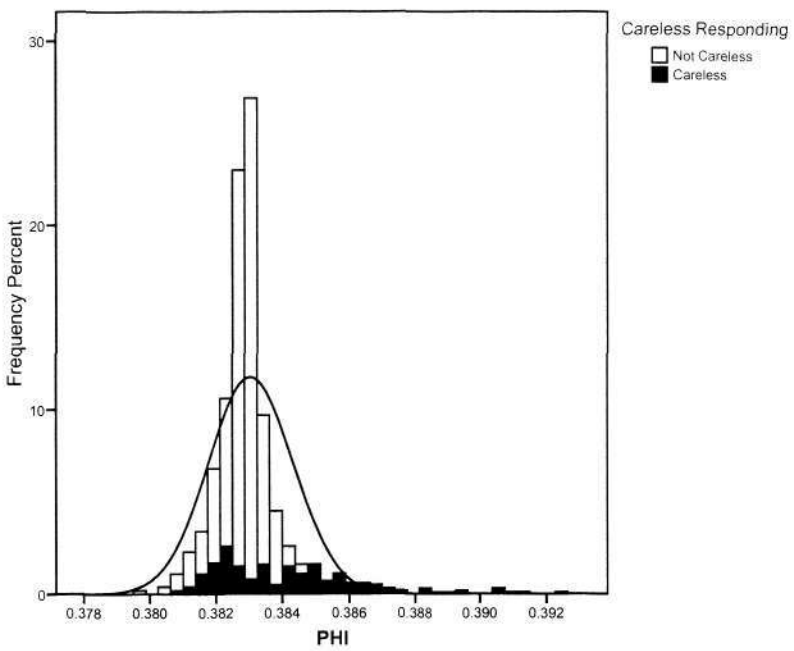


20% Careless respondents; .5 likelihood of answering negative items carelessly

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

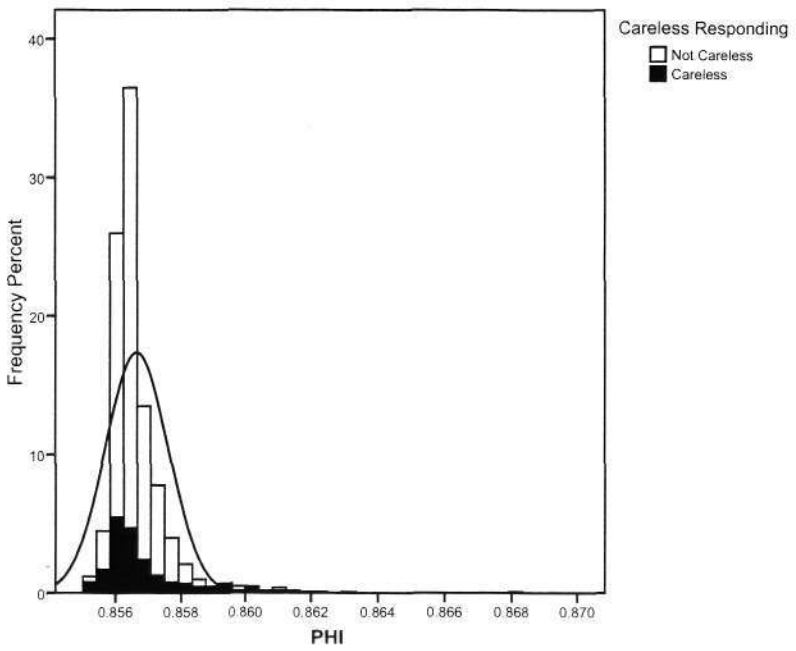


20% Careless respondents; .75 likelihood of answering negative items carelessly

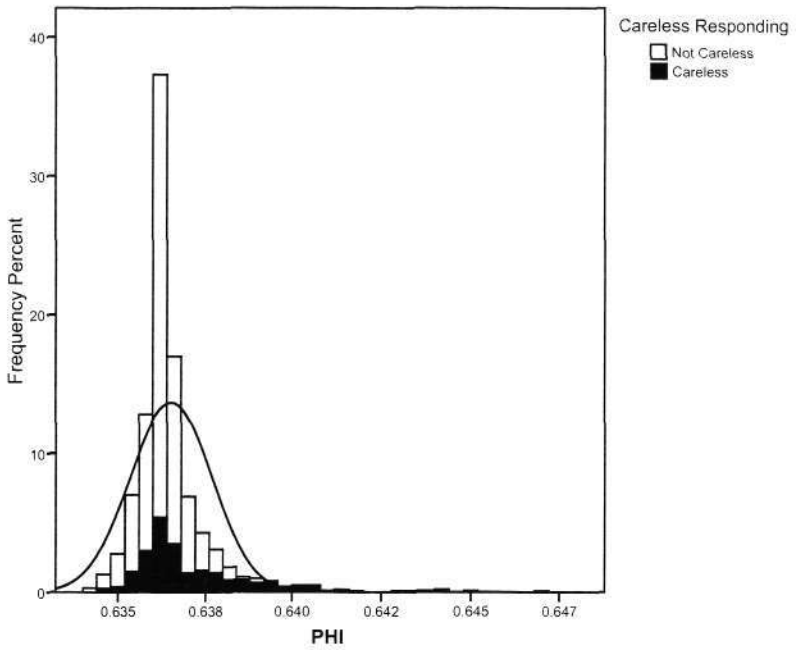


20% Careless respondents; all negative items answered carelessly

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

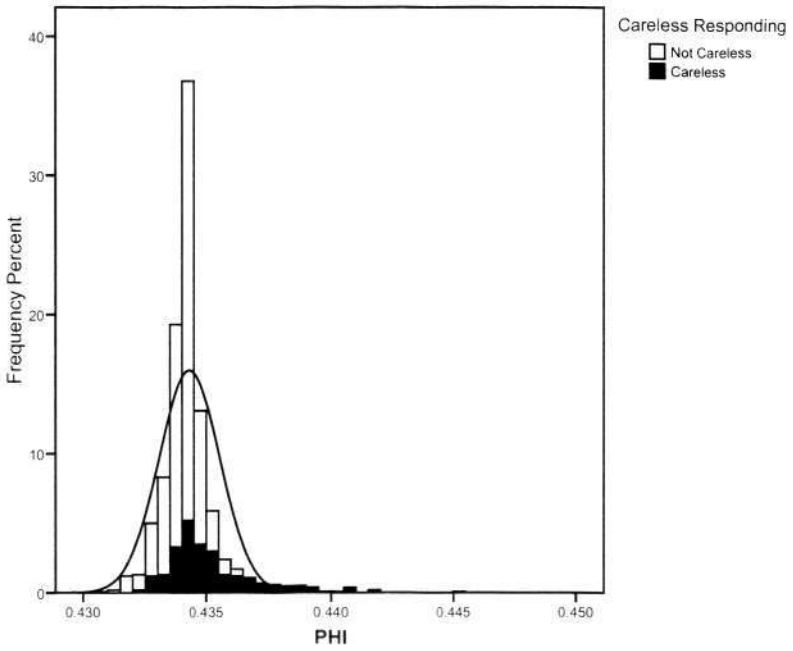


25% Careless respondents; .25 likelihood of answering negative items carelessly

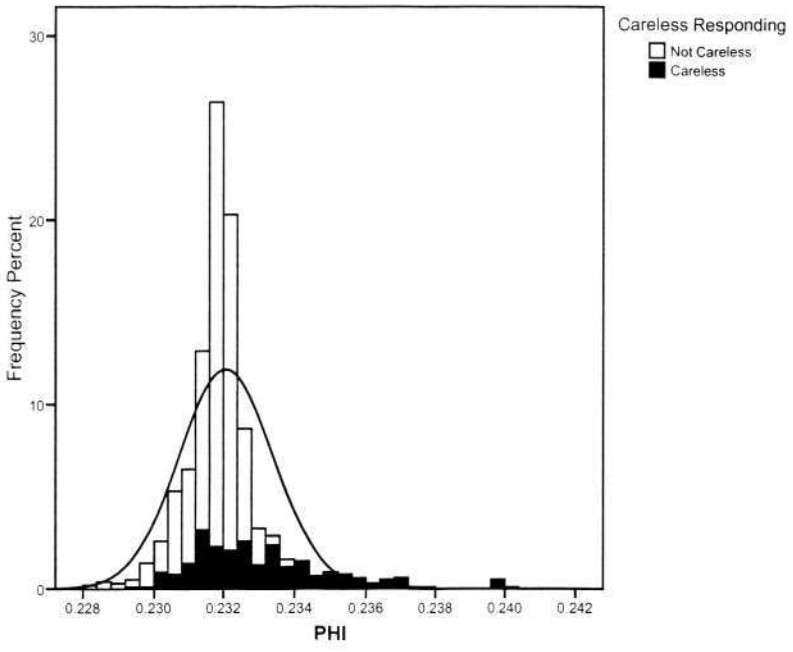


25% Careless respondents; .5 likelihood of answering negative items carelessly

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library

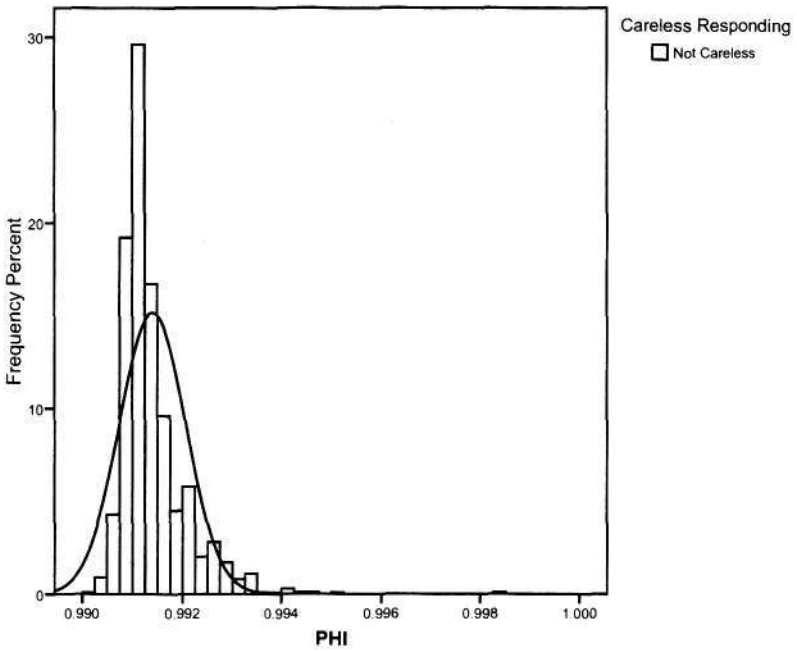


25% Careless respondents; .75 likelihood of answering negative items carelessly



25% Careless respondents; all negative items answered carelessly

ATTENTION: The Singapore Copyright Act applies to the use of this document. Nanyang Technological University Library



No careless respondents