Monte Carlo examination of common approaches for detecting response styles

Fan, Qianqian

2019

Fan, Q. (2019). Monte Carlo examination of common approaches for detecting response styles. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/89310>

<https://doi.org/10.32657/10220/47717>

**MONTE CARLO EXAMINATION OF
COMMON APPROACHES FOR
DETECTING RESPONSE STYLES**

FAN QIANQIAN

SCHOOL OF SOCIAL SCIENCES

2019

**MONTE CARLO EXAMINATION OF COMMON
APPROACHES FOR DETECTING RESPONSE STYLES**

FAN QIANQIAN

SCHOOL OF SOCIAL SCIENCES

A thesis submitted to the Nanyang Technological University

in partial fulfillment of the requirement for the degree of

Doctor of Philosophy

2019

Statement of Originality

I certify that all work submitted for this thesis is my original work. I declare that no other person's work has been used without due acknowledgement. Except where it is clearly stated that I have used some of this material elsewhere, this work has not been presented by me for assessment in any other institution or University. I certify that the data collected for this project are authentic and the investigations were conducted in accordance with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

25/11/2019
.....
Date

Fan Qianqian
.....
Fan Qianqian

Supervisor Declaration Statement

I have reviewed the content of this thesis and to the best of my knowledge, it does not contain plagiarised materials. The presentation style is also consistent with what is expected of the degree awarded. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accordance with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

25 / 1 / 2019

.....
Date

.....
Associate Professor HO Moon-ho

Authorship Attribution Statement

This statement does not apply as my thesis does not contain published material.

25/11/2019
.....
Date

Fan Qianqian
.....
Fan Qianqian

Acknowledgements

I would like to express my deepest gratitude to people who guided and supported me during my graduate study at Nanyang Technological University.

First, I would like to sincerely thank Professor Ringo Ho Moon-Ho, for his excellent guidance, support, and patience throughout my graduate training. It has been a great time working with him. He introduced world of psychological methodology, and also helped me complete a research project on response style. I would never have completed this dissertation without Dr. Ho's help.

I would like to thank my friends and colleagues. In particular, I would like to thank Ms. Neo Hwee Chin, who has always been there for me whenever I am in need. I would like to thank Lim Jie Xin for the help with my study design and data analysis, Wang Zhe for her constant encouragement.

Finally, I would like to thank my parents for their unconditional support throughout all my life.

Table of Contents

Chapter	Page
1. Introduction.....	1
1.1 Introduction to RS.....	1
1.1.1 Types of RS.....	2
1.1.2 Stimuli of RS.....	3
1.1.3 Importance of RS research.....	5
1.2. Overview of approaches for identifying RS	6
1.2.1 Count Procedure.....	6
1.2.2 Representative Indicators Response Style (RIRS).....	8
1.2.3 Representative Indicators Response Style Means and Covariance Structure (RIRSMACS).....	10
1.2.4 Correction of RS by Count Procedure, RIRS, and RIRSMACS	11
1.2.5 Limitations of previous research on Count Procedure, RIRS, and RIRSMACS	12
1.2.6 Item Response Theory (IRT)	14
1.2.6.1 Nominal Response Model (NRM).....	15
1.2.6.2 Partial Credit Model (PCM) - NRM with constraints.....	16
1.2.6.3 Multidimensional Nominal Response Model (MNRM)	17
1.2.6.3.1 MNRM with constraints to model RS	18
1.2.6.3.2 Previous research on MNRM and its limitations	23
2. Aims of Study and Study Design.....	28
2.1 Aim of study	28
2.2 Method	29
2.2.1 Study design.....	29
2.2.2 Independent variables	30

2.2.3 Analysis software.....	32
2.2.4 Data generation	32
3. Results and Discussion of Count Procedure	39
3.1 Analysis models	39
3.2 Evaluation criteria.....	39
3.3 Results of Count Procedure	39
3.3.1 Results of Count Procedure – NO RS condition.....	39
3.3.2 Results of Count Procedure – ARS condition.....	41
3.3.3 Results of Count Procedure – ERS condition.....	43
3.3.4 Results of Count Procedure – MRS condition.....	45
3.4 Discussion of Count Procedure.....	47
4. Results and Discussion of RIRS and RIRSMACS	50
4.1 Analysis models	50
4.1.1 Analysis models for RIRS approach.....	50
4.1.2 Analysis models for RIRSMACS approach	54
4.2 Evaluation criteria.....	58
4.2.1 Model fit.....	58
4.2.2 Parameter recovery	59
4.3 Results of RIRS.....	59
4.3.1 Results of RIRS – NO RS condition.....	59
4.3.2 Results of RIRS – ARS condition.....	62
4.3.3 Results of RIRS – ERS condition	65
4.3.4 Results of RIRS – MRS condition.....	68
4.4 Discussion of RIRS	71
4.5 Results of RIRSMACS	74
4.5.1 Results of RIRSMACS – NO RS condition	74
4.5.2 Results of RIRSMACS – ARS condition	77

4.5.3 Results of RIRSMACS – ERS condition.....	81
4.5.4 Results of RIRSMACS – MRS condition.....	84
4.6 Discussion of RIRSMACS	88
5. Results and Discussion of MNRM	91
5.1 Analysis models	91
5.2 Model identification.....	93
5.3 Evaluation criteria.....	96
5.4 Results of MNRM.....	97
5.4.1 Results of MNRM – NO RS condition.....	97
5.4.2 Results of MNRM – ARS condition.....	101
5.4.3 Results of MNRM – ERS condition	105
5.4.4 Results of MNRM – MRS condition	109
5.5 Discussion of MNRM	113
6. General Discussion	118
6.1 Summary of present study	118
6.2 Limitations of present study.....	123
6.3 Future directions	126
References.....	128

List of Tables

Table 1. Sample dataset for assessment of RS using Count Procedure	8
Table 2. Slope parameters in MNRM (recoded) for content factors and RS factors.....	34
Table 3. Category probabilities (recoded) for individual items for NO RS (Cho (2013)).....	36
Table 4. Intercept parameters (recoded) for individual items.....	37
Table 5. Mean and standard deviation (SD) of RS scores calculated by counting in NO RS condition.....	41
Table 6. Mean and standard deviation (SD) of RS scores calculated by Count Procedure in ARS condition	42
Table 7. Correlations between true and estimated ARS scores in Count Procedure	43
Table 8. Mean and standard deviation (SD) of RS scores calculated by Count Procedure in ERS condition.....	44
Table 9. Correlations between true and estimated ERS scores in Count Procedure	45
Table 10. Mean and standard deviation (SD) of RS scores calculated by Count Procedure in MRS condition.....	46
Table 11. Correlations between true and estimated MRS scores by Count Procedure	46
Table 12. Type I error rate in rejection of true model NO RS when compared to ARS, ERS, and MRS in RIRS	61
Table 13. The number of replications that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (no RS) in RIRS.....	61
Table 14. Type II error rate in failure to reject wrong model- NO RS when	

compared to true model – ARS in RIRS.....	64
Table 15. The number of replications that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (ARS) in RIRS	64
Table 16. Type II error rate in failure to reject wrong model- NO RS when compared to true model – ERS in RIRS	67
Table 17. The number of replications that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (ERS) in RIRS	67
Table 18. Type II error rate in failure to reject wrong model- NO RS when compared to true model –MRS in RIRS	69
Table 19. The number of replications that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (MRS) in RIRS	70
Table 20. Type I error rate in rejection of true model NO RS when compared to ARS, ERS, and MRS in RIRSMACS	76
Table 21. The number of replications (out of total number of convergence) that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (NO RS) in RIRSMACS	76
Table 22. Type II error rate in failure to reject wrong model- NO RS when compared to true model – ARS in RIRSMACS	79
Table 23. The number of replications (out of total number of convergence) that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (ARS) in RIRSMACS.....	79
Table 24. Type II error rate in failure to reject wrong model- NO RS when compared to true model – ERS in RIRSMACS.....	82
Table 25. The number of replications (out of total number of convergence) that	

one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (ERS) in RIRSMACS	83
Table 26. Type II error rate in failure to reject wrong model- NO RS when compared to true model – MRS in RIRSMACS.....	86
Table 27. The number of replications (out of total number of convergence) that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (MRS) in RIRSMACS	86
Table 28. Type I error rate in rejection of true model NO RS when compared to ARS, ERS, and MRS in MNRM	98
Table 29. The number of replications that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (NO RS) in MNRM	99
Table 30. Means of squared difference in intercept parameter among NO RS (true model), ARS, ERS, and MRS models.....	100
Table 31. Type II error rate in failure to reject wrong model- NO RS when compared to true model – ARS in MNRM.....	102
Table 32. The number of replications that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (ARS) in MNRM	103
Table 33. Means of squared difference in intercept parameter among NO RS, ARS (true model), ERS, and MRS models.....	104
Table 34. Type II error rate in failure to reject wrong model- NO RS when compared to true model – ERS in MNRM	106
Table 35. The number of replications that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (ERS) in MNRM	106
Table 36. Mean of squared difference in intercept parameter among NO RS,	

ARS, ERS (true model), and MRS models.....	108
Table 37. Type II error rate in failure to reject wrong model- NO RS when compared to true model – MRS in MNRM	110
Table 38. The number of replications that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (MRS) in MNRM	111
Table 39. Mean of squared difference in intercept parameter among NO RS, ARS, ERS, and MRS (true model) models.....	112
Table 40. Summary of approaches identifying RS in the present study	118

List of Figures

Figure 1. One-factor CFA model with four items a1, a2, a3, and a4.....	7
Figure 2. RIRS model	10
Figure 3. RIRSMACS model.....	11
Figure 4. ICC of one content factor NRM on a five-point scale item	19
Figure 5. ICC of MNRM with one content factor and ERS	20
Figure 6. ICC of MNRM with one content factor and ARS	21
Figure 7. ICC of MNRM with one content factor and MRS	22
Figure 8. ICC of MNRM with one content factor and RS of different values ..	26
Figure 9. Data generation model.....	33
Figure 10. RIRS model for NO RS, ARS, ERS, and MRS.....	51
Figure 11. RIRSMACS model for NO RS, ARS, ERS, and MRS	55
Figure 12. MNRM model for NO RS, ARS, ERS, and MRS.....	91
Figure 13. Squared difference between true and estimated intercept term in ARS.....	94
Figure 14. Squared difference between true and estimated intercept term in ERS	95
Figure 15. Squared difference between true and estimated intercept term in MRS	96
Figure 16. ICC Under 2D-MNRM at Three Levels of Response Style.....	125

Abstract

Response styles (RS) are a respondent's tendency to respond to questions in certain ways regardless of the content. RS received increasing interest, as it exerted negative impact on assessment of true scores by inflating or deflating observed values. A variety of approaches were proposed to identify various RS, including: Count Procedure, Representative Indicators Response Style (RIRS), Representative Indicators Response Style Means and Covariance Structure (RIRSMACS), and Multidimensional Nominal Response Model (MNRM), etc. However, there is lack of systematic simulation studies to examine performance of Count Procedure, RIRS, and RIRSMACS; examination of MNRM in prior simulation studies is not without limitations. Therefore, the present study explored performance of RS approaches under various simulation conditions, including different sample size (300 vs. 800), test length (10 vs. 20), and RS (No RS, Acquiescent RS, Extreme RS, vs. Mid-point RS). Results suggested that across 4 approaches explored in the present study, it was easier to detect MRS than ARS and ERS. Among 4 RS approaches, MNRM performed better than other approaches in picking out the correct RS. Results of simulation studies were discussed and recommendations of employing these approaches to identify RS were provided.

Chapter 1

Introduction

1.1 Introduction to RS

In psychological research, self-report questionnaires have been widely used for data collection. A common way of measuring individuals' attitudes and behaviors in questionnaires is using rating scales (Saal, Downey, & Lahey, 1980). In a rating scale, a set of items attempt to quantify individual differences in a latent trait, which is measured by how respondents choose the response category of these items in the scale (Andrich, 1996). It is assumed that results from a scale represent the true characteristics of individuals, regardless different formats of scales, number of response categories, item wording, and so forth. That is, all of the respondents are supposed to understand and interpret each item in the same way, so their responses would not be influenced by other factors irrelevant to the intended-to-be-measured traits. However, researchers argued that ratings given by respondents may not reflect their true opinions (Roberts, 1994).

According to classic test theory, a test score is the sum of true score and measurement error (Bechger, Maris, Verstralen, & Béguin, 2003). Correspondingly, a response variance can be decomposed into true and error variances (Smith, 2011). The error variances in general consist of systematic and random errors. In prior research, response bias refers to systematic errors, and it might affect estimate of intended-to-be-measured traits (Mellenbergh, 1996). It was proposed that systematic errors may refer to response sets and response styles (Clarke, 2001). A response set is a pattern of responses that is not independent of the content of the response scale (Rorer, 1965). For

example, socially desirable responding (SDR) refers to tendency to give positive self-descriptions (Paulhus, 2002). Presence of SDR depend on the item content. Response styles (RS) are a respondent's tendency to respond to questions in certain ways regardless of the content (van Vaerenbergh & Thomas, 2012). Some previous studies used response sets and RS interchangeably. In the present dissertation, response sets and RS follow the definitions above. The present study focuses on RS which is independent of item content.

1.1.1 Types of RS

There are several types of RS in survey scales. The most widely observed and investigated RS are acquiescence response style (ARS) and extreme response style (ERS). ARS was defined as the tendency to agree with items regardless of content, so the highest response categories are endorsed more often than others (Ray, 1983). Extreme response style (ERS) refers to the tendency to use the highest and lowest response categories of a rating scale (Greenleaf, 1992). In addition to ARS and ERS, there are other types of RS, such as disacquiescence response style (DARS), mid-point response style (MRS), and mild response style (MLRS). DARS refers to the tendency to disagree with items regardless of content, so the lowest response categories are chosen more often (Couch, & Keniston, 1960). MRS refers to the tendency to use the middle response category of a rating scale, regardless of the content (Baumgartner & Steenkamp, 2001). MLRS was regarded as the counterpart of ERS, referring to the tendency to avoid the highest and lowest response categories of a rating scale (Hurley, 1998). Among the five commonly observed

RS, the present dissertation will focus on ARS, ERS, and MRS. Because DARS is the opposite response style to ARS, and MLRS could be regarded as non ERS response style, examining ARS and ERS would provide implications to understand DARS and MLRS in empirical research.

1.1.2 Stimuli of RS

As introduced above, RS is the respondent's systematic tendency to respond to a range of survey items on a different basis from what the items are designed to measure (Paulhus, 1991). Previous researchers have proposed a variety of factors that could affect RS. These factors can be classified into external circumstances versus internal dispositions of respondents (Weijters, 2006). External circumstances include layout or format of questions, modes of data collection, language of questionnaires, cognitive load, experience of interviewers, and so on. For example, Kieruj and Moors (2010) found that MRS emerged when a large number of 9 or 10 response categories offered, relative to a smaller number of response categories. They also found that endpoint-labelled scale formats are more likely to elicit ERS compared to a fully-labelled scale. In addition, modes of data collection was found to exert different effects on RS. For example, among paper-and-pencil questionnaires, telephone interviews, and online questionnaires, telephone interviews led to lower MRS and higher ARS, online data collection lowered DRS and ERS (Weijters, Schillewaert, & Geuens, 2008). Furthermore, cross-cultural studies have found that answering a questionnaire in a second language would lead to lower levels of ARS and ERS but higher levels of MRS and RR than in a native language (Harzing, 2006). Respondents tend to use the scale's mid-point if the scale is worded in their

second language. It was also suggested that questionnaires requiring high cognitive load and less experienced interviewers are more likely to lead to RS (Knowles & Condon, 1999; Olson & Bilgen, 2011).

In addition to external stimuli of RS, the other major source of RS is characteristics and personality of respondents themselves. For example, a series of studies on education suggested that education was negatively related to presence of RS (Meisenberg & Williams, 2008). It was further found that education was positively correlated to IQ, and IQ was negatively correlated to ERS. Therefore, without information of IQ, positive relationships between education and ERS could be considered as a preliminary evidence for presence of ERS (Matarazzo & Herman, 1984). Previous research also found relationships between respondents' personality and RS. For instance, prior research has found that ERS is positively correlated to intolerance of ambiguity, simply thinking and decisiveness (Naemi, Beal, & Payne, 2009), as well as extraversion and conscientiousness measured by the Big Five (Austin, Deary, & Egan, 2006). Furthermore, many studies have investigated the relationships between RS and cultural differences. Even though the conclusions are not consistent regarding relationship between RS and individualism, researchers agreed that there is significant relationship between individualism and ARS, ERS, and MRS (Chen, Lee, & Stevenson, 1995; van Herk, Poortinga, & Verhallen, 2004). In addition, other demographic characteristics, such as race, age, gender, and SES also contributed to RS (e.g., Billiet & McClendon, 2000; Moors, 2008).

1.1.3 Importance of RS research

Researchers showed increasing interests to understand effects of RS on estimation of intended-to-be-measured traits. It was found that RS exerted negative impact on assessment of true scores by inflating or deflating observed values. On one hand, RS affects measurement of means and variances. For instance, one study showed that women are more likely to endorse the highest and lowest response categories than men (Moors, 2012). Therefore, observed gender differences in personal traits may not be present if effects of RS are controlled, or vice versa. Reynolds and Smith (2010) found that RS partially accounted for cultural differences in service quality. For example, Chinese consumers had the lowest desired expectations before RS was taken into consideration. However, Africans scored the lowest after the effects of RS were accounted for. On the other hand, RS affects correlations between variables. Many statistical methods, such as regression analysis, factor analysis, and structural equation modelling, which are based on relationships between variables, would generate inaccurate results and conclusions if RS is not correctly detected and appropriately controlled. Hence, spurious correlations would be observed between personal traits because of similar RS while answering some of the items measuring these traits instead of true correlations between these traits. A previous study found that correlations between variables may be inflated or deflated due to RS (Baumgartner & Steenkamp, 2001). Therefore, in order to obtain accurate results and meaningful research implications regarding traits of interest, it is important to identify and remedy the effects of RS.

1.2 Overview of approaches for identifying RS

To detect RS and control the impact of RS on intended-to-be-measured traits or relationships between variables, several methods have been proposed in the literature. The present dissertation focuses on Count Procedure and item response theory model (IRT; e.g., De Jong, Steenkamp, Fox, & Baumgartner, 2008; Bolt & Newton, 2011). Count procedure can be divided into traditional count approach (e.g., Bachman & O'Malley, 1984; de Beuckelaer, Weijters, & Rutten, 2010) and modified approaches, including Representative Indicators for Response Styles (RIRS; Baumgartner & Steenkamp; 2001) and Representative Indicators Response Style Means and Covariance Structure (RIRSMACS; Weijters, Geuens, & Schillewaert, 2008). IRT approach has received the most research interest as it allows for individual differences in RS and it is possible to detect as well as correct effects of RS on latent traits of interest. One of IRT models - Multidimensional Nominal Response Model (MNRM) are investigated in present dissertation.

1.2.1 Count Procedure

Count Procedure is a traditional method to detect RS. The procedure is to count the number of agreements, disagreements, extreme responses, and/or mid-point response on substantive measures across an entire questionnaire, and then to calculate the proportion of responses associated with a RS relative to number of all responses (Bachman & O'Malley, 1984).

Take 5-response-category items for example:

$$ARS = [f(4)*1 + f(5)*1]/(j*n) \quad (1)$$

$$ERS = [f(1)*1 + f(5)*1]/(j*n) \quad (2)$$

$$MRS = [f(3)]/(j*n) \quad (3)$$

where $f(k)$ refers to the frequency of response option k , n refers to number of respondents, and j refers to number of items.

Figure 1 presented Count Procedure in a SEM framework. The advantage of Count Procedure is easy to use, and it does not require additional items to calculate RS. Besides, all of the RS could be obtained by Count Procedure, including ARS, DARS, ERS, MRS, NARS, etc. However, it was argued that this approach was not sensitive to latent traits of respondents and psychometric characteristics of survey items. That is, if a respondent is high on the intended-to-be-measured latent trait, he may also score high on ARS in that this respondent is more likely to endorse agreement (Bolt & Johnson, 2009). In order to accurately estimate RS, Count Procedure works better when items are heterogeneous, that is, items are not significantly correlated to each other. This requirement could not be always met because items of a scale measure the same substantive trait or similar concept, so it's not always possible to obtain heterogeneous items within one scale or one set of questionnaires to count RS.

Figure 1. One-factor CFA model with four items a1, a2, a3, and a4

Here below in Table 1 is an example dataset for demonstration of using Count Procedure to calculate RS. It is a four-item ($j = 4$) questionnaire on a five-point Likert scale from 1 to 5 ($k = 5$), and the number of respondents is six ($n = 6$).

Table 1. Sample dataset for assessment of RS using Count Procedure

Item Respondent	a1	a2	a3	a4
1	1	2	5	5
2	1	2	5	4
3	2	1	1	3
4	1	4	3	3
5	2	3	4	5
6	1	2	4	5

According to equation (1), (2), and (3):

$$ARS = [f(4)*1 + f(5)*1]/(j*n) = (4 + 5)/ 4*6 = .375$$

$$ERS = [f(1)*1 + f(5)*1]/(j* n) = (6 + 5)/ 4*6 = .458$$

$$MRS = [f(3)]/(j*n) = 4/4*6 = .167$$

According to results of Count Procedure above, ERS score (0.458) was higher than ARS (0.375) and MRS (0.167) scores, so it indicates the presence of ERS in this sample dataset.

1.2.2 Representative Indicators Response Style (RIRS)

To address the limitation of Count Procedure that correlations between items result in inaccurate estimation of RS, RIRS approach was developed (Baumgartner & Steenkamp, 2001). RIRS requires either adding a number of heterogeneous items, or to draw a random sample of survey items from a wide range of multi-item scales. These items are supposed to be uncorrelated, so that they could generate more accurate RS scores than those in Count Procedure (Weijters, 2006). The steps to apply RIRS are as follows: 1) Calculating RS scores based on extra items which are heterogeneous in content. 2) Regressing scores of content items on RS scores. 3) Conducting subsequent analyses with the effects of RS controlled. Therefore, RIRS can be employed to compute RS scores, and also correct the effects of RS on other main research analyses.

Figure 2 showed RIRS approach with three RS scores as covariates in a SEM framework. Compared to Figure 1, the major difference between RIRS and Count Procedure is that in RIRS, RS scores were calculated from extra items which are uncorrelated and different from content items, relative to Count Procedure in which RS scores were calculated from content items.

As found in Baumgartner and Steenkamp's study (2001), ARS was positively correlated with ERS; MRS was negatively correlated with ARS and ERS. Therefore, correlations between ARS, ERS, and MRS were specified as present and were estimated in RIRS.

Figure 2. RIRS model

1.2.3 Representative Indicators Response Style Means and Covariance

Structure (RIRSMACS)

Extended from RIRS, researchers further developed RIRSMACS (Weijters, Schillewaert, & Geuens, 2008). In RIRSMACS, additional uncorrelated items are required to add into questionnaires, and they are treated as observed variables and used to calculate RS scores. These RS scored are treated as latent covariates in subsequent analyses, so that content factors of scales and RS factors could be differentiated. For example, one content factor was measured by four observed variables, and 30 uncorrelated items were added to calculate RS scores. The 30 items are randomly divided into three blocks, and one indicator for each RS is calculated from 10 items of each block. In this case, each RS has three indicators from three blocks. Hence, the major difference between RIRS and RIRSMACS is that RS is treated as latent factors in RIRSMACS, and considered as observed RS scores in RIRS. RIRSMACS model presented in a SEM framework was shown in Figure 3.

According to Weijters et al. (2008), latent factors ERS, ARS, and MRS are correlated in RIRSMACS. Similarly, observed RS scores from the same block of items are treated as correlated. Therefore, in RIRSMACS, ERS1, ARS1, and MRS1 are correlated; ERS2, ARS2, and MRS2 are correlated; ERS3, ARS3, and MRS3 are correlated.

Figure 3. RIRSMACS model

1.2.4 Correction of RS by Count Procedure, RIRS, and RIRSMACS

The main purpose of investigating RS is not only to identify RS in empirical dataset, but also to correct the effects of RS on intended-to-be-measured traits, in order to obtain more accurate estimates of content factors and relationships between content factors of interest.

In RIRS and RIRSMACS, detection of RS requires heterogeneous items. Therefore, a commonly used correction procedure is regressing item scores on observed RS scores in RIRS and on latent RS factors in RIRSMACS, and subsequent statistical analyses could be conducted with the effects of RS controlled (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). For example, Thomas, Abts, Stroecken, and Weyden (2015) found that factor analysis results of institutional trust scale were biased by RS, and controlling the effects of RS by RIRSMACS could obtain more accurate estimates of factors of interest and relationships between latent factors.

1.2.5 Limitations of previous research on Count Procedure, RIRS, and RIRSMACS

Count Procedure is not widely used and not considered as an optimal method for identifying RS. One reason is that Count Procedure is a descriptive approach to detect RS, and no guidelines were provided on interpretation of results of Count Procedure. Cautions should be taken when it comes to interpretation. Previous research (Baumgartner & Steenkamp, 2001) suggested employing heterogeneous items with low/no correlations would generate more accurate results by Count Procedure. Some of previous studies (e.g., Marín, Gamba, & Marín, 1992, Watson, 1992) employed items that are heterogeneous in content, but this requires a big research project with multiple scales. In addition, no systematic simulations were conducted to support this assumption. That is, previous studies could use Count Procedure to generate RS scores in a specific sample, but validity and reliability of the results were not examined because true RS scores in a sample remain unknown. Therefore, the present

study aims to address these limitations, in order to generate more clear guidelines on how to interpret results of Count Procedure, and to explore performance of Count Procedure when correlations between items vary.

RIRS and RIRSMACS requires adding extra items that are heterogeneous in content and uncorrelated with each other and traits-of-interest. However, these two approaches have not been commonly applied in previous research because adding additional items may make experiment more time-consuming and more complex. In addition, as suggested by previous research, cognitive load may also affect the way respondents answer survey items, and adding more items could affect cognitive load (van Vaerenbergh & Thomas, 2012). In some studies, researchers selected uncorrelated items to create RS indicators, but it requires a large study with a good number of survey items. Therefore, it is not realistic to add extra items for calculating RS for all studies.

Furthermore, convergent validity of Count Procedure, RIRS, and RIRSMACS were not well established. One previous research has examined convergent validity between Count Procedure and RIRS (de Beuckelaer et al., 2010). Count Procedure was conducted among items of the food-related lifestyle measure, whereas RIRS was conducted among items randomly selected from a pool of marketing scales. Hence, items for Count Procedure are supposed to be correlated, and items drawn from marketing scales are heterogeneous in content. It was found that the proportion of ARS is the same for Count Procedure and RIRS, but the correlation between the two approaches was low, that is, respondents who tend to agree with the items of food-related lifestyle measures in Count Procedure do not necessarily show the same tendency in the other set of items randomly selected for RIRS approach. For

ERS, the correlation between the two methods was in moderate range (around 0.5), but the proportion of extreme responses differed across the two methods. The inconsistent results between Count Procedure and RIRS for ARS and ERS suggested that correlations between items play an important role in estimating RS accurately. However, it was not clear which approach - Count Procedure or RIRS was more reliable in detecting RS, because accuracy of RS approaches has not been examined in simulation studies. In addition, subjects in this study were selected from University students and survey items were drawn from marketing scales, so it remains unknown whether results could be generalized to other populations and items of different fields.

In addition, there is lack of systematic simulation research on RIRS and RIRSMACS. There was one simulation study showing that the performance of Count Procedure for ERS got worse as the correlation between items within one scale got bigger (de Jong et al., 2008). However, little was known on the performance of Count Procedure, RIRS and RIRSMACS for other types of RS under various conditions.

Therefore, the current study intended to address limitations of RIRS and RIRSMAS above, so to further explore the performance of RIRS and RIRSMACS in detection of different RS in simulation studies.

1.2.6 Item Response Theory (IRT)

IRT is a psychometric theory and family of associated mathematical models that relate latent traits of interest (also referred to as content factors in this dissertation) to the probability of responses to items on the assessment (van der Linden, & Hambleton, 1997). Basically, IRT model is employed to obtain

the likelihood of endorsing a response category of an item as a function of latent traits of interest and characteristics of the item.

IRT models developed for polytomous data can be classified into two groups for analysing nominal data and ordinal data. Nominal Response Model (NRM; Bock, 1972) was developed for responses on a nominal scale, and Partial Credit Model (PCM) was developed for responses on an ordinal scale. NRM was extended to Multidimensional NRM (Bock, Gibbons, & Muraki, 1988) when there are multiple latent traits. Recently, MNRM was employed to study RS (Bolt & Johnson, 2009). In MNRM, RS was treated as a continuous latent factor similar as other content factors.

1.2.6.1 Nominal Response Model (NRM)

NRM (Bock, 1972) are latent trait models for polytomous data where item responses are on a nominal scale. NRM specifies the probability that a respondent with a certain value of the latent trait selects response option (Ostini & Nering, 2006). Specifically, NRM could be employed to calculate the probability of a respondent with latent trait θ selecting a response category k on a certain item j .

$$P(U_j = k | \theta) = \frac{\exp(a_{jk}\theta + c_{jk})}{\sum_{h=1}^K \exp(a_{jh}\theta + c_{jh})} \quad (4)$$

where U_j is the observed response to item j ; θ is a latent trait of interest;

a_{jk} is the slope parameter for item j , response category k ; c_{jk} is intercept

parameter for item j , response category k . a_{jk} is also known as discrimination

parameter of item j . Larger a_{jk} values indicates items are more discriminating

among respondents, that is, respondents with different latent trait θ would be

more likely to endorse different responses. c_{jk} is the intercept parameter for item j , response category k . c_{jk} is considered as easiness parameter¹ of item j . Larger c_{jk} values means that items are easier for respondents to pass two categories of an item.

1.2.6.2 Partial Credit Model (PCM) - NRM with constraints

In psychological research, ordinal scale is widely used in questionnaire and survey items. For example, an item asks respondents to rate “How often have you felt nervous and stressed?” on a scale from “never”, “rarely”, “sometimes”, “very often”, to “always”; or respondents are asked how much they agree with a survey item “When I am getting upset, I know how to calm myself.” on a scale from “strongly disagree”, “disagree”, “neutral”, “agree”, to “strongly agree”. The partial credit model (PCM; Masters, 1982) was developed for polytomous items on an ordinal scale. As latent trait gets larger, the probability of choosing higher order response category increases. Conversely, as latent trait gets smaller, the probability of choosing lower order response category increases. PCM is a special case of NRM with constraints on slope parameters.

¹ In model 5, the sign before c_{jk} is positive, so c_{jk} is easiness parameter in the model. That is, the larger the value c_{jk} , the easier the item is. In some other IRT research, IRT model could also be represented in a different way, where the sign before c_{jk} is negative (Lord, 1980). In that case, c_{jk} is considered as difficulty parameter, the larger the value c_{jk} , the difficult the item is.

1.2.6.3 Multidimensional Nominal Response Model (MNRM)

NRM presented above contains one latent trait, and it is also known as unidimensional NRM. However, unidimensional NRM is limited in situations when there are multiple latent traits to be included in the model. Bock et al. (1988) then developed MNRM to include multiple latent traits. Later, Bolt and Johnson (2009) introduced RS factor to MNRM for modelling RS:

$$P(U_j = k | \theta_1, \theta_2, \dots, \theta_d) = \frac{\exp(a_{jk1}\theta_1 + a_{jk2}\theta_2 + \dots + a_{jkd}\theta_d + c_{jk})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1 + a_{jh2}\theta_2 + \dots + a_{jhd}\theta_d + c_{jh})} \quad (5)$$

where U_j is the observed response to item j ; θ is a latent trait; a_{jk} is the slope parameter for item j , response category k ; c_{jk} is intercept parameter for item j , response category k , and d represented number of latent traits.

As MNRM could be applied when there is more than one latent trait, it has been used to model RS. RS factor is treated as a continuous latent trait and represented by a separate factor in addition to content factors in MNRM. For example, below is a MNRM with one content factor and one RS factor (For MNRM with content factors and RS factor, content factors were named content factor in the following parts for simplicity purpose.)

$$P(U_j = k | \theta_1, \theta_{RS}) = \frac{\exp(a_{jk1}\theta_1 + a_{jk2}\theta_{RS} + c_{jk})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1 + a_{jh2}\theta_{RS} + c_{jh})} \quad (6)$$

where U_j is the observed response to item j ; θ_1 is content factor; θ_{RS} is RS factor; a_{jk} is the slope parameter for item j , response category k ; c_{jk} is intercept parameter for item j , response category k .

Previous research has employed MNRM to examine multiple content factors and ERS. It was suggested that MNRM can estimate both the respondent's content factors as well as ERS (Bolt & Johnson, 2009). To

distinguish between content factors and RS factors, different slope parameters are used to indicate different relationships between content factors and response categories. The following section discussed how to model various RS by imposing constraints on corresponding slopes.

1.2.6.3.1 MNRM with constraints to model RS

For items on an ordinal scale, constraints could be imposed on slope parameters of content factors in PCM. As discussed above, a simple set of equally spaced slope parameters represent ordinal relationships. Figure 4 presented item category characteristic curves (ICC) of a unidimensional PCM with slope parameters of (-2, -1, 0, 1, 2). This set of slopes, for example, indicated an ordinal likert scale which is commonly used in psychological studies. That is, subjects who are high on a content factor would be more likely to endorse a larger response category on a likert scale. Conversely, subjects who are low on a content factor would be more likely to choose a smaller response category. Figure 4 presents ICC of an item with medium difficulty ($c_k = 0$) based on equation (4). That is, as content factor gets larger, respondents are more likely to pick category 5 (the probability of picking category 5 is close to 1 as content factors gets close to 3); as content factor gets smaller, respondents are more likely to pick category 1 (the probability of picking category 1 is close to 1 as content factor gets close to -3); when content factor is 0, the probability of picking each category is the same (0.2).

Figure 4. ICC of one content factor NRM on a five-point scale item

Different sets of slope constraints can be specified to capture different RS. For example, ERS factor can be introduced by adding constraints on slope parameters (e.g., 1, 0, 0, 0, 1, Bolt & Newton, 2011) associated with ERS factor in MNRM, so that probability of choosing extreme response categories are higher relative to when there was no ERS. According to Model (6), below presents a 2 dimensional-MNRM (1 content factor, 1 ERS factor, 5 response categories) in Figure 5. Mean of ERS factor score equal to 1 indicates a strong response tendency to endorse extreme response categories. In Figure 5, slope parameters for content factor are (-2, -1, 0, 1, 2), for ERS are (1, 0, 0, 0, 1). Compared to NRM with no RS factor shown in Figure 4, respondents are more likely to endorse extreme responses (1 and 5), especially when they have higher or lower scores on content factors. Even when content factor is 0, the probability of choosing responses 1 and 5 is higher, relative to 2, 3, and 4.

Figure 5. ICC of MNRM with one content factor and ERS

Similarly, ARS – the tendency to agree with the item and choose response category 4 and 5 on a five-point scale, can be introduced by employing another set of constraints on slope parameters (e.g., 0, 0, 0, 1, 1) associated with ARS factor in MNRM, so that the probability of choosing agreements are higher than when there was no RS. As shown in Figure 6, respondents are more likely to agree with items (choosing response categories 4 and 5). Mean of ARS factor score equal to 1 indicates a strong response tendency to agree with items.

Figure 6. ICC of MNRM with one content factor and ARS

Similar to ARS and ERS, MRS – the tendency to use the midpoint response category 3 on a five-point scale item, could be introduced by employing another set of constraints on slope parameters (e.g., 0, 0, 1, 0, 0) associated with MRS factor, so that probability of choosing mid-point response is higher than when there is no RS. As shown in Figure 7, respondents are more likely to choose the middle category 3. Mean of MRS factor score equal to 1 indicates a strong response tendency to select mid-point response category.

Figure 7. ICC of MNRM with one content factor and MRS

Bolt and Newton (2011) employed MNRM with 2 content factors and one RS factor to examine the effects of multiple content factors on estimation of RS. The model in their study was shown below:

$$P(U_j = k | \theta_1, \theta_2, \theta_{RS}) = \frac{\exp(a_{jk1}\theta_1 + a_{jk2}\theta_2 + a_{jk3}\theta_{RS} + c_{jk})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1 + a_{jh2}\theta_2 + a_{jh3}\theta_{RS} + c_{jh})} \quad (7)$$

where U_j is the observed response to item j ; θ_1 and θ_2 are two content factors; θ_{RS} is RS factor; a_{jk} is the slope parameter for item j , response category k ; c_{jk} is intercept parameter for item j , response category k .

Results of their study suggested MNRM performed better in estimating ERS through simultaneous analysis of multiple content factors, indicated by smaller difference between true and estimated ERS trait and higher correlations between true and estimated ERS trait. In addition, it was found that the improvements were not affected by correlations between content factors of various scales. This finding not only supported the stability of ERS across

multiple scales, but also provided implications for making use of MNRM to obtain a more accurate ERS with multiple scales.

1.2.6.3.2 Previous research on MNRM and its limitations

Compared to other RS detection approaches such as Count Procedure, RIRS and RIRSMACS, MNRM has some advantages. Different from Count Procedure, MNRM does not require a heterogeneous scale for a better performance. In addition, MNRM doesn't require additional items for calculating RS scores as RIRS and RIRSMACS. Influence of RS factors on estimates of content factors could be controlled by separating trait factors and RS in the model. Furthermore, RS is considered as a trait for individuals in MNRM, so it allows for individual differences in RS factor scores. Therefore, researchers have showed increasing interest in development of MNRM model for RS analysis. Johnson and Bolt (2010) demonstrated that ignoring individual differences in ERS would lead to incorrect assessment of content factors in a simulation study. As correlation between content factors and ERS increased, the estimation bias between estimated and true item response curves got larger if ERS was ignored. Another study simulated data from MNRM (Bolt & Newton, 2011) and found that with a content factor and ERS in the model, adding another content factor improved recovery of ERS, and the improvement was not affected by the correlations between the two content factors. It was also found that adding ERS factor helped with estimation of content factor, compared to the effects of adding another content scale. That is, improvement in estimation of a content factor is more accounted for by ERS than another content factor. Therefore, Bolt and Newton's (2011) simulation study has further supported the

significance of accounting for ERS in the model, which can help to estimate content factors more accurately.

However, employment of MNRM in RS research is not without limitations. Previous research of MNRM has only been focused on ERS (van Vaerenbergh & Thomas, 2012). Even though some researchers have suggested that testing other RS should follow similar procedure and generate similar conclusions as ERS (e.g., Jin & Wang, 2014; Johnson, 2003), this argument has not been fully examined and investigated in previous studies. It remains a question whether estimating other RS in MNRM could help with more accurate assessment of content factors like ARS. Hence, one objective of present study is to explore model performance of MNRM on other RS, such as ARS and MRS.

In MNRM, both content factor θ and RS factor θ_{RS} are treated as continuous latent variables following a normal distribution. For example, in de Jong et al.' (2008) study, ERS was simulated from a normal distribution $N(0, 1)$. In Bolt and Newton's (2011) simulation study, data was simulated from a multivariate normal distribution with mean equal to 0. Even though researchers realized that ERS may not follow a normal distribution $N(0, 1)$ (Johnson & Bolt, 2010), previous simulation studies were conducted with the assumption of normal assumption. If ERS follows a normal distribution with mean 0, ERS factor score can take a positive or a negative value. A positive ERS value represents the tendency to use the highest and lowest response categories of a rating scale, whereas a negative ERS value refers to the tendency to avoid selecting the highest and lowest response categories, and it is consistent with the definition of MLRS.

This can be demonstrated in Figure 8 based on model (6). Figure 8(a) presented ICC with $\theta_{ERS} = 1$. Figure 8(b) presented ICC when $\theta_{ERS} = 0$. Figure 8(c) presented ICC when $\theta_{ERS} = -1$. Figure 8(d) presented ICC when $\theta_{MLRS} = 1$. In Figure 8(b), $\theta_{ERS} = 0$ is considered as reference, meaning there is no ERS in the sample. Figure 8(a) refers to a strong ERS as ERS value was 1. Figure 8(c) refers to a strong response pattern to choose fewer extreme responses 1 and 5, and more mild responses 2, 3, and 4, because ERS values was -1. Figure 8(d) presented strong MLRS as MLRS value was 1.

As shown in Figure 8(c) and 8(d), these two figures had similar pattern that probabilities of choosing extreme responses 1 and 5 were small than when there was no RS, and probabilities of these two figures were consistent with definition of MLRS, so that mild responses 2, 3, and 4 were more likely to be endorsed relative to no RS condition. Therefore, if ERS follows a normal distribution with mean 0, this represents ERS and MLRS together rather than ERS only. However, previous researchers did not address this question clearly. In prior RS studies, a positive θ_{ERS} was considered as high ERS, and a negative θ_{ERS} as low ERS. For example, Bolt and Johnson (2009) presented item category characteristic curves (ICC) for a certain item from a scale. The 0.75 ERS factor score was noted as high ERS, and -0.75 ERS factor score was noted as low ERS. However, as discussed above, -0.75 ERS is actually MLRS – the tendency to avoid using extreme response categories.

Figure 8. ICC of MNRM with one content factor and RS of different values

8a.

8b.

8c.

8d.

Chapter 2

Aims of Study and Study Design

2.1 Aims of study

This dissertation explores performance of Count Procedure, RIRS and RIRSMACS, and MNRM in simulation studies. Specific questions addressed in present study are as below:

For Count Procedure:

1. to evaluate performance of Count Procedure in identifying various RS through simulation studies;
2. to explore effects of correlations between items on performance of Count Procedure;

For RIRS and RIRSMACS:

To evaluate performance of RIRS and RIRSMACS in identifying various RS through simulation studies. Specifically :

1. to explore whether correct model could be picked out based on model-fit indices in RIRS and RIRSMACS;
2. to explore parameter recovery that whether correlations between two content factors could be well-recovered in RIRS and RIRSMACS;

For MNRM:

To evaluate performance of MNRM in identifying various RS in simulation studies. Specifically:

1. to explore whether correct model could be picked out based on model-fit indices;
2. to explore parameter recovery of MNRM, including intercept parameter, variance of RS, and correlations between two content factors of MNRM.

2.2 Method

2.2.1 Study design

In the present study, number of response categories was fixed to five, which is commonly observed in survey research. Number of content factors was fixed at two, and correlation between two content factors was fixed at 0.5. In prior simulations studies, it was found estimates of ERS can be improved through simultaneous analysis of multiple scales. Moreover, the improvements was unaffected by correlations (0.3 vs. 0.5 vs. 0.8) between content factors (Bolt & Newton, 2011). Therefore, present study included 2 content factors with correlation of 0.5 in simulation studies to explore performance of MNRM in RS detection. Content factors were generated from a normal distribution with mean of 0 and standard deviation of 1. All the three RS factors – ERS, ARS, and MRS were generated from a normal distribution with mean of 1 and standard deviation of 0.25. This ensures all RS values generated are positive. For example, when ERS was studied, all generated ERS true scores follow a normal distribution with mean of 1 and standard deviation of 0.25, and this ensures all ERS scores are all positive values above 0. As was discussed earlier, negative ERS values is actually another type of RS – MLRS. Therefore, generating all positive ERS values makes it possible to examine one RS – ERS

instead of two RS – ERS and MLRS .In addition, the number of replications were fixed at 100 for each condition.

2.2.2 Independent variables

The present study explored the performance of Count Procedure, RIRS, RIRSMACS, and MNRM under various conditions as follows: sample size (300 vs. 800); number of items per content factor (10 vs. 20); RS (NO RS, ARS, ERS, and MRS); correlations between items (items from two correlated content factors only vs. 15 items from 5 uncorrelated factors).

Levels of independent variables were determined based on prior research and simulation studies on RS. In psychological research, means of sample sizes may range from less than 100 (Journal of Experimental Psychology) to close to 1000 (Journal of Applied Psychology) (Holmes, 1983). Furthermore, it was found sample sizes in Journal of Applied Psychology and Journal of Experimental Psychology stayed relatively constant from 1977 to 2006; and sample sizes in Journal of Abnormal Psychology and Journal of Developmental Psychology increased slightly from 1977 to 2006 (Marszalek, Barber, Kohlhart, & Holmes, 2011). Therefore, the present study picked two sample sizes, 300 and 800 indicating small and large sample sizes in psychology studies.

For number of items per content factor, two levels were chosen: 10 vs. 20 items. The reason is that it is common to have a small number of items per factor in psychological measurement. Furthermore, number of items in most of previous simulation studies of RS vary from a small number of 4 to 20 (e.g., Bolt & Newton, 2011; Cho, 2013). Therefore, present study examined two

conditions of number of items per content factor, 10 vs. 20 (that is 20 items vs. 40 items in total due to 2 content factors) to represent short and long scale length.

Previous researchers suggested Count Procedure may perform better in detection of RS using heterogeneous items. However, this assumption has not been examined in simulation studies. Therefore, present study explored effects of correlations between items on detection of RS. Two conditions were examined: correlated items and uncorrelated items. Correlated items consisted of items from 2 correlated content factors with correlation coefficient at 0.5; uncorrelated items were from 5 uncorrelated factors. In psychological research, it is common to have a few number of correlated content factors. Items from one content factor or correlated content factors are considered as correlated items. Therefore, performance of Count Procedure were examined by correlated items from 2 content factors. In addition to items from 2 correlated factors, the present study also examined performance of Count Procedure with 15 uncorrelated items from 5 uncorrelated factors (3 items from one factor). These 5 factors were uncorrelated to each other, and uncorrelated with two content factors as well. On one hand, this condition allowed for examination of approaches when there was low correlation between items; on the other hand, there was low correlations between these 15 extra items, so it was more similar to empirical research in that it would be more realistic to obtain extra items with low correlation in a large research study rather than items with no correlations at all.

2.2.3 Analysis software

Simulated datasets were generated from Matlab (MATLAB, 2013), and estimated in Mplus (Muthén & Muthén, 2013). Mplus is widely used for structural equation models because it is flexible to add constraints and can generate model-fit indices, including log likelihood, AIC, BIC, and adj BIC. Researchers explored how to estimate NRM with constraints in Mplus (Huggins-Manley & Algina, 2015). Therefore, Mplus was employed to analyze simulated datasets with RIRS, RIRSMACS, and MNRM approaches.

2.2.4 Data generation

Data was simulated developed from MNRM as shown in Figure (9). Content items were simulated by MNRM with 2 correlated content factors with/out RS factor based on equation (7) and (8). MNRM simulating content items can be expressed as:

$$P(U_j = k | \theta_1, \theta_2, \theta_{RS}) = \frac{\exp(a_{jk1}\theta_1 + a_{jk2}\theta_2 + a_{jk3}\theta_{RS} + c_{jk})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1 + a_{jh2}\theta_2 + a_{jh3}\theta_{RS} + c_{jh})} \quad (7)$$

$$P(U_j = k | \theta_1, \theta_2) = \frac{\exp(a_{jk1}\theta_1 + a_{jk2}\theta_2 + c_{jk})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1 + a_{jh2}\theta_2 + c_{jh})} \quad (8)$$

where U_j is the observed response to item j; θ_1 and θ_2 are two content factors; θ_{RS} is RS factor; a_{jk} is the slope parameter for item j, response category k; c_{jk} is intercept parameter for item j, response category k.

Extra heterogeneous items were simulated by 5 uncorrelated factors with/out RS factor.

Figure 9. Data generation model

9a. Data generation model for No RS

9b. MNRM model for ARS/ERS/MRS

In figure 9, A and B represented 2 content factors, and each consisted of 10/20 items. 2 content factors were correlated and correlation coefficient was free to estimate. ARS/ERS/MRS was a latent variable. Path coefficients from ARS/ERS/MRS variables to content items were free to estimate. E1 to E5 represented 5 uncorrelated extra factors. 3 items were simulated from each of the 5 extra factors. In total, 15 heterogeneous items were simulated.

In the present study, Mplus was employed to analyse simulated data. For factors which are on ordinal scale, there are different ways to specify slope parameters as long as it indicates an ordinal relationship. For example, 0, 1, 2, 3, 4, or 4, 3, 2, 1, 0. However, for RS factors, it could be tricky. Take ERS for example, slopes can be fixed at 0 for non-extreme response categories, and at 1 for extreme response categories to indicate a higher possibility of ERS. Furthermore, Mplus has a default that fixes slope and intercept parameters for the last response category to 0 (Muthén & Muthén, 2013). To model the 3 types RS, only category 2 can be chosen as reference category because slopes and intercepts of category 2 can be fixed to 0, as shown in Table 2. Therefore, in the present study, data was recoded ($1 \rightarrow 1$, $2 \rightarrow 5$, $3 \rightarrow 2$, $4 \rightarrow 3$, $5 \rightarrow 4$) in the way that original response category 2 was recoded to 5 - the last response category, so recoded category 2 was treated as reference category in Mplus.

Table 2. Slope parameters in MNRM (recoded) for content factors and RS factors

Traits	Category 1	Category 2	Category 3	Category 4	Category 5
θ_1/θ_2	-1	0	1	2	3

θ_{ARS}	0	0	0	1	1
θ_{ERS}	1	0	0	0	1
θ_{MRS}	0	0	1	0	0

For ARS, MNRM for 5 response categories (recoded) are as follows:

$$P(U_j = 1 | \theta_1, \theta_2, \theta_{ARS}) = \frac{\exp((-1)*\theta_1 + (-1)*\theta_2 + 0*\theta_{RS} + c_{j1})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1 + a_{jh2}\theta_2 + a_{jh3}\theta_{RS} + c_{jh})} \quad (9a)$$

$$P(U_j = 2 | \theta_1, \theta_2, \theta_{ARS}) = \frac{\exp(0*\theta_1 + 0*\theta_2 + 0*\theta_{RS} + c_{j2})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1 + a_{jh2}\theta_2 + a_{jh3}\theta_{RS} + c_{jh})} \quad (9b)$$

$$P(U_j = 3 | \theta_1, \theta_2, \theta_{ARS}) = \frac{\exp(1*\theta_1 + 1*\theta_2 + 0*\theta_{RS} + c_{j3})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1 + a_{jh2}\theta_2 + a_{jh3}\theta_{RS} + c_{jh})} \quad (9c)$$

$$P(U_j = 4 | \theta_1, \theta_2, \theta_{ARS}) = \frac{\exp(2*\theta_1 + 2*\theta_2 + 1*\theta_{RS} + c_{j4})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1 + a_{jh2}\theta_2 + a_{jh3}\theta_{RS} + c_{jh})} \quad (9d)$$

$$P(U_j = 5 | \theta_1, \theta_2, \theta_{ARS}) = \frac{\exp(3*\theta_1 + 3*\theta_2 + 1*\theta_{RS} + c_{j5})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1 + a_{jh2}\theta_2 + a_{jh3}\theta_{RS} + c_{jh})} \quad (9e)$$

For ERS, MNRM for 5 response categories (recoded) are as follows:

$$P(U_j = 1 | \theta_1, \theta_2, \theta_{ARS}) = \frac{\exp((-1)*\theta_1 + (-1)*\theta_2 + 1*\theta_{RS} + c_{j1})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1 + a_{jh2}\theta_2 + a_{jh3}\theta_{RS} + c_{jh})} \quad (10a)$$

$$P(U_j = 2 | \theta_1, \theta_2, \theta_{ARS}) = \frac{\exp(0*\theta_1 + 0*\theta_2 + 0*\theta_{RS} + c_{j2})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1 + a_{jh2}\theta_2 + a_{jh3}\theta_{RS} + c_{jh})} \quad (10b)$$

$$P(U_j = 3 | \theta_1, \theta_2, \theta_{ARS}) = \frac{\exp(1*\theta_1 + 1*\theta_2 + 0*\theta_{RS} + c_{j3})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1 + a_{jh2}\theta_2 + a_{jh3}\theta_{RS} + c_{jh})} \quad (10c)$$

$$P(U_j = 4 | \theta_1, \theta_2, \theta_{ARS}) = \frac{\exp(2*\theta_1 + 2*\theta_2 + 0*\theta_{RS} + c_{j4})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1 + a_{jh2}\theta_2 + a_{jh3}\theta_{RS} + c_{jh})} \quad (10d)$$

$$P(U_j = 5 | \theta_1, \theta_2, \theta_{ARS}) = \frac{\exp(3*\theta_1 + 3*\theta_2 + 1*\theta_{RS} + c_{j5})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1 + a_{jh2}\theta_2 + a_{jh3}\theta_{RS} + c_{jh})} \quad (10e)$$

For MRS, MNRM for 5 response categories (recoded) are as follows:

$$P(U_j = 1 | \theta_1, \theta_2, \theta_{ARS}) = \frac{\exp((-1)*\theta_1 + (-1)*\theta_2 + 0*\theta_{RS} + c_{j1})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1 + a_{jh2}\theta_2 + a_{jh3}\theta_{RS} + c_{jh})} \quad (11a)$$

$$P(U_j = 2|\theta_1, \theta_2, \theta_{ARS}) = \frac{\exp(0*\theta_1+0*\theta_2+0*\theta_{RS}+c_{j2})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1+a_{jh2}\theta_2+a_{jh3}\theta_{RS}+c_{jh})} \quad (11b)$$

$$P(U_j = 3|\theta_1, \theta_2, \theta_{ARS}) = \frac{\exp(1*\theta_1+2*1*\theta_{RS}+c_{j3})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1+a_{jh2}\theta_2+a_{jh3}\theta_{RS}+c_{jh})} \quad (11c)$$

$$P(U_j = 4|\theta_1, \theta_2, \theta_{ARS}) = \frac{\exp(2*\theta_1+2*\theta_2+0*\theta_{RS}+c_{j4})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1+a_{jh2}\theta_2+a_{jh3}\theta_{RS}+c_{jh})} \quad (11d)$$

$$P(U_j = 5|\theta_1, \theta_2, \theta_{ARS}) = \frac{\exp(3*\theta_1+3*\theta_2+1*\theta_{RS}+c_{j5})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1+a_{jh2}\theta_2+a_{jh3}\theta_{RS}+c_{jh})} \quad (11e)$$

Intercept parameters were decided according to simulation studies in Cho (2013). In Cho's (2013) studies, probabilities of response categories follow a normal distribution when there was no RS. Therefore, probabilities for each of 5 response categories are: 14% for category 1, 22% for category 2, 28% for category 3, 22% for category 4, and 14% for category 5. Then, category probabilities for each item were manipulated to be different among items. Table 3 presented 5 category probabilities for 10 items.

Table 3. Category probabilities (recoded) for individual items for NO RS (Cho (2013))

Item	Category1	Category2	Category3	Category4	Category5
1	0.1478	0.2245	0.2556	0.2245	0.1479
2	0.1539	0.2061	0.2801	0.2061	0.1539
3	0.1244	0.2413	0.2685	0.2413	0.1244
4	0.1069	0.2412	0.3038	0.2412	0.1069
5	0.1233	0.2354	0.2825	0.2354	0.1233

6	0.1657	0.2071	0.2543	0.2071	0.1657
7	0.1332	0.2308	0.2721	0.2308	0.1332
8	0.1550	0.2124	0.2653	0.2124	0.1550
9	0.1501	0.2065	0.2867	0.2065	0.1501
10	0.1416	0.1904	0.3360	0.1904	0.1416
Mean	0.1402	0.2196	0.2805	0.2196	0.1402

The next step was to calculate intercept parameters from probabilities according to IRT model. According to results of Cho (2013), intercept parameters for 10 items were presented in Table 4 below. In the condition with 20 items, intercept parameters were repeated. As noted earlier, category 2 was recoded as a reference category with 0 as intercept parameters, and intercepts of other response categories were recoded accordingly.

Table 4. Intercept parameters (recoded) for individual items

Item	Category1	Category2	Category3	Category4	Category5
1	-1.5181	0	0.5998	0.1	-1.3181
2	-1.2924	0	0.6768	-0.1	-1.4924
3	-1.8123	0	0.6265	0.2	-1.4123
4	-1.7632	0	0.551	-0.2	-2.1632

5	-1.8469	0	0.7522	0.3	-1.2469
6	-1.1232	0	0.475	-0.3	-1.7232
7	-1.7999	0	0.7849	0.4	-0.9999
8	-1.1654	0	0.4421	-0.4	-1.9654
9	-1.6191	0	0.998	0.5	-0.6191
10	-1.0966	0	0.7379	-0.5	-2.0966

Chapter 3

Results and Discussion of Count Procedure

3.1 Analysis models

Results of Count Procedure were obtained by equation (1), (2), and (3). RS scores were calculated from items: (1). content items only; (2) 15 heterogeneous items only; (3). content items and 15 heterogeneous items.

3.2. Evaluation criteria

For Count Procedure, evaluation criteria included: RS scores calculated by counting; Pearson and Spearman correlations between true RS and estimated RS scores. It was expected that in correct model, correlations would be higher, so true RS scores for individuals could be better predicted by estimated RS scores.

3.3 Results of Count Procedure

3.3.1 Results of Count Procedure – NO RS condition

RS scores calculated with Count Procedure were presented in three conditions in Table 5, Table 6, Table 8, and Table 10: content items only (items from 2 correlated content factors), uncorrelated extra items only (15 items from 5 uncorrelated factors), and all the items including content items and 15 extra items. 15 items from 5 uncorrelated factors were considered as heterogeneous items. ARS, ERS, and MRS scores calculated in NO RS condition with heterogeneous items were considered as benchmark values. In the present study, one assumption was that responses follow a normal distribution. Therefore, when there was no RS, ARS, ERS, and MRS values are non-zero

because there would be agreements, extreme responses, and mid-point responses. For example, when there is no RS, ERS score would be a positive value because there would be individuals choosing extreme response categories due to high score on that content factor, MRS score would be a positive value because there would be individuals choosing mid-point response category due to their score on content factor. Therefore, it is difficult to tell if a certain RS score indicates presence of RS. Therefore, Benchmark values for ARS, ERS, and MRS were established to confirm presence of RS.

When there was no RS, results of Count Procedure were presented in Table 5. As shown in Table 5, ARS score calculated by equation (1) was around 0.340 (range from 0.337 to 0.352), ERS score calculated by equation (2) was close to 0.230 (range from 0.232 to 0.233), and MRS score calculated by equation (3) is around 0.320 (range from 0.313 to 0.325). Performance of Count Procedure did not change much as sample size increased from 300 to 800, and it did not change much as number of items increased from 20 to 40. In addition, correlations between items did not affect Count Procedure when there was no RS, as RS scores did not differ much in condition of all content items and 15 heterogeneous items. This was inconsistent with prior assumptions that Count Procedure performed better with heterogeneous items.

Table 5. Mean and standard deviation (SD) of RS scores calculated by counting in NO RS condition

NO RS									
Sample size/ test length	content items			15 uncorrelated items			content + 15 uncorrelated items		
	ARS	ERS	MRS	ARS	ERS	MRS	ARS	ERS	MRS
	mean (SD)	mean (SD)	mean (SD)	mean (SD)	mean (SD)	mean (SD)	mean (SD)	mean (SD)	mean (SD)
300/20	0.340 (0.015)	0.233 (0.008)	0.325 (0.008)	0.352 (0.009)	0.233 (0.008)	0.314 (0.008)	0.345 (0.009)	0.233 (0.006)	0.320 (0.006)
800/20	0.338 (0.014)	0.232 (0.007)	0.325 (0.006)	0.351 (0.010)	0.233 (0.007)	0.314 (0.007)	0.342 (0.011)	0.233 (0.006)	0.322 (0.005)
300/40	0.339 (0.009)	0.233 (0.005)	0.324 (0.005)	0.352 (0.005)	0.233 (0.004)	0.314 (0.004)	0.345 (0.005)	0.233 (0.004)	0.320 (0.003)
800/40	0.337 (0.008)	0.233 (0.004)	0.324 (0.004)	0.352 (0.006)	0.233 (0.004)	0.313 (0.005)	0.341 (0.006)	0.233 (0.004)	0.321 (0.003)

3.3.2 Results of Count Procedure – ARS condition

When there was ARS present, results of Count Procedure showed that ARS scores were much higher than other RS and benchmark value of ARS (0.350). As shown in Table 6, ARS calculated by equation (1) was around 0.500 (range from 0.498 to 0.515), ERS score calculated by equation (2) is close to 0.250 (0.256 to 0.258), and MRS score calculated by equation (3) is around 0.220 (range from 0.212 to 0.223). ARS score in Count Procedure did not change much as sample size increased from 300 to 800, and as number of items increased from 20 to 40. In addition, ARS scores increased slightly when there was low correlations between items from close to 0.500 (based on content items) to 0.515 (based on 15 uncorrelated items).

Furthermore, Pearson and Spearman correlations between true ARS and estimated ARS scores by Count Procedure were calculated and presented in

Table 7. Results showed that correlations between true and estimated ARS scores calculated from 15 heterogeneous items was a little higher (above 0.200) relative to from correlated content items (about 0.150). However, all correlation values were below 0.300, so it was not advisable to predict ARS scores for individuals from estimated ARS scores from Count Procedure. In addition, the slight difference in correlation coefficient between content items and 15 uncorrelated items did not support the assumption that Count Procedure performed better by using uncorrelated items.

Table 6. Mean and standard deviation (SD) of RS scores calculated by Count Procedure in ARS condition

ARS									
Sample size/ test length	content items			15 uncorrelated items			content + 15 uncorrelated items		
	ARS	ERS	MRS	ARS	ERS	MRS	ARS	ERS	MRS
	mean (SD)	mean (SD)	mean (SD)	mean (SD)	mean (SD)	mean (SD)	mean (SD)	mean (SD)	mean (SD)
300/20	0.499 (0.014)	0.257 (0.008)	0.222 (0.007)	0.515 (0.013)	0.258 (0.008)	0.212 (0.007)	0.506 (0.009)	0.257 (0.006)	0.217 (0.005)
800/20	0.500 (0.016)	0.257 (0.009)	0.222 (0.006)	0.514 (0.009)	0.257 (0.007)	0.212 (0.007)	0.504 (0.012)	0.257 (0.007)	0.219 (0.005)
300/40	0.501 (0.008)	0.256 (0.005)	0.222 (0.004)	0.515 (0.006)	0.258 (0.004)	0.212 (0.004)	0.507 (0.005)	0.257 (0.003)	0.218 (0.003)
800/40	0.498 (0.009)	0.256 (0.005)	0.223 (0.003)	0.515 (0.006)	0.258 (0.004)	0.212 (0.004)	0.503 (0.007)	0.256 (0.004)	0.220 (0.003)

Table 7. Correlations between true and estimated ARS scores in Count

Procedure

ARS						
Sample size/ test length	content items	Pearson r		content items	Spearman r	
		15 uncorrelated items	content + 15 uncorrelated items		15 uncorrelated items	content + 15 uncorrelated items
300/20	0.157	0.246	0.246	0.155	0.236	0.236
800/20	0.159	0.243	0.211	0.155	0.235	0.202
300/40	0.153	0.241	0.240	0.139	0.231	0.228
800/40	0.158	0.238	0.209	0.155	0.228	0.202

3.3.3 Results of Count Procedure – ERS condition

For ERS condition, results of Count Procedure suggested ERS scores were the largest among all RS, and ERS scores were higher than benchmark of ERS (0.230). As shown in Table 8, ARS score calculated by equation (1) was around 0.380 (range from 0.373 to 0.389), ERS score calculated by equation (2) was around 0.420 (range from 0.419 to 0.423), and MRS score calculated by equation (3) is around 0.250 (range from 0.241 to 0.251). Differences between ARS and ERS scores were small. Therefore, it would be more difficult to detect ERS from ARS correctly by Count Procedure, and ARS is likely to be picked out when correct RS is ERS. Performance of Count Procedure did not change much as sample size increased from 300 to 800, and as number of items increased from 20 to 40. In addition, ERS scores did not increase much when there was low correlations between 15 extra items compared to using correlated content items.

Furthermore, as shown in Table 9, correlations between true and estimated ERS scores calculated from 15 uncorrelated items was a little higher (about 0.350) relative to correlated content items (about 0.300), and the correlations were the highest when combining content items and uncorrelated items. This result implicated estimation of ERS was more accurate with uncorrelated items, and more items helped with more accurate estimation of ERS. Though it was not advisable to predict ERS scores for individuals from estimated ERS scores of Count Procedure due to low correlation coefficients, it was implicated that including more items could help better predict true ERS score.

Table 8. Mean and standard deviation (SD) of RS scores calculated by Count Procedure in ERS condition

Sample size/ test length	ERS								
	content items			15 uncorrelated items			content + 15 uncorrelated items		
	ARS mean (SD)	ERS mean (SD)	MRS mean (SD)	ARS mean (SD)	ERS mean (SD)	MRS mean (SD)	ARS mean (SD)	ERS mean (SD)	MRS mean (SD)
300/20	0.373 (0.016)	0.420 (0.011)	0.251 (0.007)	0.389 (0.010)	0.423 (0.010)	0.241 (0.006)	0.380 (0.010)	0.421 (0.008)	0.246 (0.005)
800/20	0.374 (0.014)	0.419 (0.009)	0.251 (0.006)	0.387 (0.009)	0.422 (0.008)	0.241 (0.007)	0.378 (0.010)	0.420 (0.008)	0.249 (0.005)
300/40	0.375 (0.010)	0.420 (0.006)	0.251 (0.004)	0.389 (0.006)	0.422 (0.006)	0.241 (0.004)	0.381 (0.006)	0.421 (0.004)	0.247 (0.003)
800/40	0.373 (0.009)	0.420 (0.006)	0.251 (0.003)	0.389 (0.006)	0.422 (0.005)	0.242 (0.004)	0.378 (0.007)	0.420 (0.005)	0.249 (0.003)

Table 9. Correlations between true and estimated ERS scores in Count

Procedure

ERS						
Sample size/ test length	Pearson correlation			Spearman correlation		
	content items	15 uncorrelat ed items	content + 15 uncorrelat ed items	content items	15 uncorrelat ed items	content + 15 uncorrelat ed items
300/20	.300	.352	.424	.297	.340	.412
800/20	.300	.348	.421	.297	.334	.408
300/40	.333	.352	.416	.336	.340	.414
800/40	.323	.351	.407	.326	.338	.403

3.3.4 Results of Count Procedure – MRS condition

When there was MRS present, results of Count Procedure showed that MRS scores were much higher than other RS, and MRS scores were higher than benchmark value of MRS (0.310). As shown in Table 10, ARS score calculated by equation (1) was above 0.200 (range from 0.226 to 0.240), ERS score calculated by equation (2) was below 0.200 (range from 0.166 to 0.168), and MRS score calculated by equation (3) was above 0.500 (range from 0.534 to 0.548). Performance of Count Procedure did not change much as sample size increased from 300 to 800, and as number of items increased from 20 to 40. In addition, MRS scores did not change significantly as correlations between items vary.

Furthermore, as shown in Table 11, correlations between true and estimated MRS scores calculated from uncorrelated extra items did not increase much relative to correlated content items (all below 0.400). In addition,

correlations between true and estimated MRS scores were highest when content items and uncorrelated items were combined (close to 0.500). This result implicated necessity to include more items to predict true MRS score more accurately.

Table 10. Mean and standard deviation (SD) of RS scores calculated by Count Procedure in MRS condition

MRS									
Sample size/ test length	content items			15 uncorrelated items			content + 15 uncorrelated items		
	ARS	ERS	MRS	ARS	ERS	MRS	ARS	ERS	MRS
	mean (SD)	mean (SD)	mean (SD)	mean (SD)	mean (SD)	mean (SD)	mean (SD)	mean (SD)	mean (SD)
300/20	0.227 (0.011)	0.167 (0.009)	0.546 (0.011)	0.240 (0.008)	0.168 (0.007)	0.534 (0.009)	0.232 (0.007)	0.167 (0.006)	0.541 (0.008)
800/20	0.228 (0.011)	0.166 (0.007)	0.548 (0.008)	0.238 (0.008)	0.167 (0.007)	0.536 (0.009)	0.231 (0.009)	0.166 (0.006)	0.544 (0.007)
300/40	0.227 (0.007)	0.166 (0.005)	0.546 (0.006)	0.239 (0.005)	0.168 (0.005)	0.535 (0.006)	0.232 (0.005)	0.167 (0.003)	0.542 (0.004)
800/40	0.226 (0.007)	0.166 (0.004)	0.547 (0.006)	0.239 (0.005)	0.168 (0.004)	0.535 (0.005)	0.230 (0.006)	0.166 (0.003)	0.544 (0.005)

Table 11. Correlations between true and estimated MRS scores by Count

Procedure

MRS						
Sample size/ test length	Pearson correlation			Spearman correlation		
	content items	15 uncorrelat ed items	content + 15 uncorrelat ed items	content items	15 uncorrelat ed items	content + 15 uncorrelat ed items
300/20	.336	.382	.465	.335	.368	.454
800/20	.334	.374	.460	.333	.362	.449

300/40	.385	.377	.472	.396	.364	.473
800/40	.373	.372	.464	.384	.358	.462

3.4 Discussion of Count Procedure

The most traditional and easily used approach is Count Procedure. It is easy to use in software, such as excel, SPSS, etc, because it only involves counting frequencies of responses. In addition, all kinds of RS can be calculated by counting, and it did not require adding extra items. However, researchers suggested Count Procedure performed better with heterogeneous items because content items are correlated and content factors may confound detection of RS (Bolt & Johnson, 2009). This assumption has not been examined in simulation studies. Therefore, aims of present study regarding to Count Procedure were: to evaluate performance of Count Procedure in identifying various RS through simulations; and to explore whether heterogeneous items could improve performance of Count Procedure when true model was MNRM.

First of all, baseline of various RS scores by Count Procedure in simulation studies were established. Baseline RS scores were calculated by 15 heterogeneous items when there was no RS (NO RS). ARS was around 0.350, ERS was close to 0.230, and MRS is around 0.310. RS scores did not change much as sample size and number of items changed. It was suggested even when there is no RS, RS scores are non-zero because they are proportion of responses associated with a particular RS relative to total number of all responses. For example, ARS scores are proportion of agreements relative to total number of responses. ARS scores would be a non-zero value under assumption of normal

distribution. Therefore, RS scores can be compared to these benchmark values to tell whether a certain RS is present.

When there was ARS, results of Count Procedure indicated presence of ARS in that ARS scores calculated from correlated content items (around 0.500) were higher than benchmark of ARS (0.350) calculated from 15 heterogeneous items with no RS, and also higher than ERS (around 0.250) and MRS (around 0.220). ARS scores increased slightly when using 15 uncorrelated items. Correlations between ARS scores by Count Procedure and true ARS scores were higher (about 0.240) in uncorrelated items than in correlated content items (below 0.160). Therefore, it was suggested using uncorrelated items improves performance of Count Procedure in detection of ARS slightly. However, correlations were below 0.300, so it was not advised to predict true ARS scores for individuals from those estimated from Count Procedure.

When there was ERS, results of Count Procedure indicated presence of ERS in that ERS scores calculated from correlated content items (around 0.420) were higher than benchmark of ERS (0.230) calculated from 15 heterogeneous items with no RS, and also higher than MRS (around 0.250). ARS scores increased slightly when using 15 uncorrelated items. However, difference between ARS (around 0.370) and ERS (around 0.420) was not large, and ARS may be incorrectly identified as present when there was ERS only. Correlations between ERS scores by Count Procedure and true ERS scores were higher (about 0.350) in uncorrelated items than in correlated content items (about 0.300). Furthermore, it was found correlations increased slightly with more items when all the content items and uncorrelated items were combined. Therefore, it was suggested using uncorrelated items improves performance of

Count Procedure in detection of ERS slightly. However, correlations were around 0.300, so it was not advised to predict true ERS scores for individuals from those estimated from Count Procedure.

When there was MRS, results of Count Procedure indicated presence of MRS in that MRS scores calculated from correlated content items (around 0.550) were higher than benchmark of MRS (0.310) calculated from 15 heterogeneous items with no RS, and also higher than ARS (around 0.250) and ERS (around .360). MRS scores did not change much when using 15 uncorrelated items. Correlations between MRS scores by Count Procedure and true MRS scores were higher (about 0.380) in uncorrelated items than in correlated content items (about 0.330) only in condition 300/20 and 800/20, but did not differ much when number of items increased to 40. Similar to ERS condition, it was found correlations increased slightly with more items when all the content items and uncorrelated items were combined. Therefore, it was suggested using uncorrelated items improves performance of Count Procedure in detection of MRS slightly when number of items were small. However, correlations were below 0.400, so it was not advised to predict true MRS scores for individuals from those estimated from Count Procedure.

Among four conditions, NO RS, ARS, ERS, and MRS, Count Procedure worked better for NO RS, ARS and MRS condition, relative to ERS condition. In general, correlations between estimated and true RS scores were in low range, so it was not recommended to predict individual RS scores.

Chapter 4

Results and Discussion of RIRS and RIRSMACS

4.1 Analysis model

4.1.1 Analysis model for RIRS approach

RIRS model for data analysis was presented in Figure 10 below. Figure 10a, Figure 10b, Figure 10c, and Figure 10d describes RIRS model for NO RS, ARS, ERS, and MRS respectively. Take 10b ARS model as an example, A and B represented 2 content factors, and each consisted of 10 items. 2 content factors were correlated and correlation coefficient was free to estimate. ARS, ERS, and MRS are observed scores calculated from 15 extra items. As there was no consistent findings about relationships between various RS, ARS, ERS, and MRS were treated as correlated variables. Path coefficients from ARS variables to content items were specified at the same value for all content items, and path coefficients from ERS and MRS to content items were fixed as 0 (When ERS is true model, path coefficients from ERS variables to content items were specified at the same value for all content items, and path coefficients from ARS and MRS to content items were fixed as 0; When MRS is true model, path coefficients from MRS variables to content items were specified at the same value for all content items, and path coefficients from ARS and ERS to content items were fixed as 0.) In addition, data was simulated from MNRM in which content items were treated as categorical. Therefore, content items were also treated as categorical in RIRS and RIRSMACS in order to minimize differences between data simulation model and analysis model.

Figure 10. RIRS model for NO RS, ARS, ERS, and MRS

10a. RIRS model for NO RS

Note: Observed ARS, ERS, and MRS scores were calculated from 15 extra uncorrelated items.

In RIRS model for NO RS, all the path from RS scores to items were fixed to 0, because there was no RS.

10b. RIRS model for ARS

Note: Observed ARS, ERS, and MRS scores were calculated from 15 extra uncorrelated items.

In RIRS model for RS, RS was considered as independent on item content. Therefore, relationships between a particular RS and all the items were assumed to be equal. As presented in Figure 10b, all the path from ERS and MRS to items were fixed to 0, and path from ARS to all the items were fixed to be equal to be estimated.

10c. RIRS model for ERS

Note: Observed ARS, ERS, and MRS scores were calculated from 15 extra uncorrelated items.

Similarly, in RIRS model for ERS, all the path from ARS and MRS to items were fixed to 0, and path from ERS to all the items were fixed to be equal (As shown in Figure 10c).

10d. RIRS model for MRS

Note: Observed ARS, ERS, and MRS scores were calculated from 15 extra uncorrelated items.

In RIRS model for MRS, all the path from ARS and ERS to items were fixed to 0, and path from MRS to all the items were fixed to be equal (As shown in Figure 10d).

4.1.2 Analysis model for RIRSMACS approach

RIRSMACS model for data analysis was presented in Figure 11 below. Take 11(b) for example, A and B represented 2 content factors, and each consisted of 10 items. 2 content factors were correlated and correlation coefficient was free to estimate. ARS, ERS, and MRS are latent variables

calculated from 15 extra items. Path coefficients from ARS factors to content items were specified at the same value to be estimated in model, and path coefficients from ERS, MRS to content items were fixed at 0.

Figure 11. RIRSMACS model for NO RS, ARS, ERS, and MRS

11a. RIRSMACS model for NO RS

Note: Observed ARS1-ARS3, ERS1-ERS3, and MRS1-MRS3 scores were calculated from 15 extra uncorrelated items.

Similar to RIRS, in NO RS model, all the path from RS latent factors to content items were fixed to 0.

11b. RIRSMACS model for ARS

Note: Observed ARS1-ARS3, ERS1-ERS3, and MRS1-MRS3 scores were calculated from 15 extra uncorrelated items.

Similar to RIRS, all the path from ERS and MRS to content items were fixed to 0, and all the path from ARS latent factors to content items were fixed to be equal.

11c. RIRSMACS model for ERS

Note: Observed ARS1-ARS3, ERS1-ERS3, and MRS1-MRS3 scores were calculated from 15 extra uncorrelated items.

Similar to RIRS, all the path from ARS and MRS to content items were fixed to 0, and all the path from ERS latent factors to content items were fixed to be equal.

11d. RIRSMACS model for MRS

Note: Observed ARS1-ARS3, ERS1-ERS3, and MRS1-MRS3 scores were calculated from 15 extra uncorrelated items.

Similar to RIRS, all the path from ARS and ERS to content items were fixed to 0, and all the path from MRS latent factors to content items were fixed to be equal.

4.2 Evaluation criteria

4.2.1 Model fit

For each simulation condition, four competing models were estimated: NO RS, ARS, ERS, and MRS. One evaluation criteria of model fit was number of nonconvergence out of 100 simulations each condition. It was expected that correct model would have the smallest number of nonconvergence.

The second criteria was model fit indices including AIC, BIC, and adj BIC. Models with the smallest indices were regarded as the best-fitting model.

For each model-fit index, the number of replications that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices was counted, and the model with the largest number was considered as the best-fitting model.

In addition, log likelihood difference test were employed to do pairwise model comparisons, NO RS vs. ARS, NO RS vs. ERS, and NO RS vs. MRS. According to model specification in RIRS, RIRSMACS, and MNRM, NO RS was a simpler model nested in ARS, ERS, or MRS, and differences in $-2*LL$ between NO RS and ARS follow a chi-sq distribution. Results of pairwise model comparisons indicated a better fitting model. Proportions of better fitting model were calculated for all the four models, and the one with the largest proportion was considered as the best-fitting model.

4.2.2 Parameter recovery

Correlation between 2 content factors was reported in the current study. True correlation coefficient between 2 content factors is 0.5. Therefore, it was expected that correlation in correct model would be closer to true value 0.5 relative to incorrect models.

4.3 Results of RIRS

4.3.1 Results of RIRS – NO RS condition

Performance of RIRS when there was no RS was examined in simulation studies. One evaluation criteria was model fit. First, number of nonconvergence (NC) was compared among four competing models. True

model is NO RS, so it was expected that NO RS condition had the smallest number of NC. Results showed that all simulations converged.

In addition to number of convergence in each condition, four competing models were fit to pick out the best-fitting model. NO RS model was a simpler model nested into any of the three RS models, ARS, ERS, and MRS. Therefore, LL difference test was conducted to compare NO RS vs. ARS, ERS, or MRS models. Results were reported in the form of type I error rate, that is, the probability of incorrect rejection of true model - NO RS compared to ARS, ERS, or MRS. As shown in Table 12, type I error rate was less than 8%, indicating NO RS model can be selected correctly above 92% when compared to ARS, ERS, and MRS. The result did not change as sample size increased from 300 to 800, and number of items increased from 20 to 40. This result suggested correct model – NO RS can be picked out when compared to other RS models when NO RS was the true model.

In addition, another set of model-fit indices, AIC, BIC, and adj BIC were employed to compare four competing models. It was suggested that smallest model-fit indices indicated best-fit model. As shown in Table 13, correct model – NO RS were selected the most (around 60 out of 100 replications) relative to other RS models based on AIC. However, performance of AIC in selecting the correct model did not improve much as sample size increased from 300 to 800, or as number of items increased from 20 to 40. BIC performed better than AIC to pick out the correct model – NO RS, close to 100 out of 100 replications in each condition. The results did not change much as sample size increased from 300 to 800, and not change much as number of items increased. Performance of adj BIC was better than AIC that it can select

the correct model more likely than AIC (above 70 out of 100 replications), but worse than BIC. Altogether, BIC was more reliable than AIC and adj BIC to pick out correct model when there was no RS.

Table 12. Type I error rate in rejection of true model NO RS when compared to ARS, ERS, and MRS in RIRS

Sample size/ test length	300/20	800/20	300/40	800/40
ARS vs. NO RS	0.07	0.07	0.04	0.08
ERS vs. NO RS	0.02	0.06	0.06	0.03
MRS vs. NO RS	0.05	0.06	0.06	0.07

Table 13. The number of replications that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (no RS) in RIRS

	Sample size/ test length	NO RS (true)	ARS (wrong)	ERS (wrong)	MRS (wrong)
AIC	300/20	59	17	10	14
	800/20	63	16	13	8
	300/40	64	11	12	13
	800/40	64	11	12	13
BIC	300/20	97	1	0	2

	800/20	97	1	1	1
	300/40	98	2	0	0
	800/40	98	2	0	0
adj	300/20	74	10	7	9
BIC	800/20	80	8	8	4
	300/40	73	8	9	10
	800/40	73	8	9	10

Parameter recovery was also examined for RIRS - correlations between two content factors. In condition 300/20, correlation was 0.499 for all the four competing models; in condition 800/20, correlation was 0.498 for all the four competing models; in condition 300/40, correlation was 0.502 for all the four competing models; in condition 800/40, correlation was 0.499 for all the four competing models. It was suggested this parameter did not differ in correct – NO RS model and wrong models. Correlations between two content factors did not change much as number items or sample size increased.

4.3.2 Results of RIRS – ARS condition

Performance of RIRS when there was ARS was examined in simulation studies. One evaluation criteria was model fit. First, number of nonconvergence (NC) was compared among four competing models. True model is ARS, so it was expected that ARS condition had the smallest number of NC. Results showed that all simulations converged.

Secondly, LL difference test was employed to do pairwise comparison. When compared to NO RS model, it was expected that true model - ARS would fit better. Results were reported in the form of type II error rate, that is, the probability of failure to reject of wrong model - NO RS compared to true model – ARS based on Log Likelihood difference test. As shown in Table 14, type II error rate was above 0.89, indicating correct model - ARS model cannot be selected correctly when compared to NO RS. The result did not improve as sample size increased from 300 to 800, and it did not improve as number of items increased from 20 to 40.

In addition, a set of model-fit indices, AIC, BIC, and adj BIC, were employed to compare four competing models. As shown in Table 15, incorrect model – NO RS were selected the most based on AIC (around 60 out of 100 replications) relative to correct model – ARS and other RS models. Performance of AIC in selecting the correct model did not improve as sample size increased from 300 to 800, and not improve as number of items increased from 20 to 40. BIC performed worse than AIC to pick out the correct model – ARS, as it selected the wrong model – NO RS close to 100 out of 100 replications in each condition. The results did not change much as sample size increased from 300 to 800, and not change much as number of items increased from 20 to 40. Performance of adj BIC was worse than AIC that it was more likely to select incorrect model - NO RS (above 70 out of 100 replications), but better than BIC. Altogether, when correct model was ARS, log likelihood difference test and model-fit indices were not sensitive to pick out correct model, and more likely to select NO RS as correct model.

Table 14. Type II error rate in failure to reject wrong model- NO RS when compared to true model – ARS in RIRS

Sample size/ test length	300/20	800/20	300/40	800/40
ARS vs. NO RS	0.89	0.94	0.95	0.95

Table 15. The number of replications that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (ARS) in RIRS

	Sample size/ test length	NO RS (wrong)	ARS (true)	ERS (wrong)	MRS (wrong)
AIC	300/20	57	17	19	7
	800/20	63	11	15	11
	300/40	70	16	6	8
	800/40	60	13	13	14
BIC	300/20	94	1	5	0
	800/20	96	1	2	1
	300/40	95	3	0	2
	800/40	95	0	1	4
adj	300/20	62	15	16	7
BIC	800/20	83	6	8	3

300/40	76	11	5	8
800/40	80	4	7	9

In addition to model fit, another evaluation criteria was parameter recovery – recovery of correlations between two content factors. In condition 300/20, correlation was 0.504 for all the four competing models; in condition 800/20, correlation was 0.503 for all the four competing models; in condition 300/40, correlation was 0.505 for all the four competing models; in condition 800/40, correlation was 0.500 for all the four competing models. It was suggested this parameter did not differ in correct – ARS model and wrong models. Correlations between two content factors were closer to true value 0.5 as number items and sample size increased.

4.3.3 Results of RIRS – ERS condition

Performance of RIRS when there was ERS was examined in simulation studies. One evaluation criteria was model fit. First, number of nonconvergence (NC) was compared among four competing models. True model is ERS, so it was expected that ERS condition had the smallest number of NC. Results showed that all simulations converged.

Secondly, LL difference test was employed to pick out better-fitting model. When compared to NO RS model, it was expected that true model – ERS would fit better. Results were reported in the form of type II error rate, that is, the probability of failure to reject of wrong model - NO RS compared to true model – ERS based on Log Likelihood difference test. As shown in Table 16, type II error rate was above 0.71 except for condition 800/40, indicating correct

model - ERS model cannot be selected correctly when compared to NO RS, esp. when sample size and number of items were small. The performance of LL difference test in detection of ERS improved as sample size increased from 300 to 800, and as number of items increased from 20 to 40. Only when sample size is 800, and number of items was 40, type II error rate decreased to 0.44. This result suggested it was difficult to pick out correct model – ERS compared to NO RS model, especially when sample size and number of items were small.

In addition, a set of model-fit indices, AIC, BIC, and adj BIC, were employed to compare four competing models. As shown in Table 17, when sample size was small as 300, number of items was 20, NO RS is more likely to be selected as correct model based on AIC. As sample size increased to 800, and number of items increased to 40, correct model – ERS was more likely to be identified as correct model (more than half of the replications). BIC performed worse than AIC to pick out the correct model – ERS, as it was more likely to select the wrong model – NO RS especially when sample size and number of items was small. The results did not change much as sample size increased from 300 to 800, and not change much as number of items increased. Performance of adj BIC was worse than AIC, but better than BIC to select the correct model. Altogether, when correct model was ERS, model-fit indices required large sample size and number of items in order to pick out correct model.

Table 16. Type II error rate in failure to reject wrong model- NO RS when compared to true model – ERS in RIRS

Sample size/ test length	300/20	800/20	300/40	800/40
ERS vs. NO RS	0.87	0.68	0.71	0.44

Table 17. The number of replications that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (ERS) in RIRS

	Sample size/ test length	NO RS (wrong)	ARS (wrong)	ERS (true)	MRS (wrong)
AIC	300/20	46	14	22	18
	800/20	33	3	42	12
	300/40	40	14	41	6
	800/40	21	5	55	19
BIC	300/20	90	3	2	5
	800/20	87	1	8	4
	300/40	80	1	18	1
	800/40	62	0	28	10
adj	300/20	54	11	21	14
BIC	800/20	58	6	28	8

300/40	47	11	36	6
800/40	34	2	50	14

In addition, parameter recovery - correlations between two content factors was examined. In condition 300/20, correlation was 0.488 for all the four competing models; in condition 800/20, correlation was 0.499 for all the four competing models; in condition 300/40, correlation was 0.496 for all the four competing models; in condition 800/40, correlation was 0.498 for all the four competing models. It was suggested this parameter did not differ in correct – ERS model and wrong models. Correlations between two content factors were closer to true value 0.5 as number items and sample size increased.

4.3.4 Results of RIRS – MRS condition

Performance of RIRS when there was MRS was examined in simulation studies. One evaluation criteria was model fit. First, number of nonconvergence (NC) was compared among four competing models. True model is NO RS, so it was expected that NO RS condition had the smallest number of NC. Results showed that all simulations converged.

Secondly, LL pairwise comparison was employed to identify better-fitting model. When compared to NO RS model, it was expected that true model - MRS would fit better. Results were reported in the form of type II error rate, that is, the probability of failure to reject of wrong model - NO RS compared to true model – MRS based on Log Likelihood difference test. As shown in Table 18, type II error rate was below 0.40 except for condition 300/20, indicating correct model - MRS model can be selected as true model

when compared to NO RS. Performance of LL difference test in selecting true model - MRS when compared to NO RS improved as sample size increased from 300 to 800, and number of items increased from 20 to 40. This result suggested correct model – MRS performed better than NO RS, esp. when sample size and number of items was large.

In addition, a set of model-fit indices, AIC, BIC, and adj BIC were employed to compare four competing models. As suggested in Table 19, correct model - MRS was more likely to be selected as correct model based on AIC, and performance of model-fit indices improved as sample size and number of items increased. BIC performed worse than AIC to pick out the correct model – MRS, as it was more likely to select the wrong model – NO RS especially when sample size and number of items was small. The results improved as sample size increased from 300 to 800, and number of items increased from 20 to 40. Performance of adj BIC was similar to AIC, and better than BIC to select the correct model. Altogether, when correct model was MRS, model-fit indices required large sample size and number of items to pick out correct model.

Table 18. Type II error rate in failure to reject wrong model- NO RS when compared to true model –MRS in RIRS

Sample size/ test length	300/20	800/20	300/40	800/40
MRS vs. NO RS	0.54	0.21	0.40	0.02

Table 19. The number of replications that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (MRS) in RIRS

	Sample size/ test length	NO RS (wrong)	ARS (wrong)	ERS (wrong)	MRS (true)
AIC	300/20	20	12	14	54
	800/20	9	13	2	76
	300/40	12	17	4	67
	800/40	2	1	0	97
BIC	300/20	71	4	4	21
	800/20	41	6	0	53
	300/40	51	8	1	40
	800/40	10	1	0	89
adj	300/20	31	9	12	49
BIC	800/20	16	10	1	73
	300/40	18	15	3	64
	800/40	2	1	0	97

In addition, another evaluation criteria was parameter recovery – recovery of correlations between two content factors. In condition 300/20, correlation was 0.493 for all competing models except for ARS model (0.492);

in condition 800/20, correlation was 0.500 in all the four competing models; in condition 300/40, correlation was 0.502 in all the four competing models; in condition 800/40, correlation was 0.500 in all the four competing models. It was suggested this parameter did not differ in correct – MRS model and wrong models. Correlations between two content factors were closer to true value 0.5 as number items and sample size increased.

Based on results of log likelihood difference test and model-fit indices, RIRS performed the best for NO RS condition compared to the other 3 RS conditions. Among the three RS conditions, RIRS performed best in selecting the correct model for MRS model, and worst for ARS model. In general, performance of LL difference test and model-fit indices for all the three RS conditions got better as sample size and number of items increased. For NO RS condition, BIC is more sensitive to pick out correct model. For 3 RS conditions, AIC and adj BIC are more sensitive to pick out correct model.

According to results of correlations between two content factors in NO RS, ARS, ERS, and MRS conditions above, it was suggested that performance of RIRS improved (indicated by correlations between two content factors getting closed to true value 0.5) as number of items increased from 20 to 40, and as sample size increased from 300 to 800.

4.4 Discussion of RIRS

Previous research suggested low correlations between items may help estimation of RS more accurately, so RIRS approach was developed by adding heterogeneous items to identify RS. In RIRS approach, RS variables were treated as observed which was calculated by extra uncorrelated items. However,

previous research has not examined performance of RIRS in simulation studies. Therefore, the present study aims to explore: whether correct model could be picked out based on model-fit indices; and parameter recovery that whether correlations between two content factors could be well-recovered.

Model-fit indices included Log Likelihood difference test, AIC, BIC, and adj BIC. log likelihood was obtained for each of the four competing models, NO RS, ARS, ERS, and MRS. Model NO RS was nested into ARS, ERS, and MRS, and NO RS was considered as a simpler model. Therefore, pairwise comparison between NO RS vs. ARS, ERS, and MRS were conducted to pick out better-fitting model. When NO RS was true model, results of log likelihood difference test suggested that true model in most replications (over 92 out of 100 replications) could be picked out when compared to other 3 RS models. That is, type I error rate of incorrectly rejecting true model – NO RS was smaller than 0.08. When ARS was true model, results of log likelihood difference test suggested that true model could not be selected when compared to NO RS (less than 15 out of 100 replications), so NO RS is more likely to be selected as correct model when true model is ARS. Type II error rate of failure to reject wrong models are above 0.89. When ERS was true model, results of log likelihood difference test showed that ERS model did not fit better than NO RS model when sample size and number of items were small (type II error rate was bigger than 0.68). Only in condition 800/40, true model was correctly identified in more than half replications (type II error rate was below 0.5). In MRS condition, results of log likelihood difference test suggested true model was more likely to be selected, especially when sample size and number of items were large (type II error rate was 0.02 in condition 800/40). In general,

when true model is NO RS or MRS, results of log likelihood difference test could be employed to select correct model, and performance of RIRS indicated by log likelihood difference test improved as sample size and number of items increased. However, in condition ARS and ERS, it was not recommended to select correct model by results of log likelihood difference test due to low accuracy.

Another set of model-fit indices included AIC, BIC, and adj BIC. Model with the smallest value of these indices is considered as the best-fitting model. Therefore, in each condition, the number of replications that one of the four competing models identified as the best-fitting model indicated by smallest AIC, BIC, and adj BIC were reported. In NO RS condition, BIC (correctly selected true model NO RS in 97 to 98 out of 100 replications) performed better than AIC (correctly selected true model in 59 to 64 replications) and adj BIC (correctly selected true model in 73 to 80 replications) in selecting the correct model. In ARS condition, all the three model-fit indices were not sensitive to pick out correct model (correctly selected true model ARS in less than 20 replications out of 100), and results suggested NO RS was more likely to be incorrectly picked out as correct model according to AIC, BIC, and adj BIC when true model was ARS. In ERS condition, these three model-fit indices did not perform well to detect correct RS (correctly selected true model ERS in less than 55 replications by AIC, less than 28 replications by BIC, and less than 50 replications by adj BIC). In MRS condition, true model was more likely to perform better than other competing models by AIC, BIC, and adj BIC. Among the three model-fit indices, AIC (correctly selected true model from 54 to 97 replications among 4 conditions) and adj BIC (correctly selected true model

from 49 to 97 replications among 4 conditions) performed better than BIC (correctly selected true model from 21 to 89 replications among 4 conditions). In general, BIC performed better than AIC and adj BIC in NO RS condition; whereas in MRS, AIC and adj BIC performed better than BIC. In addition, performance of these model-fit indices in detection of NO RS, ARS, ERS, and MRS improved as sample size and number of items increased.

Furthermore, parameter recovery - correlations between two content factors was another evaluation criteria for performance of RIRS. Results showed that correlations did not differ in correct models and the other three wrong models. In general, correlations between two content factors was closer to true value 0.5 as number of items increased from 20 to 40, and as sample size increased from 300 to 800.

4.5 Results of RIRSMACS

4.5.1. Results of RIRSMACS – NO RS condition

Performance of RIRSMACS when there was no RS was examined in simulation studies. One evaluation criteria was number of nonconvergence (NC) in each condition. True model is NO RS, it was expected that NO RS condition had the smallest number of NC. However, results showed that there was no big difference between NO RS to other RS in number of NC. For condition 300/20, range of number of NC was from 8 to 25; for 800/20, number of NC ranged from 0 to 10; for 300/40, number of NC ranged from 5 to 27; for 800/40, number of NC ranged from 0 to 4. As sample size increased, number of NC got smaller, but number of NC did not change much as number of items increased.

Four competing models were fit to pick out the best-fitting model. NO RS model was a simpler model nested into any of the three RS models, ARS, ERS, and MRS. Therefore, log likelihood difference test was conducted to compare NO RS vs. ARS, ERS, or MRS models. Results were reported in the form of type I error rate, that is, the probability of incorrect rejection of true model - NO RS compared to ARS, ERS, or MRS. As shown in Table 20, type I error rate was less than 0.22, indicating NO RS model can be selected correctly above 78% when compared to ARS, ERS, and MRS. The results improved as sample size increased, but not as number of items increased. This result suggested correct model – NO RS can be picked out when compared to RS models when there was no RS by Log Likelihood difference test.

In addition, model-fit indices were employed to compare four competing models. Model-fit indices included AIC, BIC, and adj BIC, and smallest model-fit indices indicated best-fitting model. As suggested by AIC, correct model – NO RS were selected the most (around 60 out of 100 replications) relative to other RS models. Performance of AIC in selecting the correct model increased as sample size increased slightly from 300 to 800, but not increased significantly when number of items increased from 20 to 40. Furthermore, there was no significant differences between RS models. BIC performed better than AIC to pick out the correct model – NO RS, especially when sample size was as large as 800, but BIC did not change much as number of items increased. Performance of adj BIC was better than AIC that it can select the correct model more likely than AIC, but worse than BIC. Performance of adj BIC improved as sample size increased from 300 to 800, but did not change much as number of items increased from 20 to 40.

Altogether, BIC was more reliable than AIC and adj BIC to pick out correct model when there was no RS.

Table 20. Type I error rate in rejection of true model NO RS when compared to ARS, ERS, and MRS in RIRSMACS

Sample size/ test length	300/20	800/20	300/40	800/40
ARS vs. NO RS	0.06	0.06	0.05	0.09
ERS vs. NO RS	0.13	0.05	0.22	0.07
MRS vs. NO RS	0.13	0.05	0.19	0.04

Table 21. The number of replications (out of total number of convergence) that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (NO RS) in RIRSMACS

	Sample size/ test length	NO RS (true)	ARS (wrong)	ERS (wrong)	MRS (wrong)
AIC	300/20	59/97	11/97	13/97	14/97
	800/20	69/100	16/100	6/100	9/100
	300/40	58/98	8/98	23/98	9/98
	800/40	68/100	16/100	8/100	8/100
BIC	300/20	83/97	1/97	5/97	8/97
	800/20	95/100	2/100	1/100	2/100

	300/40	77/98	4/98	9/98	8/98
	800/40	96/100	2/100	1/100	1/100
adj	300/20	65/97	9/97	11/97	12/97
BIC	800/20	86/100	7/100	2/100	5/100
	300/40	65/98	7/98	18/98	8/98
	800/40	80/100	10/100	6/100	4/100

In addition, correlations between two content factors were also reported in Mplus output. In condition 300/20, correlation was 0.496 for true model NO RS and ranged from 0.495 to 0.500 for other RS models. In condition 800/20, correlation was 0.498 for true model NO RS and ranged from 0.498 to 0.500 for other RS models. In condition 300/40, correlation was 0.507 in true model NO RS, and ranged from 0.503 to 0.510 in other RS models. In condition 800/40, correlation was 0.500 in true model NO RS, and ranged from 0.499 to 0.500 in other RS models. It was suggested correlation between two content factors did not differ much between true model and other 3 competing models. In general, correlations between two content factors was closer to true value 0.5 as sample size increased from 300 to 800, but not as number of items increased from 20 to 40.

4.5.2 Results of RIRSMACS – ARS condition

Performance of RIRSMAS when there was ARS was examined in simulation studies. Number of nonconvergence (NC) was compared among four

competing models. True model is ARS, it was expected that ARS condition had the smallest number of NC. However, results showed that ERS had the fewest NC (range from 0 to 8), and there was no big difference between NO RS, ARS, and MRS in number of NC. For NO RS, ARS, and MRS, number of NC was around 20 in condition 300.20; for 300/40, number of NC was around 15. In condition 800/20 and 800/40, number of NC is close to 0. Therefore, correct model – ARS did not have the smallest number of NC relative to other competing models. In general, number of NC got smaller as sample size increased, but did not change much as number of items increased.

Four competing models were fit to pick out the best-fitting model. NO RS model was a simpler model nested into RS model. Therefore, log likelihood difference test were conducted to compare NO RS vs. true model - ARS. Results were reported in the form of type II error rate, that is, the probability of failure to reject of wrong model - NO RS compared to true model – ARS based on Log Likelihood difference test. As shown in Table 22, type II error rate was above 0.88, indicating correct model - ARS model cannot be selected correctly when compared to NO RS. The result did not change much as sample size increased from 300 to 800, and as number of items increased from 20 to 40. This result suggested correct model – ARS cannot be picked out when compared to NO RS models. Compared to ERS and MRS, ARS was less likely to be selected as correct model.

In addition, a set of model-fit indices, AIC, BIC, and adj BIC were employed to compare four competing models. As suggested by AIC, incorrect model – NO RS were selected the most (around 60 out of 100 replications) relative to correct model – ARS and other RS models. Performance of AIC in

selecting the correct model did not improve as sample size increased from 300 to 800, and not improve as number of items increased from 20 to 40. BIC performed worse than AIC to pick out the correct model – ARS, as it selected the wrong model – NO RS close to 100 out of 100 replications in each condition. The results did not change much as sample size increased from 300 to 800, and not change much as number of items increased. Performance of adj BIC was worse than AIC that it was more likely to select incorrect model - NO RS (around 70 out of 100 replications), but better than BIC. Altogether, when correct model was ARS, log likelihood difference test and model-fit indices were not sensitive to pick out correct model, and was more likely to select NO RS as correct model.

Table 22. Type II error rate in failure to reject wrong model- NO RS when compared to true model – ARS in RIRSMACS

Sample size/ test length	300/20	800/20	300/40	800/40
ARS vs. NO RS	0.88	0.94	0.92	0.96

Table 23. The number of replications (out of total number of convergence) that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (ARS) in RIRSMACS

	Sample size/ test length	NO RS (wrong)	ARS (true)	ERS (wrong)	MRS (wrong)
AIC	300/20	46/98	16/98	21/98	14/98

	800/20	66/100	16/100	13/100	5/100
	300/40	61/100	14/100	18/100	7/100
	800/40	63/100	15/100	14/100	8/100
BIC	300/20	74/98	5/98	13/98	6/98
	800/20	97/100	1/100	1/100	1/100
	300/40	80/100	3/100	13/100	4/100
	800/40	98/100	0/100	2/100	0/100
adj	300/20	58/98	11/98	16/98	13/98
BIC	800/20	88/100	7/100	4/100	1/100
	300/40	68/100	10/100	16/100	6/100
	800/40	86/100	3/100	6/100	5/100

In addition, correlations between two content factors were also examined. In condition 300/20, correlation was 0.507 for true model ARS and ranged from 0.503 to 0.506 for other 3 competing models. In condition 800/20, correlation was 0.503 for true model ARS and ranged from 0.503 to 0.504 for other 3 competing models. In condition 300/40, correlation was 0.505 in true model ARS, and ranged from 0.505 to 0.507 in other 3 competing models. In condition 800/40, correlation was 0.500 in true model ARS, and 0.500 in other 3 competing models. It was suggested correlation between two content factors did not differ much between true model and other 3 competing models. In general, correlations between two content factors was closer to true value 0.5 as

sample size increased from 300 to 800 and as number of items increased from 20 to 40.

4.5.3 Results of RIRSMACS – ERS condition

Performance of RIRSMACS when there was ERS was examined in simulation studies. Number of nonconvergence (NC) was compared among four competing models. True model is ERS, it was expected that ERS condition had the smallest number of NC. Results showed that ERS had smallest number (less than 4) of NC, but there was no big difference from other models, and number of NC ranged from 0 to 16 in all conditions. Generally, number of NC got smaller as sample size increased, but did not change much as number of items increased.

Four competing models were fit to pick out the best-fitting model. NO RS model was a simpler model nested into RS models. Therefore, log likelihood difference test was conducted to compare NO RS vs. true model - ERS. Results were reported in the form of type II error rate, that is, the probability of failure to reject wrong model - NO RS compared to true model – ERS based on Log Likelihood difference test. As shown in Table 24, type II error rate was above 0.66 except for condition 800/40, indicating correct model - ERS model cannot be selected correctly when compared to NO RS. The performance of RIRSMACS in detecting ERS improved as sample size increased from 300 to 800, and improved as number of items increased from 20 to 40. Only when sample size is 800, and number of items was 40, type II error rate decreased to 0.41. This result suggested correct model – ERS performed better than NO RS only when sample size and number of items were large.

In addition, a set of model-fit indices, AIC, BIC, and adj BIC were employed to compare four competing models. As suggested by AIC, in condition 300/20, NO RS is more likely to be selected as correct model. As sample size and number of items increased, correct model – ERS was more likely to be identified as correct model, but still less than half of 100 replications. BIC performed worse than AIC to pick out the correct model – ERS, as it was more likely to select the wrong model – NO RS. The results did not change much as sample size and number of items increased. Performance of adj BIC was worse than AIC, but better than BIC to select the correct model. Altogether, when correct model was ERS, model-fit indices required large sample size and number of items to pick out correct model, but ERS did not perform better than NO RS and MRS.

Table 24. Type II error rate in failure to reject wrong model- NO RS when compared to true model – ERS in RIRSMACS

Sample size/ test length	300/20	800/20	300/40	800/40
ERS vs. NO RS	0.81	0.66	0.66	0.41

Table 25. The number of replications (out of total number of convergence) that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (ERS) in RIRSMACS

	Sample size/ test length	NO RS (wrong)	ARS (wrong)	ERS (true)	MRS (wrong)
AIC	300/20	46/100	10/100	24/100	19/100
	800/20	43/100	10/100	29/100	18/100
	300/40	30/100	13/100	26/100	31/100
	800/40	26/100	5/100	42/100	27/100
BIC	300/20	85/99	3/99	5/99	6/99
	800/20	81/100	1/100	10/100	8/100
	300/40	70/100	1/100	12/100	17/100
	800/40	60/100	0/100	21/100	19/100
adj	300/20	52/99	9/99	20/99	18/99
BIC	800/20	57/100	8/100	22/100	14/100
	300/40	44/100	8/100	23/100	25/100
	800/40	36/100	4/100	36/100	24/100

In addition, correlations between two content factors were also examined. In condition 300/20, correlation was 0.488 for true model ERS and ranged from 0.488 to 0.489 for other 3 competing models. In condition 800/20,

correlation was 0.499 for true model ERS and 0.499 for other 3 competing models. In condition 300/40, correlation was 0.495 in true model ERS, and ranged from 0.497 to 0.498 in other 3 competing models. In condition 800/40, correlation was 0.498 in true model ERS, and ranged from 0.497 to 0.498 in other 3 competing models. It was suggested correlation between two content factors did not differ much between true model and other 3 competing models. In general, correlations between two content factors was closer to true value 0.5 as sample size increased from 300 to 800 and as number of items increased from 20 to 40.

4.5.4 Results of RIRSMACS – MRS condition

Performance of RIRSMACS when there was MRS was examined in simulation studies. Number of nonconvergence (NC) was compared among four competing models. True model is MRS, it was expected that ERS condition had the smallest number of NC. Results showed that there was no big difference between MRS and other models, because number of NC was less than 3 in all conditions.

Four competing models were fit to pick out the best-fitting model. NO RS model was a simpler model nested into RS models. Therefore, log likelihood difference test was conducted to compare NO RS vs. true model - MRS. Results were reported in the form of type II error rate, that is, the probability of failure to reject of wrong model - NO RS compared to true model - MRS based on Log Likelihood difference test. As shown in Table 26, type II error rate was below 0.30 except for condition 300/20, indicating correct model - MRS model can be selected as true model when compared to NO RS.

Performance of log likelihood difference test in selecting MRS compared to NO RS as sample size increased from 300 to 800, and number of items increased from 20 to 40. This result suggested correct model – MRS performed better than NO RS, esp. when sample size and number of items was large.

In addition, a set of model-fit indices, AIC, BIC, and adj BIC were employed to compare four competing models. As suggested by AIC, correct model - MRS was more likely to be selected as correct model (selected as best-fitting model in 40 to 84 replications), and performance of AIC improved as sample size increased, but did not improve much as number of items increased. BIC performed worse than AIC to pick out the correct model – MRS, as it was more likely to select the wrong model – NO RS especially when sample size and number of items was small. The results improved as sample size and number of items increased. Performance of adj BIC was similar to AIC, and better than BIC to select the correct model. Altogether, when correct model was MRS, model-fit indices required large sample size and number of items to pick out correct model with RIRSMACS approach.

Based on results of log likelihood difference test and model-fit indices, RIRSMACS performed the best for NO RS condition compared to RS conditions. Among the three RS conditions, RIRSMACS was best for MRS, and worst for ARS. Performance of log likelihood difference test and model-fit indices in detecting RS improved as sample size and number of items increased.

Table 26. Type II error rate in failure to reject wrong model- NO RS when compared to true model – MRS in RIRSMACS

Sample size/ test length	300/20	800/20	300/40	800/40
MRS vs. NO RS	0.53	0.29	0.02	0.09

Table 27. The number of replications (out of total number of convergence) that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (MRS) in RIRSMACS

	Sample size/ test length	NO RS (wrong)	ARS (wrong)	ERS (wrong)	MRS (true)
AIC	300/20	26/100	8/100	26/100	40/100
	800/20	14/100	7/100	11/100	68/100
	300/40	19/100	11/100	24/100	46/100
	800/40	3/100	1/100	12/100	84/100
BIC	300/20	69/100	3/100	9/100	19/100
	800/20	48/100	4/100	8/100	13/100
	300/40	54/100	4/100	13/100	29/100
	800/40	16/100	0/100	12/100	72/100
adj	300/20	32/100	7/100	25/100	36/100
BIC	800/20	24/100	6/100	10/100	60/100

300/40	21/100	11/100	24/100	44/100
800/40	7/100	1/100	12/100	80/100

In addition, correlations between two content factors were also examined. In condition 300/20, correlation was 0.492 for true model MRS and ranged from 0.491 to 0.493 for other 3 competing models. In condition 800/20, correlation was 0.499 for true model MRS and ranged from 0.499 to 0.500 for other 3 competing models. In condition 300/40, correlation was 0.502 in true model MRS, and ranged from 0.501 to 0.502 in other 3 competing models. In condition 800/40, correlation was 0.500 in true model MRS, and 0.500 in other 3 competing models. It was suggested correlation between two content factors did not differ much between true model and other 3 competing models. In general, correlations between two content factors was closer to true value 0.5 as sample size increased from 300 to 800 and as number of items increased from 20 to 40.

Based on results of log likelihood difference test and model-fit indices, RIRSMACS performed the best for NO RS condition compared to the other 3 RS conditions. Among the three RS conditions, RIRSMACS performed best in selecting the correct model for MRS model, and worst for ARS model. In general, performance of log likelihood difference test and model-fit indices in identifying three types of RS improved as sample size and number of items increased. For NO RS condition, BIC is more sensitive to pick out correct model. For 3 RS conditions, AIC and adj BIC are more sensitive to pick out correct model.

According to results of correlations between two content factors in NO RS, ARS, ERS, and MRS conditions above, it was suggested that performance of RIRSMACS improved (indicated by correlations between two content factors getting closed to true value 0.5) as number of items increased from 20 to 40, and as sample size increased from 300 to 800.

4.6 Discussion of RIRSMACS

Different from RIRS approach, RS variables in RIRSMACS were treated as latent variables which were measured by extra uncorrelated items. The present study further explored performance of RIRSMACS in detection of various RS through simulations, including: to explore whether correct model could be picked out based on model-fit indices; to explore parameter recovery that whether correlations between two content factors could be well-recovered.

Model-fit indices included Log Likelihood (log likelihood) difference test, AIC, BIC, and adj BIC. log likelihood was obtained for each of the four competing models, NO RS, ARS, ERS, and MRS. Model NO RS was nested into ARS, ERS, and MRS, and NO RS was considered as a simpler model. Therefore, pairwise comparison between NO RS vs. ARS, ERS, and MRS were conducted to pick out better-fitting model. When NO RS was true model, results of log likelihood difference test suggested that true model in most replications could be picked out when compared to other 3 RS models, indicated by type I error rate below 0.22. When ARS was true model, results of log likelihood difference test suggested that true model could not be selected when compared to NO RS, indicated by type I error rate above 0.88, so NO RS is more likely to be selected as correct model when true model is ARS. When

ERS was true model, results of log likelihood difference test showed that true model did not fit better than NO RS model when sample size and number of items were small, indicated by type I error rate above 0.66 in condition 300/20, 800/20, and 300/40. Only in condition 800/40, true model was correctly identified in more than half replications (type I error rate was 0.41). In MRS condition, results of log likelihood difference test suggested true model was more likely to be selected, especially when sample size and number of items were large (type I error rate was below 0.29 except for 0.53 in condition 300/20). In general, when there was NO RS or MRS, results of log likelihood difference test could be employed to select correct model, and performance of RIRSMACS indicated by log likelihood difference test improved as sample size and number of items increased. However, in condition ARS and ERS, it was not recommended to select correct model by results of log likelihood difference test due to low accuracy.

Another set of model-fit indices included AIC, BIC, and adj BIC. Model with the smallest value of these indices is considered as the best-fitting model. Therefore, in each condition, the number of replications that one of the four competing models identified as the best-fitting model indicated by smallest AIC, BIC, and adj BIC were reported. In NO RS condition, BIC (correctly selected true model NO RS in 83 to 96 replications) performed better than AIC (correctly selected true model in 58 to 69 replications) and adj BIC (correctly selected true model in 65 to 86 replications) in selecting the correct model. In ARS condition, all the three model-fit indices were not sensitive to pick out correct model (correctly selected true model ARS in less than 16 replications), and results suggested NO RS was more likely to be incorrectly picked out as

correct model according to AIC, BIC, and adj BIC when true model was ARS. In ERS condition, these three model-fit indices did not perform well to detect correct RS (correctly selected true model ERS in less than 42 replications by AIC, less than 21 replications by BIC, and less than 36 replications by adj BIC). In MRS condition, true model was more likely to perform better than other competing models by AIC, BIC, and adj BIC. Among the three model-fit indices, AIC (correctly selected true model from 40 to 84 replications among 4 conditions) and adj BIC (correctly selected true model from 36 to 80 replications among 4 conditions) performed better than BIC (correctly selected true model from 19 to 72 replications among 4 conditions). In general, model-fit indices AIC, BIC, and adj BIC were more reliable to identify NO RS and MRS relative to ARS and ERS in RIRSMACS approach. In NO RS condition, BIC performed better than AIC and adj BIC, whereas in MRS, AIC and adj BIC performed better than BIC. In addition, performance of these model-fit indices for NO RS, ARS, ERS, and MRS improved as sample size and number of items increased.

Furthermore, parameter recovery - correlations between two content factors was another evaluation criteria for performance of RIRSMACS. Results showed that correlations did not differ in correct models and the other three wrong models. In general, correlations between two content factors were closer to true value 0.5 as number of items increased from 20 to 40, and as sample size increased from 300 to 800.

Chapter 5

Results and Discussion of MNRM

5.1 Analysis model

MNRM model for data analysis was presented in Figure 12 below. A and B represented 2 content factors, and each consisted of 10 items. 2 content factors were correlated and correlation coefficient was free to estimate. ARS/ERS/MRS was a latent variable. Path coefficients from ARS/ERS/MRS variables to content items were free to estimate.

Figure 12. MNRM model for NO RS, ARS, ERS, and MRS

12a. MNRM model for No RS

12b. MNRM model for ARS

12c. MNRM model for ERS

12d. MNRM model for MRS

5.2 Model identification

Model was analyzed in Mplus. It was noted that mean of latent variables were fixed at 0 by default in Mplus (Muthén & Muthén, 2013). However, true value of RS was specified at 1 when data was simulated. Therefore, it was necessary to add constraints in Mplus input so that mean of RS was consistent with that in data simulation model. In addition, if intercept terms in equation (7) were free to estimate, mean of RS is required to be fixed at a certain value so that MNRM can be identified. Hence, mean of RS was fixed at true value – 1 when MNRM was analyzed in Mplus.

In empirical data analyses, true values of RS are unknown. Therefore, a few values of means of RS were selected to fit models in order to find out the best-fit model, using a ‘grid’ search approach. Three example datasets were randomly selected from condition 800/20 with ARS, 800/20 with ERS, and 800/20 with MRS to test whether model with correct mean can be picked out as

the best-fit model. As true mean of RS was 1, so a set of 7 means of RS bigger or smaller than true mean of RS were tested: -0.5, 0, 0.5, 1, 1.5, 2, and 2.5.

For dataset from condition 800/20 with ARS, 7 MNRM models were fitted as mean of ARS varies. Figure 13 below presented squared difference between true and estimated intercept term in these 7 models. It was suggested model with correct mean of ARS had the smallest squared difference values, indicating model with true mean of ARS fit the best.

Figure 13. Squared difference between true and estimated intercept term in ARS

For dataset from condition 800/20 with ARS, 7 MNRM models were fitted as mean of ERS varies. Figure 14 below presented squared difference between true and estimated intercept term in these 7 models. It was suggested model with correct mean of ERS had the smallest squared difference values, indicating model with true mean of ERS fit the best.

Figure 14. Squared difference between true and estimated intercept term in ERS

For dataset from condition 800/20 with ARS, 7 MNRM models were fitted as mean of MRS varies. Figure 15 below presented squared difference between true and estimated intercept term in these 7 models. It was suggested model with correct mean of MRS had the smallest squared difference values, indicating model with true mean of MRS fits the best.

Figure 15. Squared difference between true and estimated intercept term in MRS

In the present study, mean of the RS trait (i.e. ARS, ERS, and MRS) was fixed to true value – 1 in data analysis so that MNRM can be identified to detect RS.

5.3 Evaluation criteria

For MNRM, evaluation criteria included: model fit indices (AIC, BIC, adj BIC, Log Likelihood difference test) calculated for 4 models (ARS, ERS, MRS, and NO RS); Parameter recovery including: intercept parameter, variance of RS, and correlation between two content factors. As shown in model (7) and model (8), intercept parameter would affect probability of endorsing response categories, so it is important to accurately estimate intercept in order to pick out correct model. For each simulation condition, mean of squared difference between true and estimated intercepts were calculated. It was expected that correct model is supposed to generate smallest mean of squared difference.

True value of RS variance is 0.0625. It was expected that variance of RS is closer to true value than in wrong models. Similar to RIRS and RIRSMACS, true correlation between two content factors is 0.5, so it was expected that this correlation would be better recovered in correct model.

5.4 Results of MNRM

5.4.1 Results of MNRM – NO RS condition

Performance of MNRM when there was no RS was examined in simulation studies. One evaluation criteria was model fit. First, number of nonconvergence (NC) was compared among four competing models. True model is NO RS, so it was expected that NO RS condition had the smallest number of NC. Results showed that all simulations converged.

The second criteria of model fit is differences in log likelihood. Among four competing models, NO RS model was a simpler model nested into any of the three RS models, ARS, ERS, and MRS. Therefore, log likelihood difference test was employed to compare NO RS vs. ARS, ERS, or MRS models. Results were reported in the form of type I error rate, that is, the probability of incorrect rejection of true model - NO RS compared to ARS, ERS, or MRS. As shown in Table 28, type I error rate was less than 0.04, indicating NO RS model can be selected correctly above 96% when compared to ARS, ERS, and MRS. The result did not change much as sample size increased from 300 to 800, and as number of items increased from 20 to 40. This result suggested correct model – NO RS can be picked out when compared to RS models when there was no RS.

In addition, a set of model-fit indices including AIC, BIC, and adj BIC were employed to compare four competing models. Smallest model-fit indices

indicted best-fit model. As shown in Table 29, correct model – NO RS were selected the most (above 80 out of 100 replications) relative to other RS models based on AIC. Performance of AIC in selecting the correct model increased as number of items increased from 20 to 40, but not as sample size increased from 300 to 800. BIC performed better than AIC to pick out the correct model – NO RS (close to 100 out of 100 for every condition). Performance of adj BIC was better than AIC in that it can select the correct model more likely than AIC, but worse than BIC. Performance of adj BIC improved as sample size increased from 300 to 800, and as number of items increased from 20 to 40. Altogether, BIC was more reliable than AIC and adj BIC to pick out correct MNRM model when there was no RS.

Table 28. Type I error rate in rejection of true model NO RS when compared to ARS, ERS, and MRS in MNRM

Sample size/ test length	300/20	800/20	300/40	800/40
ARS vs. NO RS	0.02	0.03	0.03	0.00
ERS vs. NO RS	0.02	0.00	0.00	0.03
MRS vs. NO RS	0.04	0.01	0.03	0.00

Table 29. The number of replications that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (NO RS) in RIRS

	Sample size/ test length	NO RS (true)	ARS (wrong)	ERS (wrong)	MRS (wrong)
AIC	300/20	82	5	7	6
	800/20	83	9	3	5
	300/40	92	3	1	4
	800/40	90	2	6	2
BIC	300/20	98	0	1	1
	800/20	99	1	0	0
	300/40	98	1	0	1
	800/40	100	0	0	0
adj	300/20	84	5	5	6
BIC	800/20	95	3	1	1
	300/40	95	2	1	2
	800/40	96	0	3	1

Parameter recovery was also examined in MNRM. The first parameter examined was intercept term in equation (8). Means of squared difference between estimated and true intercepts were calculated for each condition. For

example, there was 20 items in total in condition 300/20, and each item had 4 intercepts, so in total there was 80 parameters in condition 300/20; whereas in condition 300/40, there was 160 parameters in total for 40 items. Results of means of squared difference were presented in Table 30. It was shown that means of squared difference in correct model – NO RS was the smallest, and it got larger in the other three wrong models, about 10 times of that in correct model. In true model NO RS, means of squared difference were smaller as sample size increased, but not much as number of items increased.

Table 30. Means of squared difference in intercept parameter among NO RS (true model), ARS, ERS, and MRS models

Sample size/ test length	300/20	800/20	300/40	800/40
NO RS (true)	0.059	0.022	0.058	0.021
ARS (wrong)	0.550	0.516	0.553	0.534
ERS (wrong)	0.565	0.525	0.569	0.536
MRS (wrong)	0.307	0.271	0.304	0.274

The second parameter was correlations between two content factors. In condition 300/20, correlation was 0.499 for true model NO RS and ranged from 0.497 to 0.499 for other 3 competing models. In condition 800/20, correlation was 0.498 for true model NO RS and ranged from 0.497 to 0.498 for other 3 competing models. In condition 300/40, correlation was 0.502 in true model NO RS, and ranged from 0.502 to 0.513 in other 3 competing models. In

condition 800/40, correlation was 0.499 in true model NO RS, and 0.499 in other 3 competing models. It was suggested correlation between two content factors did not differ much between true model and other 3 competing models. In general, correlations between two content factors was closer to true value 0.5 as sample size increased from 300 to 800 and as number of items increased from 20 to 40.

5.4.2 Results of MNRM – ARS condition

Performance of MNRM in ARS condition was examined in simulation studies. One evaluation criteria was model fit. First, number of nonconvergence (NC) was compared among four competing models. As the true model was ARS, it was expected that ARS condition had the smallest number of NC. Results showed that all simulations converged.

Secondly, four competing models, NO RS, ARS, ERS, and MRS were fit to pick out the best-fitting model. Therefore, log likelihood difference test were employed to compare NO RS vs. ARS, ERS, or MRS models. Results were reported in the form of type II error rate, that is, the probability of failure to reject of wrong model - NO RS compared to true model – ARS based on Log Likelihood difference test. As shown in Table 31, type II error rate was above 0.58 except for condition 800/40 (type II error rate was 0.20), indicating correct model - ARS model was less likely to be selected as the correct model, especially when sample size and number of items was small. In general, performance of log likelihood difference test increased as number of items increased from 20 to 40, and as sample size increased from 300 to 800.

In addition, a set of model-fit indices, AIC, BIC, and adj BIC were employed to compare four competing models. As shown in Table 32, compared to wrong models NO RS, ERS, and MRS, correct model ARS was more likely to be selected as best-fitting model based on AIC. However, there was only slight difference between NO RS and ARS in AIC, especially when sample size and number of items were small. Therefore, among three wrong models, NO RS was more likely to be selected as true model. BIC performed worse than AIC to pick out the correct model – ARS (less than 20 out of 100 for every condition in three conditions), and more likely to pick out NO RS as best-fitting model. Performance of adj BIC was better than BIC in that it can select the correct model more likely than BIC (close to 50 out of 100 replications), but worse than AIC. Altogether, correct model – ARS can be picked out compared to other wrong models based on model-fit indices only when sample size and number of items were large. As sample size and number of items increased, it was more likely to identify ARS with model-fit indices.

Table 31. Type II error rate in failure to reject wrong model- NO RS when compared to true model – ARS in MNRM

Sample size/ test length	300/20	800/20	300/40	800/40
ARS vs. NO RS	0.80	0.70	0.58	0.20

Table 32. The number of replications that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (ARS) in MNRM

	Sample size/ test length	NO RS (wrong)	ARS (true)	ERS (wrong)	MRS (wrong)
AIC	300/20	47	33	5	15
	800/20	35	46	8	11
	300/40	35	49	5	11
	800/40	2	85	2	11
BIC	300/20	84	10	4	2
	800/20	85	13	2	0
	300/40	74	18	1	7
	800/40	33	56	1	10
adj	300/20	55	30	5	10
BIC	800/20	62	29	4	5
	300/40	43	43	4	10
	800/40	14	75	1	10

Parameter recovery was also examined in MNRM. The first parameter examined was intercept term in equation (9). Results of means of squared difference were presented in Table 33. It was shown that means of squared

difference in correct model – ARS was the smallest, and it got larger in the other three wrong models, about 10 times of that in correct model. In true model ARS, means of squared difference were smaller as sample size increased, but not much as number of items increased.

Table 33. Means of squared difference in intercept parameter among NO RS, ARS (true model), ERS, and MRS models

Sample size/ test length	300/20	800/20	300/40	800/40
NO RS (wrong)	0.570	0.542	0.545	0.537
ARS (true)	0.063	0.023	0.062	0.024
ERS (wrong)	0.575	0.541	0.572	0.534
MRS (wrong)	0.820	0.787	0.805	0.775

The second parameter was variance of RS factor in correct model - ARS. Results showed that variance of estimated ARS factor in 300/20, 800/20, 300/40, and 800/40 conditions was 0.088, 0.061, 0.064, and 0.061. True variance of ARS was 0.0625. It was found that variance of ARS was close to true value as number of items and sample size increased.

The third parameter was correlations between two content factors. In condition 300/20, correlation was 0.501 for true model ARS and 0.504 for other 3 competing models. In condition 800/20, correlation was 0.501 for true model ARS and 0.503 for other 3 competing models. In condition 300/40, correlation was 0.503 in true model ARS, and ranged from 0.505 to 0.506 in other 3

competing models. In condition 800/40, correlation was 0.498 in true model ARS, and 0.500 in other 3 competing models. It was suggested correlation between two content factors did not differ much between true model and other 3 competing models. In general, correlations between two content factors was closer to true value 0.5 as sample size increased from 300 to 800 and as number of items increased from 20 to 40.

5.4.3 Results of MNRM – ERS condition

Performance of MNRM when there was ERS was examined in simulation studies. Mode fit was one evaluation criteria, which includes number of nonconvergence (NC) compared among four competing models. True model was ERS, so it was expected that ERS condition had the smallest number of NC. Results showed that all simulations converged.

Secondly, log likelihood difference test was conducted to compare NO RS vs. ARS, ERS, or MRS models. Results were reported in the form of type II error rate, that is, the probability of failure to reject of wrong model - NO RS compared to true model – ERS based on Log Likelihood difference test. As shown in Table 34, type II error rate was below 0.13 except for condition 300/20 (type II error rate was 0.54), indicating correct model - ERS model can be selected correctly when compared to NO RS, especially when sample size and number of items were large. The performance of log likelihood difference test improved as sample size increased from 300 to 800, and improved as number of items increased from 20 to 40.

In addition, a set of model-fit indices, AIC, BIC, and adj BIC were employed to pick out the best-fitting model. As shown in Table 35, correct

model - ERS was more likely to be selected as the best-fit model (close to 100 out of 100 replications) relative to other competing models based on AIC. Performance of model-fit indices in selecting the correct model increased as number of items increased from 20 to 40, and as sample size increased from 300 to 800. BIC performed worse than AIC to pick out the correct model – ERS, but still can pick out correct model ERS in most conditions. Performance of adj BIC was slightly better than BIC in that it can select the correct model more likely than BIC, but worse than AIC. Altogether, correct model – ERS was more likely to be picked out compared to other wrong models based on model-fit indices. As sample size and number of items increased, it was more likely to identify ERS with model-fit indices.

Table 34. Type II error rate in failure to reject wrong model- NO RS when compared to true model – ERS in MNRM

Sample size/ test length	300/20	800/20	300/40	800/40
ERS vs. NO RS	0.54	0.13	0.02	0.00

Table 35. The number of replications that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (ERS) in MNRM

Sample size/ test length	NO RS (wrong)	ARS (wrong)	ERS (true)	MRS (wrong)
AIC 300/20	26	6	64	4

	800/20	6	2	90	2
	300/40	2	1	97	0
	800/40	0	0	100	1
BIC	300/20	68	1	31	0
	800/20	28	1	69	2
	300/40	8	1	91	0
	800/40	1	0	99	0
adj	300/20	31	6	60	3
BIC	800/20	11	2	85	2
	300/40	2	1	97	0
	800/40	0	0	100	0

Parameter recovery was also examined in MNRM. The first parameter examined was intercept term in equation (10). Results of means of squared difference were presented in Table 36. It was shown that means of squared difference in correct model – ERS was the smallest, and it got larger in the other three wrong models, about 10 times of that in correct model. In true model ERS, means of squared difference were smaller as sample size increased, but not much as number of items increased.

Table 36. Mean of squared difference in intercept parameter among NO RS, ARS, ERS (true model), and MRS models

Sample size/ test length	300/20	800/20	300/40	800/40
NO RS (wrong)	0.557	0.529	0.553	0.524
ARS (wrong)	0.581	0.526	0.555	0.537
ERS (true)	0.060	0.022	0.059	0.024
MRS (wrong)	0.813	0.780	0.802	0.775

The second parameter was variance of RS factor in correct model - ERS. Results showed that variance of estimated ERS factor in 300/20, 800/20, 300/40, and 800/40 conditions was .064, .061, .071, and .061. True variance of ERS was 0.0625. It was found that variance of ERS was close to true value as sample size increased, but not much as number of items increased.

The third parameter was correlations between two content factors. In condition 300/20, correlation was 0.490 for true model ERS and ranged from 0.487 to 0.489 for other 3 competing models. In condition 800/20, correlation was 0.500 for true model ERS and ranged from 0.498 to 0.499 for other 3 competing models. In condition 300/40, correlation was 0.499 in true model ERS, and 0.496 in other 3 competing models. In condition 800/40, correlation was 0.501 in true model ERS, and 0.498 in other 3 competing models. It was suggested correlation between two content factors did not differ much between true model and other 3 competing models. In general, correlations between two

content factors was closer to true value 0.5 as sample size increased from 300 to 800 and as number of items increased from 20 to 40.

5.4.4 Results of MNRM – MRS condition

Performance of MNRM when there was MRS was examined in simulation studies. Model fit was reported to evaluate performance of MNRM. One criteria was number of nonconvergence (NC) was compared among four competing models. As the true model was MRS, it was expected that MRS condition had the smallest number of NC. Results showed that all simulations converged.

In addition, four competing models, NO RS, ARS, ERS, and MRS were fit to pick out the best-fitting model. Therefore, log likelihood difference test were employed to compare NO RS vs. ARS, ERS, or MRS models. Results were reported in the form of type II error rate, that is, the probability of failure to reject of wrong model - NO RS compared to true model – MRS based on Log Likelihood difference test. As shown in Table 37, type II error rate was below 0.06 except for condition 300/20 (type II error rate was 0.41), indicating correct model - MRS model can be selected as true model when compared to NO RS. Performance of log likelihood difference test in selecting MRS compared to NO RS as sample size increased from 300 to 800, and number of items increased from 20 to 40. This result suggested correct model – MRS performed better than NO RS, esp. when sample size and number of items was large.

Furthermore, a set of model-fit indices, AIC, BIC, and adj BIC were employed to compare four competing models. As shown in Table 38, correct

model - MRS was more likely (above 80 out of 100 replications except for condition 300/20, 64 out of 100 replications can be picked out as correct model) to be selected as the best-fit model relative to other competing models based on AIC. Performance of model-fit indices in selecting the correct model increased as number of items increased from 20 to 40, and as sample size increased from 300 to 800. BIC performed worse than AIC to pick out the correct model, but in most conditions, MRS fit better than other 3 competing models. Performance of adj BIC was better than BIC in that it can select the correct model more likely than BIC, but worse than AIC. Altogether, correct model – MRS was more likely to be picked out compared to other wrong models based on model-fit indices. As sample size and number of items increased, it was more likely to identify MRS with model-fit indices.

Table 37. Type II error rate in failure to reject wrong model- NO RS when compared to true model – MRS in MNRM

Sample size/ test length	300/20	800/20	300/40	800/40
MRS vs. NO RS	0.41	0.06	0.02	0.00

Table 38. The number of replications that one of the four competing models identified as the best-fitting model indicated by smallest model-fit indices (MRS) in MNRM

	Sample size/ test length	NO RS (wrong)	ARS (wrong)	ERS (wrong)	MRS (true)
AIC	300/20	19	14	3	64
	800/20	3	10	0	87
	300/40	0	8	1	91
	800/40	0	2	0	98
BIC	300/20	61	5	0	34
	800/20	17	8	0	75
	300/40	8	6	1	85
	800/40	0	2	0	98
adj	300/20	22	14	1	63
BIC	800/20	5	10	0	85
	300/40	0	8	1	91
	800/40	0	2	0	98

Parameter recovery was also examined in MNRM. The first parameter examined was intercept term in equation (11). Results of means of squared difference were presented in Table 39. It was shown that means of squared

difference in correct model – MRS was the smallest, and it got larger in the other three wrong models, about 10 times of that in correct model. In true model MRS, means of squared difference were smaller as sample size increased, but not much as number of items increased.

Table 39. Mean of squared difference in intercept parameter among NO RS, ARS, ERS, and MRS (true model) models

Sample size/ test length	300/20	800/20	300/40	800/40
NO RS (wrong)	0.330	0.243	0.332	0.278
ARS (wrong)	0.864	0.801	0.846	0.818
ERS (wrong)	0.860	0.799	0.857	0.800
MRS (true)	0.079	0.029	0.077	0.030

The second parameter was variance of RS factor in correct model - MRS. Results showed that variance of estimated MRS factor in 300/20, 800/20, 300/40, and 800/40 conditions was 0.057, 0.059, 0.062, and 0.063. True variance of MRS was 0.0625. It was found that variance of MRS was close to true value as sample size and number of items increased.

The third parameter was correlations between two content factors. In condition 300/20, correlation was 0.493 for true model MRS and ranged from 0.489 to 0.493 for other 3 competing models. In condition 800/20, correlation was 0.501 for true model MRS and ranged from 0.497 to 0.500 for other 3 competing models. In condition 300/40, correlation was 0.503 in true model

MRS, and ranged from 0.500 to 0.502 in other 3 competing models. In condition 800/40, correlation was 0.502 in true model MRS, and ranged from 0.499 to 0.500 in other 3 competing models. It was suggested correlation between two content factors did not differ much between true model and other 3 competing models. In general, correlations between two content factors was closer to true value 0.5 as sample size increased from 300 to 800 and as number of items increased from 20 to 40.

5.5 Discussion

MNRM was one of the IRT models which receive most attention in previous research on RS. However, there are some limitations in previous research on MNRM, and there is lack of systematic simulation studies on performance of MNRM. Therefore, present study addressed limitations in previous research and further explored performance of MNRM in identifying RS. First, previous simulation research only examined one RS – ERS as an example to testify performance of MNRM. Therefore, it remains unknown whether MNRM worked similarly for other types of RS. Secondly, in previous research, ERS was simulated from a normal distribution $N(0, 1)$. However, it was noted that only positive ERS values are consistent with definition of ERS, that is, the tendency to choose extreme response styles. Conversely, negative ERS values refer to another response styles MLRS (mild RS). This indicated that in previous research, two RS – ERS and MLRS were present in simulated dataset, so it was still unclear that whether MNRM performed well when there was only one RS.

The present study was designed to address limitations above, and to explore: 1) whether MNRM could pick out correct model with model-fit indices when there is only one RS, ARS, ERS, or MRS; 2) parameter recovery, including intercept parameters, variance of RS (true value: 0.0625), and correlations between two content factors (true value: 0.5).

Overall, there was no nonconvergence for all the simulation conditions. Model-fit indices, Log Likelihood difference test, AIC, BIC, and adj BIC were examined to pick out best-fitting model among four competing models, NO RS, ARS, ERS, and MRS. Model NO RS was nested into ARS, ERS, and MRS, and NO RS was considered as a simpler model. Therefore, pairwise comparison between NO RS vs. ARS, ERS, and MRS were conducted to pick out better-fitting model. When NO RS was true model, results of log likelihood difference test suggested that true model in most replications (over 96 out of 100 replications) could be picked out when compared to other 3 RS models, indicating type I error rate in rejection of NO RS was below 0.04. When ARS was true model, results of log likelihood difference test suggested that true model could only be identified in condition 800/40 (80 out of 100 replications). That is, type II error rate in failure to reject wrong model NO RS was 0.20 for condition 800/40. In other conditions, correct model – ARS could only be selected for less than half replications (type II error rate was above 0.5). When ERS was true model, results of log likelihood difference test showed that true model can be more likely to be picked out in condition 800/20, 300/40, and 800/40 (type II error rate was below 0.13), but only in condition 300/20 with smaller sample size and number of items, correct model is less likely to be selected compared to wrong model NO RS. In MRS condition, results of log

likelihood difference test suggested true model was more likely to be selected, especially when sample size and number of items were large (type II error rate was below 0.06 in condition 800/20, 300/40, and 800/40). In general, when there was NO RS, ERS, and MRS, results of log likelihood difference test could be employed to select correct model, and performance of MNRM according to log likelihood difference test improved as sample size and number of items increased. However, in condition ARS, it was not recommended to select correct model by results of log likelihood difference test due to low accuracy and more likely to incorrectly select NO RS as true model.

Another set of model-fit indices included AIC, BIC, and adj BIC. Model with the smallest value of these indices were considered as the best-fitting model. Therefore, in each condition, the number of replications that one of the four competing models identified as the best-fitting model indicated by smallest AIC, BIC, and adj BIC were reported. In NO RS condition, BIC (correctly selected true model NO RS in 97 to 98 out of 100 replications) performed better than AIC (correctly selected true model in 82 to 90 replications) and adj BIC (correctly selected true model in 84 to 96 replications) in selecting the correct model. In ARS condition, all the three model-fit indices were not sensitive to pick out correct model (correctly selected true model ARS in less than 50 replications out of 100 in condition 300/20, 800/20, and 300/20). Only in condition 800/40, correct model can be picked out over half replications. Results of model-fit indices suggested NO RS was more likely to be incorrectly picked out as correct model according to AIC, BIC, and adj BIC when true model was ARS. In ERS condition, AIC (correctly selected true model ERS in 64 to 100 out of 100 replications) and adj BIC (correctly selected true model

ERS in 60 to 100 out of 100 replications) performed better than BIC (correctly selected true model ERS in 31 to 99 out of 100 replications). Only in condition 300/20 by BIC, correct model was selected in less than half replications. In MRS condition, results were similar to those in ERS condition. Only in condition 300/20, correct more was selected in less than half replications (34) based on BIC. AIC and adj BIC in general performed better than BIC in selecting correct model. To summarize, model-fit indices AIC, BIC, and adj BIC were more reliable to identify NO RS, ERS and MRS relative to ARS in MNRM approach. In NO RS condition, BIC performed better than AIC and adj BIC, whereas in ERS and MRS, AIC and adj BIC performed better than BIC. In addition, performance of these model-fit indices for NO RS, ARS, ERS, and MRS improved as sample size and number of items increased.

Parameter recovery was another evaluation criteria for performance of MNRM. Parameter examined in the present study included: intercept parameters, variance of RS factor (true value is 0.0625), and correlations between two content factors (true value is 0.5). First, mean of squared difference between estimated and true intercepts were calculated to examine MNRM. For all the four conditions, NO RS, ARS, ERS, and MRS, results showed that mean of squared difference in correct model was smaller (about 10 times) than wrong models. In correct models, mean of squared difference got smaller as sample size increased, but not much as number of items increased.

Secondly, recovery of RS variance was also examined to evaluate performance of MNRM. In 3 RS condition, variance of RS was close to true value 0.0625, and it got closer to true value as sample size and number of items increased.

The third parameter examined is the estimated correlation between two content factors (true value is 0.5). In NO RS condition, correlation estimates range from 0.497 to 0.513; in ARS condition, correlation estimates range from 0.498 to 0.500; in ERS condition, correlation estimates range from 0.487 to 0.501; and in MRS condition, correlations range from 0.488 to 0.500. However, these estimates did not differ much in correct models and the other three competing models. In general, correlation estimates between two content factors was closer to true value 0.5 as sample size increased from 300 to 800 and as number of items increased from 20 to 40.

Chapter 6

General Discussion

6.1 Summary of present study

The primary objective of the current study was to explore the performance of RS approaches - Count Procedure, RIRS, RIRSMACS, and MNRM in identifying various RS under different simulation conditions. RS examined in the present study includes: ARS (acquiescent response style), ERS (extreme response style), and MRS (mid-point response style). DARS (disacquiescent response style) and MLRS (mild response styles) were not examined in the present study. However, results and conclusions of different approaches on ARS and ERS could be generalized to DARS and MLRS because ARS is opposite to DARS, and ERS is opposite to MLRS. Summary table (Table 40) was presented as below to provide general information about RS approaches examined in the present study:

Table 40. Summary of approaches identifying RS in the present study

RS approaches	Sample size / test length	NO RS (type I error)	ARS (power)	ERS (power)	MRS (power)
Count	300/20	fair	fair	bad	fair
Procedure	800/20	fair	fair	bad	fair
(descriptive)	300/40	fair	fair	bad	fair
	800/40	fair	fair	bad	fair
RIRS	300/20	(0.02, 0.07)	0.11	0.13	0.46
	800/20	(0.06, 0.07)	0.06	0.32	0.79

	300/40	(0.04, 0.06)	0.05	0.29	0.60
	800/40	(0.03, 0.08)	0.05	0.56	0.98
RIRSMACS	300/20	(0.06, 0.13)	0.12	0.19	0.47
	800/20	(0.05, 0.06)	0.06	0.34	0.71
	300/40	(0.05, 0.22)	0.08	0.34	0.98
	800/40	(0.04, 0.09)	0.04	0.59	0.91
MNRM	300/20	(0.02, 0.04)	0.20	0.46	0.59
	800/20	(0.00, 0.03)	0.30	0.87	0.94
	300/40	(0.00, 0.03)	0.42	0.98	0.98
	800/40	(0.00, 0.03)	0.80	1.00	1.00

Across all the RS approaches examined in the present study, it was found that in general, NO RS and MRS were easier to identify, compared to ERS and ARS. Among 3 types of RS, MRS is the easier to detect, and ARS is the most difficult to detect.

For Count Procedure, 2 main research questions in the current study are: to explore performance of Count Procedure in identifying RS through simulation studies; to examine whether performance of Count Procedure improves by using heterogeneous items. Results suggested that Count Procedure on a sample level, performance is good to tell presence of a certain RS if benchmark values were established. However, for individuals, Count Procedure did not perform well because correlations between true and estimated RS scores were low, so results of Count Procedure cannot be used to tell RS for individuals. Among various RS, Count Procedure performed better for NO RS, ARS, and MRS condition, relative to ERS. For ERS condition, scores of ERS

was only slightly higher than ARS, so it was possible to incorrectly identify presence of ARS in addition to true RS – ERS. In general, Count Procedure could be used to identify RS in a sample, but correlations between estimated and true RS factor scores were low under all the RS conditions.

Previous researchers suggested Count Procedure may perform better when correlations between items are low (e.g., Bolt & Johnson, 2009). However, this assumption was not supported in the present study. Performance of Count Procedure in detection of RS did not differ much between using correlated and uncorrelated items.

To sum up, Count Procedure could be employed to identify RS as a descriptive procedure. However, in real world, it is difficult to tell whether a certain RS is present according to RS scores, because there is no established benchmark values for RS scores by Count Procedure. This disadvantage limited application of Count Procedure in empirical studies to detect RS, because results of Count Procedure could only provide RS scores, without giving benchmark values of RS scores to indicate presence of RS. In the present study, Monte Carlo studies were conducted to generate benchmark values for the three types of RS scores. As shown in the present study, RS scores obtained from uncorrelated items when there was no RS were non-zero: 0.350 for ARS, 0.230 for ERS, and 0.310 for MRS. Therefore, when a certain RS score is bigger than benchmark value of that particular RS, it may indicate presence of RS. Nevertheless, benchmark values in the current study were generated under limited conditions examined, such as 5 response categories, sample size is 300 or 800, etc. Cautions should be taken to use these benchmark values when conditions are different.

RIRS and RIRSMACS are similar in that these two approaches require additional uncorrelated items to calculate RS scores. The difference is that in RIRS, RS scores are observed values, and in RIRSMACS, RS are latent factors. In previous research, there is lack of systematic simulation studies to examine performance of RIRS and RIRSMACS. The present study conducted simulation studies to explore performance of RIRS and RIRSMACS in detection of various RS. Results of RIRS and RIRSMACS are similar in that among model fit indices – AIC, BIC, and adj BIC, BIC works better for condition NO RS; whereas AIC and adj BIC work better for conditions ARS, ERS, and MRS. When using power as an evaluation criteria, it was found that power for MRS is higher than ARS and ERS conditions. Besides, power increased more as number of items increased, but not much as sample size increased. In general, performance of RIRS and RIRSMACS are better in identifying NO RS and MRS, but worse in identifying ARS and ERS.

MNRM was one commonly examined IRT model on RS detection. The present study addressed some limitations of previous research on MNRM. One is that previous research only examined performance of MNRM in detection of ERS, so present study further examined ARS and MRS, in addition to ERS. The second limitation of previous research is that RS follows a normal distribution $N \sim (0, 1)$. In the present study, RS follows $N \sim (1, 0.25)$, so that there is only one RS in each simulated dataset instead of mixing multiple RS.

Results of MNRM suggested MNRM performs better in detection of NO RS, ERS, and MRS, relative to ARS. Similar to RIRS and RIRSMACS, when there is no RS, BIC is more sensitive to select correct model. When there is a certain RS, AIC and adj BIC are more sensitive to identify the correct RS.

When it comes to power, it was found that power is high for ERS and MRS conditions, relative to ARS condition. Power increased more as number of items increased, but not much as sample size increased. Parameter recovery was also examined in MNRM. First parameter examined is intercept parameter. It was found that intercept was recovered better in correct model rather than wrong models. Therefore, as shown in MNRM model (7) and (8), better recovered intercept contributes to more accurately estimating probability of selecting a response category. The second parameter examined is RS variance. Across all the RS conditions, it was found that estimated variance of RS in the correct model is close to true value – 0.0625. In addition, as sample size and number of items increased, variance of RS got closer to true value. The last parameter is correlation between two content factors. Different from expectation, the parameter did not differ much in correct or wrong models. One possible explanation could be that: in the present study, correlations between RS and content factors were fixed to 0, so whether RS can be correctly identified did not affect relationships between content factors.

Furthermore, RS factors are treated as latent variables in MNRM and RIRSMACS. One major difference between the two approaches is that, in RIRSMACS, 15 heterogeneous items are indicators to measure latent RS factors, and content items are indicators to measure content factors; whereas in MNRM, both content factors and RS factors are estimated simultaneously through content items. Results of current study suggested that performance of MNRM in picking out the correct model is better than RIRSMACS, indicating adding heterogeneous items does not help with identification of RS. However, this result should be interpreted with caution because data generation model in

the present study is MNRM. Future research may explore performance of MNRM and RIRSMACS when data generation model is RIRSMACS.

Results and findings of the present study provided implications in real world research. As was discussed earlier, RS may affect accuracy of estimates of true scores in survey research. It is recommended to identify presence of RS, so that cautions can be taken in interpretation of findings in empirical studies. According to findings in the present simulations studies, there are some advice to take: 1. Researchers can follow steps of Count Procedure to get a general idea if there is a certain RS present in the data. As was discussed earlier, there was no established benchmark values for ARS, ERS, and MRS. Ideally, a simulation study is suggested to generate benchmark values for that specific study. However, it is not realistic. Therefore, benchmark values in the present study could be used first. In the present study, benchmark values for ARS is 0.350, for ERS is 0.230, and for MRS is 0.310. RS values bigger than corresponding benchmark values are considered as presence of RS. 2. If results of Count Procedure indicated presence of a certain RS. It is advised to employ MNRM to further examination. Compared to RIRS and RIRSMACS, MNRM does not require adding extra heterogeneous items, and results of MNMR showed higher power and low type I error to pick out correct RS.

6.2 Limitations of present study

In the present study, data was simulated from model based on MNRM, adding extra uncorrelated items, so that simulated data could be analyzed with RIRS and RIRSMACS. In addition, content items in RIRS and RIRSMACS were treated as on a nominal scale to minimize differences between data

generation model and data analysis model. Three RS factors were simulated from distribution $N \sim (1, 0.25)$. This is to guarantee all RS factors scores were positive, so there is one RS examined in each simulated dataset. However, the present study did not explore performance of RS approaches when RS is more extreme or less extreme. For example, if mean of ERS is bigger than 1, it is more likely to have a bigger probability to endorse a particular RS according to MNRM model. As shown in Figure 16, there were three ICC figures according to model (6) with different values of ERS:

$$P(U_j = k | \theta_1, \theta_{RS}) = \frac{\exp(a_{jk1}\theta_1 + a_{jk2}\theta_{RS} + c_{jk})}{\sum_{h=1}^K \exp(a_{jh1}\theta_1 + a_{jh2}\theta_{RS} + c_{jh})} \quad (6)$$

Figure 16 presented three figures with value of $ERS = 1, 2$, and 3 . As in Figure 16, when ERS is 2 or 3, the probability of choosing extreme response categories 1 and 5 is bigger than ERS is 1. As the present study only examined performance of approaches identifying RS when mean of RS is 1, so it remains unknown whether results of present study could be generalized to situations when mean of RS is a different value.

Figure 16. ICC Under 2D-MNRM at Three Levels of Response Style

(a).

(b).

(c).

In addition, one assumption of present study is that content factor follows a normal distribution. Therefore, it remains unknown how these approaches may perform if data was simulated from another distribution other than normal distribution.

Furthermore, simulation model in the current study was MNRM, and analyses models included Count Procedure, RIRS, RIRSMACS, and MNRM. As shown in results, MNRM performed better than other approaches in identifying RS. However, results did not exclude the possibility that if data was simulated from other approaches or other models, performance of these four approaches examined in the present study may not be the same. This limitation could be addressed in future studies to simulate data from different models.

6.3 Future directions

As found in the current study, different approaches may perform differently in identifying RS. For example, Count Procedure worked better for

NO RS, ERS and MRS; RIRS and RIRSMACS worked better for NO RS and MRS; MNRM worked better for NO RS, ERS, and MRS. However, it remains a question why all the approaches work better for NO RS and MRS, compared to ARS. Therefore, one direction of future research is to explore this research question.

Previous research found that it is possible to have both ARS and ERS in one dataset (e.g., Marín et al., 1992). However, previous research and current study have not explored performance of RS approaches when there are two RS. Therefore, one future direction could be to further explore performance of RS approaches when there are ARS and ERS.

As mentioned in limitations of the current study, RS follows distribution $N \sim (1, 0.25)$, and content factors follow a normal distribution. However, it was not examined whether results would be different if RS or content factors follow other distributions. For example, when there was more extreme responses and RS follows $N \sim (2, 0.25)$ or if RS does not follow a normal distribution.

In addition, future research could further examine performance of approaches in the present study when true model is not MNRM. For example, data could be simulated from RIRS, RIRSMACS, or other IRT model, to explore if performance of MNRM would be better than other RS approaches, or if data generating model would be outperform other models.

REFERENCES

- Andrich, D. A. (1996). Measurement criteria for choosing among models for graded responses. In A. von Eye and C. C. Clogg (Eds.) *Analysis of categorical variables in developmental research*. Orlando FL: Academic Press. Chapter 1, 3-35.
- Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences, 40*, 1235-1245.
- Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly, 48*, 491-509.
- Baumgartner, H., & Steenkamp, J. B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*, 143-156.
- Bechger, T. M., Maris, G., Verstralen, H. H. F. M., & Béguin, A. A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement, 27*(5), 319-334.
- Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling, 7*, 608-628.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.
- Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*(3), 261-280.

- Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response styles. *Applied Psychological Measurement, 33*(5), 335-352.
- Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement, 71*, 814-833.
- Chen, C., Lee, S. Y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science, 6*, 170-175.
- Cho, Y. (2013). The mixture distribution polytomous rasch model used to account for response styles on rating scales: A simulation study of parameter recovery and classification accuracy. Doctoral dissertation, University of Maryland, Ann Arbor, USA.
- Clark, I. III (2001). Extreme response style in cross-cultural research. *International Marketing Review, 18*, 301-324.
- Couch, A., & Keniston, K. (1960). Yeasayers and Naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology, 60*(2), 151-172.
- de Beuckelaer, A., Weijters, B., & Rutten, A. (2010). Using ad hoc measures for response styles: A cautionary note. *Quality and Quantity, 44*, 761-775.
- de Jong, M. G., Steenkamp, J. B., Fox, J. P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research, 45*, 104-115.

- Greenleaf, E. A. (1992). Measuring extreme response style. *Public Opinion Quarterly*, 56, 328-350.
- Harzing, A. W. (2006). Response styles in cross-national survey research: A 26-country study. *International Journal of Cross-Cultural Management*, 6, 243-266.
- Huggins-Manley, A. C., & Algina, J. (2015). The partial credit model and generalized partial credit model as constrained nominal response models, with applications in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-11.
- Hui, C. H., & Triandis, H. C. (1985). The instability of response sets. *Public Opinion Quarterly*, 49, 253-260.
- Hurley, J. R. (1998). Timidity as a response style to psychological questionnaires. *Journal of Psychology*, 132, 201-210.
- Jin, K., & Wang, W. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement*, 74(1), 116-138.
- Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika*, 68(4), 563-583.
- Johnson, T. R., & Bolt, D. M. (2010). On the use of factor-analytic multinomial logit item response models to account for individual differences in response style. *Journal of Educational and Behavioral Statistics*, 35(1), 92-114.
- Kieruj, N. D., & Moors, G. (2010). Variations in response style behavior by response format in attitude research. *International Journal of Public Opinion Research*, 22, 320-342.

- Knowles, E. S., & Condon, C. A. (1999). Why people say “yes”: A dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, 77, 379-386.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Marín, G., Gamba, R. J., & Marín, B. V. (1992). Extreme response style and acquiescence among Hispanics – the role of acculturation and education. *Journal of Cross-Cultural Psychology*, 23(4), 498-509.
- Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). *Sample size in psychological research over the past 30 years*, 112(2), 331-348.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- Matarazzo, J. D., & Herman, D. O. (1984). Relationship of education and IQ in the WAIS-R standardization sample. *Journal of Consulting and Clinical Psychology*, 52, 631-634.
- MATLAB. (2013). *Version 8.1.0.604 (r2013a)*. Natick, Massachusetts: The MathWorks Inc.
- Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences*, 44, 1539-1550.
- Mellenbergh, G. J. (1996). Measurement precision in test scores and item response models. *Psychological Methods*, 1(3), 293-299.
- Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality & Quantity*, 42, 779-794.

- Moors, G. (2012). The effect of response style bias on the measurement of transformational, transaction, and laissez-faire leadership. *European Journal of Work and Organizational Psychology*, 77, 261-282.
- Muthén, L. K. & Muthén, B. O. (2013). *Mplus user's guide* (7th). Log Angeles, CA: Muthén & Muthén.
- Naemi, B. D., Beal, D., & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of Personality*, 77, 261-282.
- Olson, K., & Bilgen, I. (2011). The role of interviewer experience on acquiescence. *Public Opinion Quarterly*, 75, 99-114.
- Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage.
- Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (Vol. 1, pp. 17-59). San Diego, CA: Academic Press.
- Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49-69). Mahwah, NJ: Lawrence Erlbaum.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J-K., Podsakoff, N. P. (2003). Common method biases in behavioural research: A critical review of the literature and recommended remedies, *Journal of Applied Psychology*, 88, 879-903.
- Ray, J. J. (1983). Reviving the problem of acquiescent response bias, *Journal of Social Psychology*, 121, 81-96.

- Reynolds, N., & Smith, A. (2010). Assessing the impact of response styles on cross-cultural service quality evaluation: A simplified approach to eliminating the problem. *Journal of Service Research*, 13(2), 230-243.
- Roberts, J. (1994). Rating scale functioning. *Rasch Measurement Transactions*, 8, 3, 386.
- Rorer, L. G. (1965). The great response-style myth. *Psychological Bulletin*, 63, 129-156.
- Saal, F. E., Downey, R. G., & Lahcy, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88, 413-428.
- Smith, T. W. (2011). Refining the total survey error perspective. *International Journal of Public Opinion Research*, 23, 464-484.
- Thomas, T. D., Abts, K., Stroeken, K., & Weyden, P. V. (2015). Measuring institutional trust: Evidence from Guyana. *Journal of Politics in Latin America*, 7(3), 85-115.
- van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer.
- van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology*, 35, 346-360.
- van Vaerenbergh, Y. & Thomas, T. D. (2012). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, 25(2), 195-217.

- Watson, D. (1992). Correcting for acquiescent response bias in the absence of a balanced scale – an application to class consciousness. *Sociological Methods and Research*, 21(1), 52-88.
- Weijters, B. (2006). *Response styles in consumer research*. Doctoral dissertation, Ghent University, Ghent, Belgium.
- Weijters, B., Geuens, M., & Schillewaert, N. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science*, 36(3), 409-422.
- Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science*, 36, 409-422.
- Wyer, R. S. (1969). The effects of general response styles on measurement of own attitude and the interpretation of attitude-relevant messages. *British Journal of Social and Clinical Psychology*, 8, 104-115.