ABSTRACT

Title of Document: THE MIXTURE DISTRIBUTION POLYTOMOUS RASCH MODEL USED TO ACCOUNT FOR RESPONSE STYLES ON RATING SCALES: A SIMULATION STUDY OF PARAMETER RECOVERY AND CLASSIFICATION ACCURACY

Youngmi Cho, Doctor of Philosophy, 2013

Directed By: Professor Jeffrey R. Harring
Professor George B. Macready
Department of Human Development and Quantitative Methodology

Response styles presented in rating scale use have been recognized as an important source of systematic measurement bias in self-report assessment. People with the same amount of a latent trait may in some cases be victims of biased test scores due to the construct's irrelevant effect of response styles. The mixture polytomous Rasch model has been proposed as a tool to deal with the response style problems. This model can be used to classify respondents with different response styles into different latent classes and provides person trait estimates that have been corrected for the effect of a response style.

This study investigated how well the mixture partial credit model (MPCM) recovered model parameters under various testing conditions. Item responses that characterized extreme response style (ERS), middle-category response style (MRS),

and acquiescent response style (ARS) on a 5-category Likert scale as well as ordinary response style (ORS), which does not involve distorted rating scale use, were generated.

The study results suggested that ARS respondents could be almost perfectly differentiated from other response-style respondents while the correct differentiation between MRS and ORS respondents was most difficult to attain followed by the differentiation between ERS and ORS respondents. The classifications were more difficult when the distorted response styles were presented in small proportions within the sample. Under the simulated conditions where ten-items and a sample size of 3000 were used there were reasonable item thresholds and person parameter estimates that were obtained. As the structure of mixture of response styles became more complex, increased sample size, test length, and balanced mixing proportion were needed in order to achieve the same level of recovery accuracy. Misclassification impacted the overall accuracy of person trait estimation. BIC was found to be the most effective data-model fit statistic in identifying the correct number of latent classes under this modeling approach.

The model-based correction of score bias was explored with up to four different response-style latent classes. Problems with the estimation of the model including non-convergence, boundary threshold estimates, and label switching were discussed.

THE MIXTURE DISTRIBUTION POLYTOMOUS RASCH MODEL USED TO ACCOUNT FOR RESPONSE STYLES ON RATING SCALES: A SIMULATION STUDY OF PARAMETER RECOVERY AND CLASSIFICATION ACCURACY

By

Youngmi Cho

Doctoral Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee:
Professor Jeffrey R. Harring, Chair
Professor George B. Macready, Co-Chair
Professor Robert G. Croninger
Professor Robert W. Lissitz
Professor Matthias von Davier

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

To provide the background of the problems dealt with in the current study, Chapter 1 reviews the individual differences in rating scale use. Psychological aspects of those individual differences, methodologies to address the related psychometric issues, and the findings in previous empirical studies are detailed. The chapter continues to discuss the purpose and significance of the current study.

1.1 Background of the Problem

1.1.1 Response styles.

While dichotomously-scored item format is more prevalent in cognitive assessment, items with ordered polytomous response categories have been routinely used in self-report, non-cognitive assessment including various psychological tests and attitudinal survey questionnaires. Prototypical examples of ordered polytomous item format are Likert-type rating scales (Likert,1932), of which an illustration is presented in Figure1.[1]

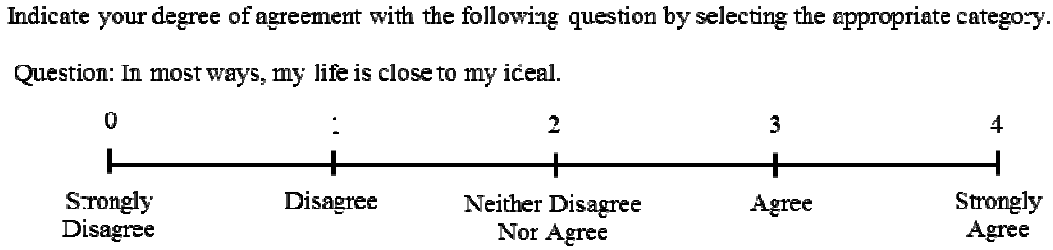

Figure 1. A Likert scale with five ordered response categories

[1] The question 'In most ways, my life is close to my ideal' is one of the five items of Satisfaction With Life Scale (SWLS) by Diener, Emmons, Larse, & Griffin (1985). SWLS intends to measure global cognitive judgments of satisfaction with one's life. The original form of SWLS uses seven response categories and does not use the graphical representation of the continuum as presented in Figure 1.

As seen in the illustrative item in Figure 1, a Likert-scale item attempts to quantify the individual differences in a continuous trait variable based on a certain number of response categories that are often associated with integer scores. It is generally assumed that if a respondent chooses a higher response category, he or she has more of the latent trait being measured by the item than a person who selects a lower response category. The formal aspects of the rating scale such as the number of response categories, category-wording and item-wording can differ in various ways.

In order to utilize the Likert-scale measures as valid indicators of a latent trait of interest and to further compare the trait level among (groups of) respondents, certain necessary conditions must first be satisfied. For example, it must be assumed that respondents' choice of a response category is solely based on the substantive meaning of the item. In other words, any content-irrelevant factor should not systematically influence the respondent's choice of response categories. Additionally, all respondents in a sample interpret the meaning of the provided response categories and use them in the same manner when they answer each item.

These assumptions, however, do not hold if respondents present different *response styles* in responding to a rating scale. A response style (also referred to as a response set or response bias) can be defined as an individual's tendency that causes a person to consistently respond to test items based on some formal aspects of the item or item connotation rather than the underlying construct the item intends to measure (Cronbach, 1946; Messick, 1991; Nunally, 1978; Paulhus, 1991). The prototypical

manifestations of the response styles in ordered polytomous response items are respondents' differential uses of response categories.

Among many others (see e.g., Baumgartner & Steenkamp, 2001 and Paulhus, 1991 for a review of various response styles), three particular patterns of response category use that are well-documented in psychometric literature (e.g., Nunally, 1978: Paulhus, 1991) are the primary focus in the current study. These are extreme response style (ERS), middle-category response style (MRS), and acquiescent response style (ARS). ERS is an individual tendency that leads a person to predominantly use extreme response categories (e.g., categories 0 and 4 in Figure 1) and avoid less extreme choices (response categories in the middle of the scale). Conversely, MRS is a tendency to select the middle category (e.g., category 2 in Figure 1) predominantly while avoiding extreme responses. ARS is a tendency to use only one side of the response scale, i.e., agreement ('*yea-saying*', e.g., categories 3 or 4 in Figure 1) or disagreement ('*nay-saying*', e.g., categories 0 and 1 in Figure 1).

1.1.2 Why response styles matter?

The presence of response styles in a data set can cause various psychometric problems. These adverse effects may invalidate test score differences, obscure true relations among traits of interest, impact test reliability, and confound the results of comparative studies at the group-level.

Response styles can invalidate the assessment of true scores by inflating or deflating observed item scores. Cronbach (1946) pointed out that response styles always reduce logical validity of a test because they permit people with equal

knowledge, identical attitude, or equal amounts of a personality trait to have different test scores. Suppose that there are two people whose true levels of 'satisfaction with life' are located around category 3 on the latent trait continuum in Figure 1. However, they are different in terms of their response styles, i.e., one is an ERS respondent and the other is a MRS respondent. If their different response styles are operating during the item response process, it is highly likely that the two people's choice of response category will not end up with the same. Instead, due to the confounding effects of their different response styles, the ERS respondent might select category 4, for example, while the MRS respondent might select category 2. Consequently, the ERS respondent would be regarded as being more satisfied with his life than the MRS respondent.

Using the observed test scores contaminated by response styles can also cause serious problems in clinical diagnostic settings (see, e.g., Gollwitzer, Eid, & Jürgensen, 2005). In clinical symptom assessments, it is common practice for the total (sum) scores to be computed by adding up the category response scores and these sum scores are compared to appropriate normative values in order to make diagnostic decisions. Without considering individual differences in response styles, this approach for assessing clinical symptoms will lead to lower sensitivity as well as lower specificity of the diagnosis.

Response styles may also give rise to spurious associations among trait domains of interest. Austin, Deary, Gibson, McGregor, and Dent (1998) assessed the consistency of response styles over items and over subscales of the NEO-FFI (NEO-Five Factor Inventory: Costa & McCrae, 1992) by using a measure of response spread

on a rating scale. They found non-trivial, highly significant correlations between unrelated, independent items. The observed spurious correlations may be attributed to the effect of response styles operating across the items because it seems unlikely that the items whose contents are not related with each other yielded such high levels of correlation. Austin et al. (1998) also pointed out that such spurious correlations could cause erroneous extraction and interpretation of latent factors in multivariate data analysis that were based on correlation matrices. Similarly, Austin et al. (2006) and Baumgartner and Steenkamp (2001) provided empirical support for the contribution of response styles inflating scale-level correlations.

The impact of response styles on test reliability can be found in a simulation study by Liu, Wu, and Zumbo (2009). They generated outlying data, which represented ERS responses under a mixture modeling framework. Their results of the bias and efficiency of Cronbach's coefficient alpha showed that outliers severely inflated the alpha coefficient as well as the standard error of the estimates of the coefficient.

Another methodological issue is that response styles tend to be manifested differentially across groups. That is, certain response styles tend to be more prevalent in a particular group than in another. This between-group variability in response styles is likely to contribute to the violation of structural invariance and, in turn, any observed group differences may simply reflect measurement artifacts due to the differences in response styles. Regarding the between-group variability, Cheung and Rensvold (2000) used multi-group confirmatory factor analysis and demonstrated that

certain types of measurement non-invariance were attributed to the manifestations of ERS and ARS. Bolt and Johnson (2009) applied a multidimensional item response theory (IRT) model and found that ERS was an underlying source of item differential functioning (DIF).

In various areas of study such as marketing, organizational and industrial psychology, education, and medicine there has been accumulating empirical evidence of between-group variability across nations, ethnic groups, and cultural regions (e.g., Baumgartner & Steenkamp, 2001; Buckley, 2009; Cheung & Rensvold, 2000; Harzing, 2006; Yang, Harkness, Chin, & Villar, 2010). For example it has been shown that ERS and ARS are more prevalent in among Hispanics/Latinos and African-Americans than among Caucasians in the U.S. (Bachman & O'Malley, 1984; Clarke III, 2000; Hui & Triandis. 1989; Marin, Gamba & Marin, 1992; Ross & Mirowsky, 1984). Japanese and Chinese respondents in the U.S. tended to use extreme responses less often than Americans in responding to positive feeling (Lee, Jones, Mineyama, & Shang, 2002). Japanese and Korean students tended to use middle categories more often than their American counterparts (Chen, Lee, & Stevenson, 1995; Lee & Green, 1991). In Europe, ERS has been shown to be more prevalent in Mediterranean countries (Italy, Spain, and Greece) than in the United Kingdom, Germany, and France (Van Herk, Poortinga, & Verhallen, 2004).

1.1.3 Response styles as meaningful constructs

Rather than perceiving response styles as a source of systematic measurement bias, one strand of research in psychology views response styles as meaningful

reflectors of psychological constructs such as personality traits and cognitive processes, or some cultural values. In those research studies, the relation between some criteria variables and specific response style were investigated. For examples, ERS appeared to be positively related to trait anxiety (Berg & Collier, 1953; Lewis & Taylor, 1955; Norman 1969), extraversion (Austin, Deary, & Egan, 2006), and conscientiousness (Austin et al., 2006; Harzing, 2006).

In cognitive process research area, Temple and Geisinger (1990) and Kulas and Stachowski (2008) found that middle category endorsements (e.g., 'neither disagree nor agree', 'no answer', or '?') exhibited longer response latencies than other category endorsements and were more frequently elicited when the given items were unclear, personally intrusive, or asked introspective questions. The results of these experimental studies have shown the evidence of increased cognitive load in processing information contained in the middle category. The implication is that response styles, in some cases, could be associated with the respondent's attempts to reduce the cognitive demand required to process the meaning of the item content and the labels of the response categories.

In cross-cultural comparative studies, the types of response style and cultural values are associated. For example, using the measures of Hofstede's cultural dimensions,[2] several studies argued that ARS seemed to be positively correlated with collectivism and femininity but negatively related to power distance and uncertainty

[2] The Hofstede's cultural dimensions theory (Hofstede, 1980) postulates four dimensions along which cultural values can be analyzed. The four dimensions are individualism-collectivism; uncertainty avoidance; power distance (strength of social hierarchy) and masculinity-femininity (task orientation versus person-orientation).

avoidance. ERS appeared to be positively correlated with individualism, power distance, uncertainty avoidance, and masculinity (see e.g., Chen, Lee, & Stevenson, 1995; de Jong, Steenkamp, Fox, & Baumgartner, 2008; Harzing, 2006; Johnson, Kulesa, Cho, & Shavitt, 2005).

1.1.4 Methodology to deal with response styles

No matter how response styles are considered, i.e., treated as a statistical nuisance that needs to be controlled for or as a meaningful construct of interest, the initial treatment of the data analysis should be the distinction of the cases that are influenced by certain response styles. Following the distinction, the identified cases can be either controlled for (by eliminating the cases from the data or applying a correction method) or related with other variables to reveal the nature of the response styles and investigate their structural relations among latent variables.

Traditional strategies dealing with response styles use simple descriptive statistics calculated for heterogeneous items and balanced scales, which are designed as "built-in control" in an instrument. Relatively recently, different latent variable models have been proposed to aid in solving this response style problems.

Heterogeneous items. Heterogeneous items refer to the items whose contents are psychologically diffused and theoretically independent of each other. In practice, a number of items that do not refer to substantively meaningful psychological construct can be used as heterogeneous items in an assessment. Alternatively, items varying widely in content can be selected from diverse set of scales that have little in common (see, e.g., Couch & Keniston, 1960). If a respondent consistently favors particular

response categories (e.g., extreme categories) across such heterogeneous items, this behavior can be taken as evidence of a response style (e.g., ERS). Response style measures for ERS, MRS, or ARS can then be derived by calculating the number or the proportion of the heterogeneous items on which a respondent selects the most extreme categories, middle category, or categories in just the upper or the lower extreme, respectively. Instead of frequency or proportion, response range as measured by the standard deviation of item scores within individuals has also been used (Austin et al., 1997; Greenleaf, 1992; Hui & Triandis, 1985).

The major weakness of using heterogeneous items is that if the substantive independence among heterogeneous items is not warranted for a given sample of respondents, which is not unusual in practice, the resulting response style measures are confounded with the respondent's trait level. In such cases, clustering respondents into different response-style groups may not be valid and inferences based on these clusters can hardly be justified. There is also a practical limitation. In the literature, it has been pointed out that the number of heterogeneous items should be large in order for a response style to have sufficient opportunity to manifest itself by permeating the responding pattern in a consistent way (Couch & Keniston, 1960; Greenleaf, 1992). If a test is lengthened due to the inclusion of heterogeneous items, it may raise some psychometric problems of a test (e.g., an increase in measurement error due to the respondent's fatigue and lowered face validity of the test) as well as the issues of time and cost needed for the administration of the inventory.

Balanced scales. A balanced scale consists of pairs of logically reversed items, i.e., one item of the pair states a construct positively while the other of the pair states the equivalent construct negatively (Couch & Keniston, 1960; Paulhus, 1991). In such a way, the scale becomes semantically balanced. If a respondent has a tendency to acquiesce and respond to a pair of such logically reversed item by *'yea-saying'* or *'nay-saying'* to both, his or her responses are conceptually conflicting. If this conflicting endorsement is repeated, it can provide strong evidence for ARS. Using a balanced scale in an assessment per se does not preclude the occurrence of ARS. A well-constructed balanced scale, however, can alleviate score distortion to some degree. By "reverse coding" item responses (mostly, responses to negatively worded items) before summing up all item scores, high or low item scores obtained by simply *'yea-saying'* or *'nay-saying'* will cancel each other out and ARS respondents will receive a moderate test score.

Mirowsky and Ross (1991) showed that the ARS inflated the variance and reliability of the trait estimates when unbalanced scales were used, leading to either an overestimation or an underestimation of the relation between the construct measured by the unbalanced scale and other constructs. Watson (1992) showed that the covariance due to ARS is extracted using structural equation modeling when an unbalanced scale is used.

Model-based approaches. Besides utilizing heterogeneous items and balanced scales in the test development stage, an increasing number of studies have attempted a

more rigorous solution to this problem by applying latent variable models into which response style effects are directly incorporated.

Within the structural equation modeling (SEM) framework, response styles are examined as group characteristics and group differences in the manifestation of response styles are statistically tested. Cheung and Rensvold (2000) applied multiple-group confirmatory factor analysis to test for the presence of ERS and ARS and determine whether cultural groups can be meaningfully compared on the basis of factor means. Group differences in ERS and ARS are operationalized as non-invariance in the factor loadings and intercepts. This study showed the utility of using the SEM approach in this matter, but also highlighted its limitations. The SEM approach was not appropriate to use when no items in the scale were invariant across groups with respect to the effects of response styles. Also, the SEM approach does not provide individual level information.

Billiet and McClendon (2000) estimated a confirmatory three-factor model that included ARS as a common "style" factor (i.e., method factor) in addition to two "content" factors. By using two sets of balanced scales measuring two independent constructs, they demonstrated that the effects of style factor can be separated from the content factors. Moors (2003, 2004, 2008) adapted the same rationale as Billiet and McClendon (2000) but within latent class factor analysis (LCFA). Moors emphasized the flexibility of this approach over multi-group CFA in that LCFA allowed response styles to be manifested within an exploratory setting in which no response style was hypothesized in a given data set. In Moors empirical studies, an ERS factor was

identified. Billiet and McClendon and Moors's approach commonly impose a restriction that the factor loadings are equal for all items. However, if the items are actually influenced differentially by the response style, assuming a constant factor loading on the style factor would lead to a model misspecification.

Within an item response theory (IRT) framework, Bolt and Johnson (2009) developed a multidimensional model that extends Bock's nominal response model (Bock, 1972) to investigate ERS. In this model, response styles were characterized as continuous trait dimensions that influenced the attractiveness of particular score categories. The item response probabilities were defined as a function of two trait dimensions, i.e., an intended substantive trait and ERS tendency. Based on the estimates for these two dimensions, observed test scores were rectified for the impact of ERS. Although this approach has been shown to be useful to help understand how both substantive and ERS traits are combined to affect item response behaviors, whether it can be successfully applied for other types of response styles (e.g., MRS and ARS) and whether the condition in which more than two response styles are presented in a sample can be handled have not yet been explored.

De Jong, Steenkamp, Fox, and Baumgartner (2008) proposed a model that extended a standard IRT model by integrating testlet models (e.g., Bradlow, Wainer, & Wang, 1999) and a structural multilevel model. The inclusion of the testlet component in the model permits a control for substantive correlations that may exist among heterogeneous items. This model allows the response styles to have differential impact across items. In addition, measurement invariant anchor items are not required

for group comparisons. This approach successfully identifies ERS, but is arguably less useful for correcting the effects of ERS on substantive trait estimates (Bolt & Newton, 2010).

Lastly, mixture polytomous IRT models, which generalize the standard polytomous IRT models to mixture distribution models, have been used by an increasing number of researchers in various disciplines compared to the other model-based approaches previously introduced. Similar to LCFA, mixture polytomous IRT models are useful for the study of response styles in an exploratory manner, which is not benefited from the SEM approach as well as the extended IRT models by Bolt and Johnson (2009) and by De Jong et al. (2008). Unlike the Bolt and Jonhson (2009) approach where response styles were treated as continuous variables (quantitative differences), mixture polytomous IRT models treat response styles as discrete variables (qualitative differences) and assign each respondent to a latent class membership that represent his or her response style. This would allow for a more flexible and effective modeling technique that can be applicable when multiple response styles are present within a sample of respondents. Not only the classification of respondents but also the individual-level estimate of latent trait is obtained with mixture polytomous IRT models, which is not available information in the studies in the SEM framework. More details of the mixture polytomous IRT models are followed in the subsequent section as well as in Chapter 2.

1.2 Mixture IRT Models in Empirical Studies

As mentioned earlier, the common manifestations of response styles, regardless of the cause of the emergence of response styles, is respondents' disproportionate usages of response categories. Different types of response styles can be characterized by different category response probabilities. For example, a sample of ERS respondents shows a high probability of endorsing the end-categories. Based on the analysis of the unique patterns of category response probabilities, mixture polytomous IRT models provide the way that can distinguish latent groupings of respondents with different response styles.

In general, mixture IRT models assume that the respondent population can be heterogeneous not only quantitatively but also qualitatively. If respondents are different with respect to how they use the response categories, this heterogeneity can possibly be captured using mixture IRT models and respondents with different response styles are classified into different latent classes. A latent trait estimate is assigned to each respondent within the identified classes and, hence, the response style effects can be controlled when latent trait levels are compared.

Mixture polytomous Rasch models are special cases of mixture IRT models where the category response probabilities are predicted by one of the logistic functions of the polytomous Rasch family such as the partial credit model (Masters, 1984), rating scale model (Andrich, 1978), mixed dispersion model (Andrich, 1982), and successive interval model (Rost, 1988).

The mixture partial credit model (MCPM) was proposed by Rost as an extension of latent class analysis that takes account of the different usage of rating scales within latent classes (Rost, 1991). When he proposed the MPCM, he suggested this model as a method for classifying people according to their item response profile, independent of the location of the profile on latent continuum. Because the MPCM is the Rasch model in which no restriction on the item parameters is imposed, it is often called the mixture (or mixed) polytomous Rasch model (Rost, 1991; von Davier & Rost, 1995). In this dissertation, the mixture partial credit model (MPCM) and the mixture polytomous Rasch model are used interchangeably.

Mixture polytomous IRT models, especially the MPCM, have been increasingly used in applied studies in personality, organizational, and clinical psychology for the analysis of Likert-scale self-report data. (e.g., Austin, Deary, & Egan, 2006; Egberink, Meijer, & Veldkamp, 2010; Eid & Rauber, 2000; Gollwitzer et al., 2005; Maij-de Meij, Kelderman, & van der Flier, 2005, 2008; Meiser & Machunski, 2008; Rost, 1991; Rost, Carstensen, & von Davier, 1997; Smith, Ying, & Brown, 2012; Wu & Huang, 2010; Zickar, Gibby, & Robie, 2004). All these referred studies used the MPCM except Maij-de Meij et al. (2005, 2008), which used the mixture nominal response model, Egberink et al. (2010), which used the mixture graded response model, and Meiser and Machunski (2008), which used the mixture rating scale model.

1.3 The Current Study

As reviewed in this chapter, the MPCM has great potential to provide solutions to the long-standing psychometric problems caused by response styles. Despite the growing interest and need in practical settings, little evidence has been provided about the accuracy of parameter estimation of the MPCM. When Rost (1991) proposed the MPCM, a one-replication simulation study was conducted in which the quality of MLE was evidenced. However, the simulation conditions were very limited, which made the results difficult to be generalized. In the MPCM, the accuracy of parameter estimates can vary depending on several factors such as the estimation algorithm, the number of items, the number of respondents, and the number and size of latent classes.

The current study, therefore, proposes to conduct a larger-scale simulation in which the quality of MPCM parameter estimation is evaluated especially under the population where different response styles coexist. Specifically, the recovery of latent class membership, item thresholds, and person trait levels will be examined. The effects of the type of mixture, mixing proportions, sample size, and test length on the parameter recovery are assessed. In addition to the parameter recovery study, the simulation study will also examine how the MPCM makes an adjustment of the latent trait estimates to compensate for the effects of different response styles on test score. The effectiveness of information criteria for the MPCM model selection is also assessed.

Given that there has been thus far no systematic simulation study that investigates the parameter recovery of the MPCM, the current study is expected to

provide some evidence regarding the soundness of the application of this model in empirical data analysis. Especially, the various mixture conditions of response styles simulated in this study will allow for the evaluation of the utility of the MPCM in dealing with particular response styles problems in real data analytic research.

Chapter 2: Literature Review

Chapter 2 starts with an introduction of the conceptual development of the Rasch model (RM) and partial credit model (PCM) as well as the unique features of the Rasch family models. The chapter continues to introduce the finite mixture distribution before presenting the model formulation of the MPCM. Estimations of the model parameters of the MPCM and the application of the information criteria for model selection are discussed. Finally, the designs and results of related empirical and simulation studies are summarized.

2.1 The Rasch Model for binary item responses

2.1.1 Presentation of the model

Item response theory (IRT) was built around the central idea that the probability of a certain answer when a person is confronted with an item ideally can be described as a simple function of the person's position on the latent trait scale and one or more parameters characterizing the particular item (Molenaar, 1995). The Rasch model for dichotomously scored item responses (RM: Rasch, 1960) is the simplest IRT model in the sense that it only needs the difficulty of an item, which indicates the location of the latent trait scale, in order to characterize an item. This simplicity allows the RM to directly compare item and person parameters to define the item response probability. The following introduces the essential idea of the Rasch measurement model applied to the comparison of the difficulty of an item i (δ_i) and person n's trait level (θ_n) on the same latent continuum.

Suppose that specific δ_i and θ_n are located at positions ξ_D and ξ_T on a latent variable continuum, respectively. In addition, P_T is the probability of observing an event *T* indicating that ξ_T exceeds ξ_D on the continuum. Similarly, P_D is the probability of observing an event *D* indicating that ξ_D exceeds ξ_T. Considering the relative locations of θ_n and δ_i on the latent continuum, P_T would imply the probability of a success on the item whereas P_D would imply the probability of a failure on the item. For dichotomous responses, P_T may be replaced as P_{ni1} representing the probability of person *n* scoring 1 on item *i*. Also, P_D may be replaced as P_{ni0} representing the probability of person *n* scoring 0 on item *i*. The RM then relates the distance between θ_n and δ_i on the continuum to the events *T* and *D* as the natural logarithm of the odds ratio presented in Equation 1.

$$\xi_T - \xi_D = \theta_n - \delta_i = \ln\left(\frac{P_T}{P_D}\right) = \ln\left(\frac{P_{ni1}}{P_{ni0}}\right). \quad (1)$$

As seen in Equation 1, the log odds of observing a success rather than a failure on item *i* is determined based on the distance between θ_n and δ_i. From Equation 1, one can easily verify that when $\theta_n = \delta_i$, $P_{ni1} = P_{ni0} = 0.5$. If $\theta_n > \delta_i$, it implies that the respondent's ability surpasses the difficulty level of the item, indicating a greater chance of success because the odds (P_{ni1} / P_{ni0}) must be greater than 1. Conversely, if $\theta_n < \delta_i$, it implies that the difficulty level of the item surpasses the respondent's ability, indicating a greater chance of failure because the odds (P_{ni1} / P_{ni0}) must be

smaller than 1. Using the inverse logistic, Equation 1 transforms with respect to P_{ni1} as presented in Equation 2.

$$P_{ni1} = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}, \quad (2)$$

where P_{ni1} is the probability that person *n* correctly answers item *i*, or the probability of scoring 1 on item *i*, θ_n is the trait level for person *n*, and δ_i is the difficulty of item *i*. This is the RM equation, which is the basic building block shared by all models within the Rasch family.

2.1.2 Item response function

Equation 2 provides a trace line that indicates the probability of a correct item response at all possible levels of θ for a given difficulty δ_i. This trace line is referred to as an item response function (IRF) or item characteristic curve (ICC). Figure 2 illustrates three ICCs that the RM produces for items with δ_i = -0.5, 0, and 0.5, respectively. As can be seen in the plot, the RM ICCs differ only with respect to the locations on the continuum indicating different levels of item difficulty. The slopes of the ICCs are parallel, which indicates that discriminations of the items are the same for the three items. As mentioned in the previous section, the direction of the response probability changes at the point that corresponds to the probability value of 0.5, which is the point of inflexion of the ICC.

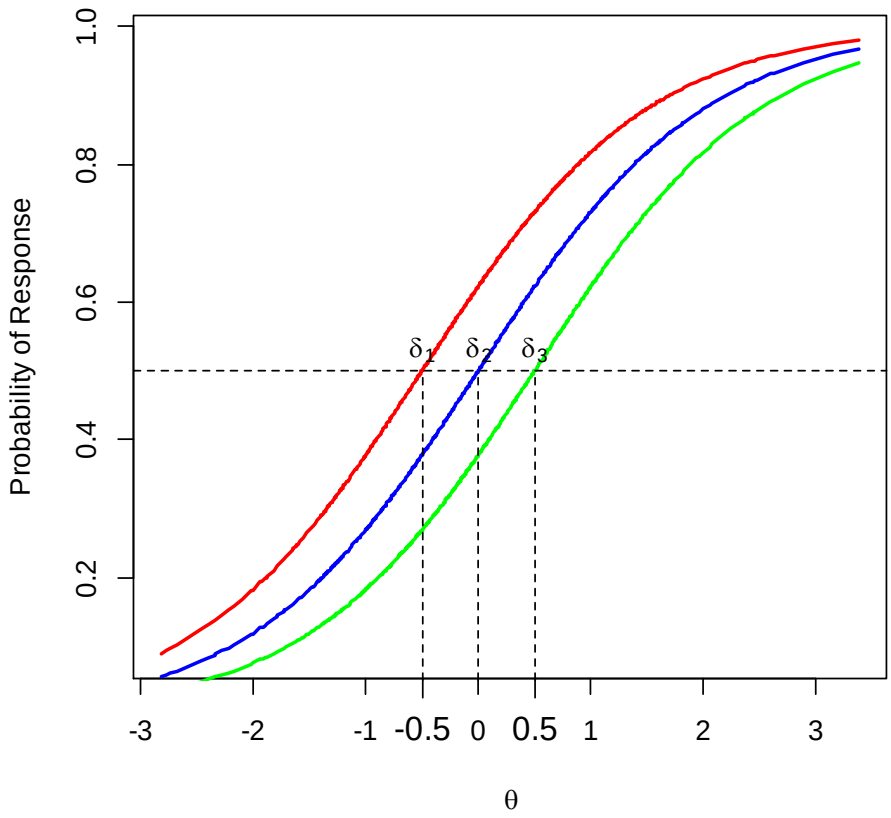

Figure 2. CCs corresponding to Rasch model items with different item difficulty

2.2 Partial Credit Model

2.2.1 Presentation of model.

Masters (1984) proposed the partial credit model (PCM) by extending the RM to polytomously-scored item responses. The fundamental idea of the PCM is that the multiple response categories are a series of pairs of adjacent categories and the RM can be applied for modeling each pair. The PCM is appropriate for the items that are subject to partial credit scoring as well as those that are obtained with a Likert-type scale.

Masters (1984) introduced the concept of *step* as he proposed the PCM. As depicted in Figure 3, a step in an item represents the transition from one category to the next. Thus, there are *k* steps in an item with *k* + 1 response categories.

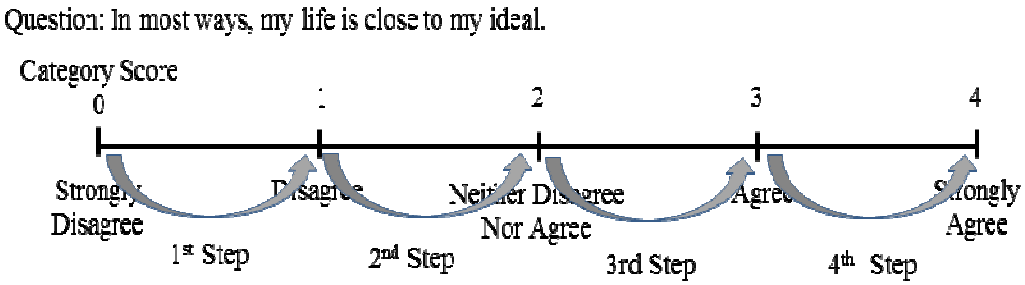

Figure 3. Five response categories and four corresponding steps

On this Likert scale, passing the *k*th step means selecting response category *k* over *k*-1 in response to the item. If a person chose 'Agree' (response category 3), for example, he or she is regarded to have selected 'Disagree' over 'Strongly disagree' (first step passed), 'Neither disagree nor agree' over 'Disagree' (second step passed), and 'Agree' over 'Neither disagree nor agree' (third step passed), but to have failed to make a transition from 'Agree' to 'Strongly Agree'. In this case, the person will earn a partial credit score of 3, i.e., the number of the steps that he or she has passed.

For dichotomously-scored items, there is only one pair of adjacent categories and, hence, only one step needs to be passed to reach the highest score or 1. Let us revisit Equation 2, which now can be considered as a special case of the PCM where

the test items are one-step items. To make this point explicit in the model presentation, Equation 2 may be rewritten using modified notations following Masters (1984):

$$P_{ni1} = \frac{\phi_{ni1}}{\phi_{ni0} + \phi_{ni1}} = \frac{\exp(\theta_n - \delta_{i1})}{1 + \exp(\theta_n - \delta_{i1})},$$

where ($\phi_{ni0} + \phi_{ni1}$) is the probability of person *n* scoring 0 or 1 on item *i* and P_{ni1} is the probability of person *n* passing the first step to score 1 rather than 0 on item *i* conditional on that only the two successive categories are considered. δ_{i1} is the first (and the only in this case) step difficulty. The details of the step difficulty will be shortly introduced in the subsequent section. For the second pair of categories, the RM is again applied:

$$P_{ni2} = \frac{\phi_{ni2}}{\phi_{ni1} + \phi_{ni2}} = \frac{\exp(\theta_n - \delta_{i2})}{1 + \exp(\theta_n - \delta_{i2})},$$

where P_{ni2} is the probability of person *n* passing the second step to score 2 rather than 1 on item *i* conditional on that only the two successive categories are considered. The general form of the step difficulty probability that person *n* passes the *k*th step to score *k* rather than *k*-1 on item *i* is then defined as:

$$P_{nik} = \frac{\phi_{nik}}{\phi_{ni,k-1} + \phi_{nik}} = \frac{\exp(\theta_n - \delta_{ik})}{1 + \exp(\theta_n - \delta_{ik})}, \; k = 1, 2, \ldots, \; h_i . \tag{3}$$

Here, note that h_i is used to indicate potentially varying number of steps in different items. In the PCM, it is assumed that person *n* must select one of the given *k*+1 categories. Therefore, the following restriction needs to be applied:

$$\phi_{ni0} + \phi_{ni1}, \ldots, + \phi_{nik} = 1.$$

Finally, combining Equation 3 and the restriction, the PCM can be written as the unconditional probability that person *n* scores *x* on item *i* over all other possible scores.

$$\phi_{nix} = \frac{\exp\left(\sum_{k=0}^{x}(\theta_n - \delta_{ik})\right)}{\sum_{g=0}^{h_i}\left(\exp\left(\sum_{k=0}^{g}(\theta_n - \delta_{ik})\right)\right)}, \; x = 0, 1, 2, \ldots, \; h_i, \tag{4}$$

where $\sum_{k=0}^{0}(\theta_n - \delta_{ik}) \equiv 0$.

To show how Equation 4 determines a category response probability, an explicit expansion of Equation 4 is demonstrated below. The illustration is to calculate the category response probability for the third category (ϕ_{ni3}) when five response categories are given.

$$\begin{aligned}
\phi_{ni3} &= \frac{\exp[0 + (\theta_n - \delta_{i1}) + (\theta_n - \delta_{i2}) + (\theta_n - \delta_{i3})]}{\exp[0] + \exp[0 + (\theta_n - \delta_{i1})] + ...} \\
&= \frac{}{\exp[0 + (\theta_n - \delta_{i1}) + (\theta_n - \delta_{i2})] + ...} \\
&= \frac{}{\exp[0 + (\theta_n - \delta_{i1}) + (\theta_n - \delta_{i2}) + (\theta_n - \delta_{i3})] + ...} \\
&= \frac{}{\exp[0 + (\theta_n - \delta_{i1}) + (\theta_n - \delta_{i2}) + (\theta_n - \delta_{i3}) + (\theta_n - \delta_{i4})] + ...} \\
&= \frac{}{\exp[0 + (\theta_n - \delta_{i1}) + (\theta_n - \delta_{i2}) + (\theta_n - \delta_{i3}) + (\theta_n - \delta_{i4}) + (\theta_n - \delta_{i5})]}.
\end{aligned} \tag{5}$$

2.2.2 Threshold parameters in the PCM

Masters (1984) used the term '*step difficulty*' to refer to δ_{ik} . The step difficulty is conceptually the same with the item difficulty in the RM. It indicates the location of a particular step on the latent trait continuum and the location of each threshold can be compared to the location of person. The probability of passing a step to select a particular response category is determined based on the relative locations of these two locations (i.e., step and person) on the latent continuum. In the IRT literature, several alternative terms have been used such as *category intersection* (see, e.g., Embretson & Reise, 2000), *category transition location* (de Ayala, 2009), and *threshold* (Rost, 1991; von Davier & Rost, 1995). Hereafter the step difficulty δ_{ik} is referred to as the *threshold*.

The mean of the thresholds within an item is often used to indicate the global/general location of the given item.[3] In the current study, this is referred to as the *item location*. The item location β_i is defined as follows:

$$\beta_i = \sum_{k=1}^{h} \delta_{ik} / h_i , \quad k = 1, 2, \ldots, h_i ,$$

where δ_{ik} is the *k*th threshold for item *i* and h_i is the number of thresholds of item *i*.

[3] The PCM can be reformulated so that the threshold is decomposed into item location (β_i) and the difference between threshold and item location (τ_{ik}). Equation 2 can be rewritten as follow:

$$P_{nik} = \frac{\exp(\theta_n - \beta_i - \tau_{ik})}{1 + \exp(\theta_n - \beta_i - \tau_{ik})} .$$

Among a set of *k* steps within a PCM item, some steps may be easier to pass than others. If a particular step is easier to pass than others, the threshold value associated with that step will be lower than those associated with more difficult steps. One of the important features of the PCM is that the model does not assume that there is an underlying sequential step process to achieve a partial score. Although the response category scores (e.g., 0, 1, 2, 3, and 4) should be ordered to reflect increasing θ level, the estimated thresholds are not restricted to follow a specified order. When the thresholds are disordered, for example, $\delta_1 = 0.45$, $\delta_2 = 0.74$, $\delta_3 = -0.74$, and $\delta_4 = -0.45$, instead of being ordered as in the following example, $\delta_1 = -0.74$, $\delta_2 = -0.45$, $\delta_3 = 0.45$, and $\delta_4 = 0.74$, it is often called a *reversal* of the thresholds.

2.2.3 Category characteristic curves and the presence of response styles

Understanding how the order of thresholds and distances between thresholds are related to the category response probabilities in the PCM is fundamental to simulate item response patterns contaminated by different types of response styles in the current study. Continuing previous sections, this section further explains the relations between thresholds and category response probabilities by introducing graphical representations of the relations.

Similar to the ICCs in Figure 2, the category response probability of a polytomous item (ϕ_{nix} in Equation 4) can be depicted as a trace line called a category

characteristic curve (CCC).[4] A CCC relates the probability of choosing a particular response category given a specific θ value. While only one ICC is needed for a dichotomously-scored item, as many CCCs as the number of response categories are required to present probabilities for each category response for a polytomously-scored item. Note that each category response probability can be calculated by following Equation 5. Figures 4 to 8 present different patterns of CCCs that have hypothetical threshold values estimated for the groups of different response-style respondents. In these CCC plots, four trace lines representing threshold probabilities (P_{nik} in Equation 3) are overlaid. The black lines present the threshold probabilities while the colored lines present CCCs. In the plots, it is commonly seen that the thresholds correspond to the points of inflexion of threshold probabilities and those points are the intersections of two adjacent CCCs. This indicates that when the item difficulty level is at *k*th threshold, the probability of choosing *k* and that of *k*-1 are the same at 0.5. As the item difficulty increases from *k*, the probability of choosing *k* becomes higher while the probability of choosing *k*-1 becomes higher as the difficulty decrease from *k* in this group of respondents.

Ordered thresholds and the implication for response styles. In the following Figure 4, the CCCs and threshold probabilities are dictated by a set of four thresholds δ_{i1} = -1.7, δ_{i2} = -0.6, δ_{i3} = 0.6, and δ_{i4} =1.7. Apparently, the four thresholds are in a strict order from low to high values on the θ continuum and the distances between thresholds are fairly evenly spaced. The latent trait space is divided into five segments

[4] CCC is sometimes called as category response curve, category response function, option response function, or operating characteristic curve.

within each of which one of the five categories has the greatest probability to be selected than the others. For example, respondents with the lowest level of θ would be most likely to choose the response category 0 (see the CCC in orange color) while respondents within the next higher θ range, between δ_1 and δ_2, would choose category 1 with the highest probability than any other categories (see the CCC in brown color).

Figure 4 shows that every category is used properly in accordance with the respondent's θ level. In this group of respondents, no response category is avoided and the item categories seem to function well as they are designed to differentiate individual's trait level. Related to the issue of response styles, this pattern of CCCs and threshold probabilities is likely to be observed in a measurement situation where respondents would not present a particular response style such as ERS, MRS, or ARS but respond to the item solely conditional on their θ level. This "normal" responding pattern is referred to as *ordinary response style* (ORS) in the current study.

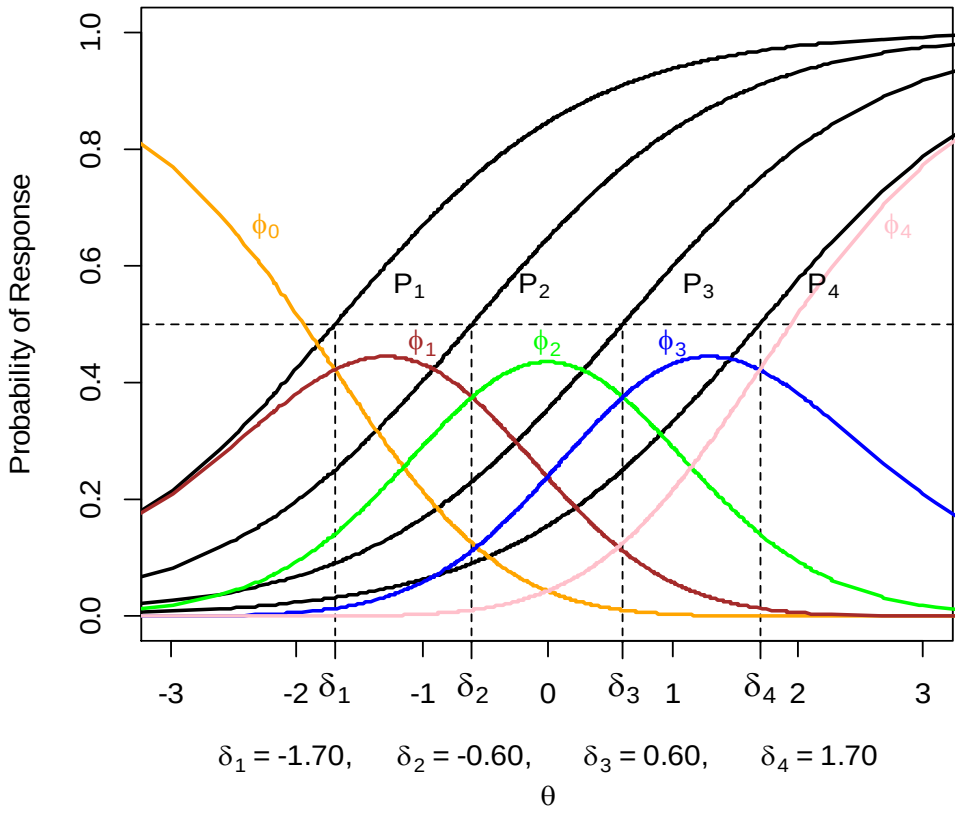

Figure 4. CCC and threshold probabilities for a PCM item with thresholds (-1.7, -0.6, 0.6, and 1.7)

Figure 5 also shows a set of ordered thresholds (δ_1 = -1.85, δ_2 = -1.24, δ_3 = 1.34, and δ_4 =1.95) but compared to Figure 4, the distances between thresholds are uneven. The distance between the second and third threshold is longer than the distances between other thresholds, which links to the relatively high probability for category 2 to be selected within a wide range of values on the θ continuum. In this plot, Category 1 and 3 are still the most favorable category within the ranges from δ_1 to δ_2 and from δ_3 to δ_4, respectively. The pattern of CCCs in Figure 5 may be observed in a sample of MRS respondents.

If the distance between δ_2 and δ_3 becomes longer, in other words, if the number of people in the sample who select the middle category increases, then the CCC for the middle category will peak more distinctively and the order of δ_1 and δ_2 as well as that of δ_3 and δ_4 can be reversed. An illustrative plot is shown in Figure 6 in which a larger proportion of respondents respond to the middle category and accordingly the reversals occur.

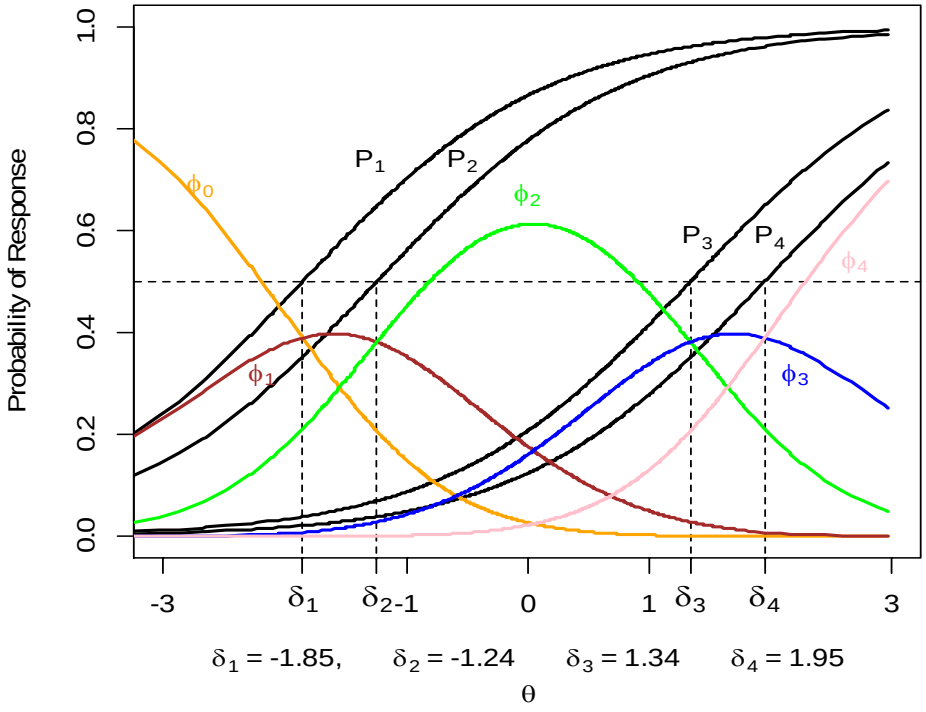

Figure 5. CCCs and threshold probabilities for a PCM item with thresholds (-1.85, -1.24, 1.34, and 1.95)

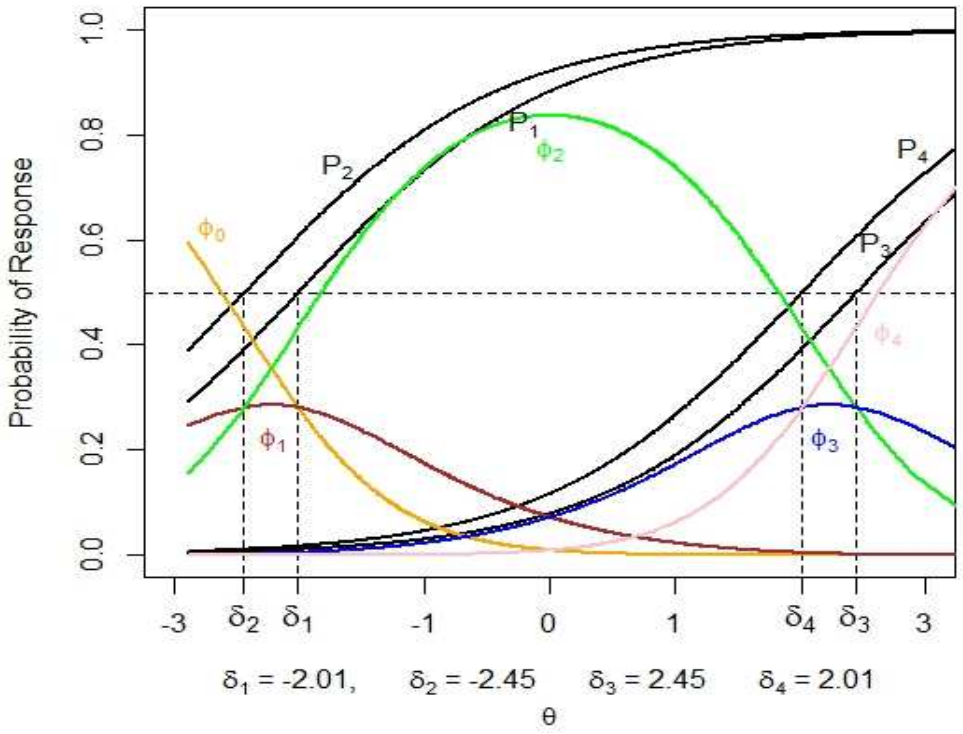

Figure 6. CCCs and threshold probabilities for a PCM item with thresholds (-2.01, -2.45, 2.45, and 2.01)

Reversed thresholds and the implication for response styles. The following Figure 7 shows a dramatically different array of CCCs and threshold probabilities from the previous figures. In this case, a reversal occurs (δ_1 = 0.45, δ_2 = 0.74, δ_3 = -0.74, and δ_4 = -0.45) and the latent trait space is predominantly taken by the first and the last CCCs. Category 1, 2, and 3 are never be the most likely category to be selected at any θ level. If a respondent in this sample has a higher level of θ than zero (i.e., the point where the first and the fifth CRCs intersect), category 4 has the highest probability to be chosen. Conversely, if a respondent has a lower level of θ than zero, category 0 has the highest probability of being selected. Category 1, 2, and

3 will rarely be selected. If estimated CCCs show this pattern of distortion, this may be evidence that item responses from this sample of respondents are contaminated by ERS.

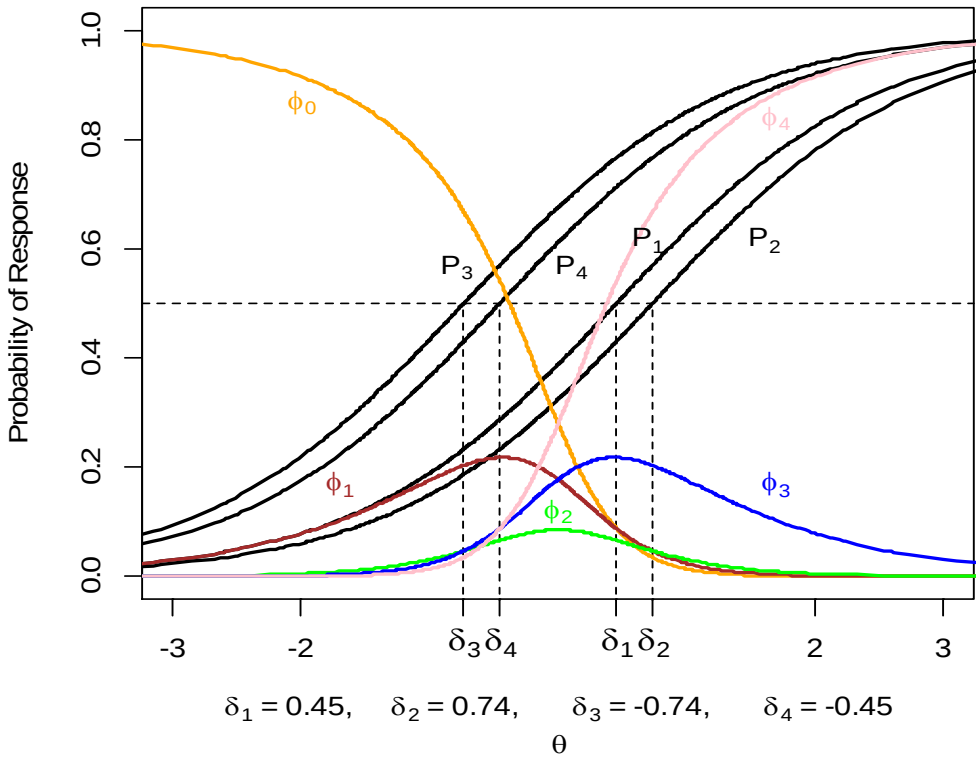

Figure 7. CCCs and threshold probabilities for a PCM item with thresholds (0.45, 0.74, -0.74, and -0.45)

The last example presented in Figure 8 depicts CCCs with δ_{i1} = -1.51, δ_{i2} = -1.64, δ_{i3} = -2.42, and δ_{i4} = -0.93 and corresponding threshold probabilities. For this item, reversals also occurs and category 1 and 2 do not have the highest probability to be selected at any level of θ. The extremely high response probability for category 4 results in all item thresholds being located at lower levels on the θ continuum. This can happen when the item content is too easy for the respondents and, therefore, most

of the respondents pass the highest threshold. Irrespective of the difficulty of the content of the item, however, if a group of respondents manifests acquiescent response style (ARS) in response to the item, this pattern of CCCs can also occur.

The CCCs plots illustrated above show that threshold distances contain important information about response category use. As a rule, if the threshold parameters are ordered within an item, every response category is the most likely option at least at one θ level. In this case, each response category is linked to an area on the latent continuum where it has a larger response probability than the other categories. In contrast, disordered thresholds indicate that certain response categories are avoided or the relation between trait and category choice is improperly specified. In this case, there is no area in which the CCCs of one or more categories are larger than the CCC of the other items.

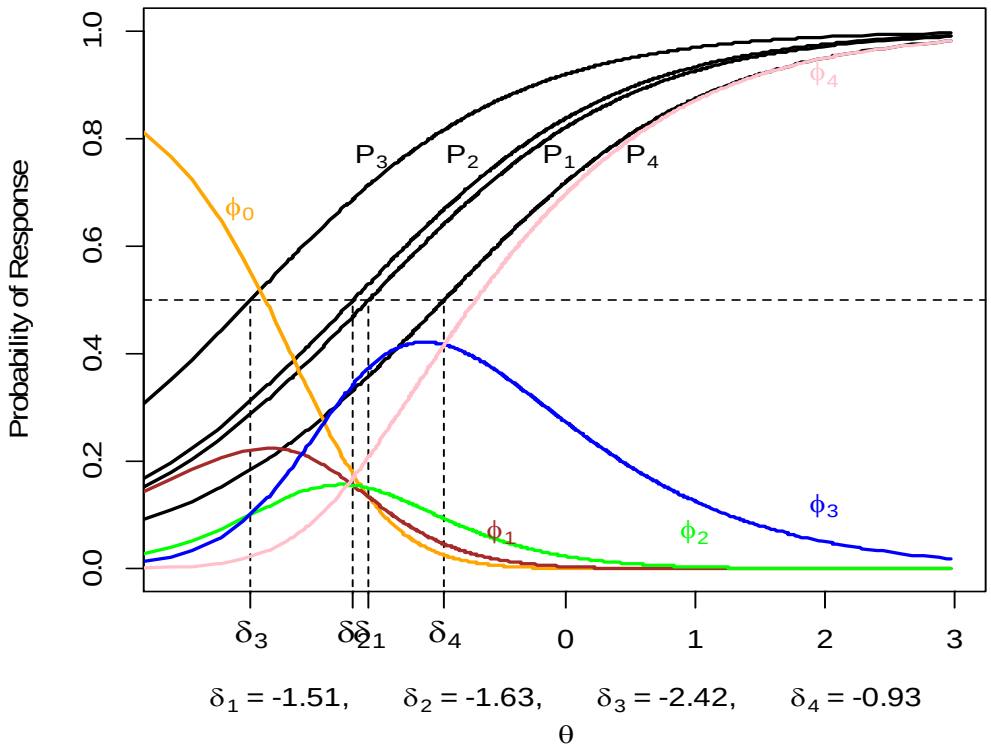

Figure 8. CCCs and threshold probabilities for a PCM item with thresholds (-1.51, -1.63, -2.42, and -0.93)

2.3 Unique Features of the Rasch Models

The models in the Rasch family are distinguished from other IRT models by a fundamental statistical characteristic: separable person and item parameters and hence sufficient statistics (Masters & Wright, 1984). It is said that a sufficient statistic exists when no other information from the data is required to estimate a parameter.

Suppose an $N \times I$ data matrix (N is the number of people and I is the number of items) with elements x_{ni} being 0 or 1 for dichotomous Rasch model cases or being the number of thresholds passed for polytomous Rasch model cases. Then, the total score

(i.e., the row sum of the data matrix, $\upsilon_n = \sum_{i=1}^{I} x_{ni}$) is a sufficient statistic for the estimation of person trait parameters (θ_n) and the item score (i.e., the column sum, $\varepsilon_i = \sum_{n=1}^{N} x_{ni}$) is a sufficient statistic for the estimation of item difficulty parameters (δ_i).

Once the sufficiency of total scores is established, the unknown parameter θ can be eliminated by conditioning on the person's total score υ during the course of item parameter estimation. All different response vectors (patterns) that yield the same total score υ have the same trait estimate. Therefore, increasing sample size does not increase the number of person parameters to be estimated and item characteristics do not have an impact on trait estimation. Consequently, the consistency of item parameter estimates can be achieved.

Also, once the sufficiency of item scores is established, by conditioning on the observed vector of item score ε , the item parameters are eliminated. This means that under the PCM, a simple count of respondents passing each threshold of an item contains all information about the threshold difficulty.

2.4 Mixture Distribution Models

The model of interest in the current study, the mixture partial credit model, (MPCM: Rost, 1991; von Davier & Rost, 1995) can be viewed as a generalization of the PCM to a finite mixture distribution model. In this section, mixture distribution is introduced followed by the latent class model (LCM), which is the simplest discrete

mixture distribution model and closely related to the MPCM. Lastly, the general idea of integrating the IRT and LC models is discussed.

2.4.1 Continuous and discrete mixture distribution

A mixture distribution refers to a composite of several subpopulation distributions (see e.g., McLachlan & Peel, 2000). The basic assumption of the model based on a mixture distribution is that the distribution of an observed random variable is not adequately described by a single probability function, but by a number of conditional probability functions.

In a research setting where the observed sample can be seen as being drawn from two or more subpopulations with distinctive features, a mixture distribution model can possibly model this heterogeneity by combining conditional probability functions across subpopulations. These subpopulations are alternatively called mixture components or latent classes. A mixture distribution can be either continuous or discrete depending on the nature of the mixing variable on which the probability is conditioned. In a general form, the continuous mixture distribution can be presented as follow:

$$f(\mathbf{x}) = \int_{-\infty}^{\infty} f(\mathbf{x} \mid \theta) d\theta ,$$

where $f(\mathbf{x})$ is the unconditional probability density of an *I*-dimensional random vector $\mathbf{x} = \{x_1, ..., x_I\}$ and is obtained by integrating over the component densities $f(\mathbf{x} \mid \theta)$ conditional on a continuous mixing variable θ. The previously reviewed RM and PCM can be viewed as continuous mixture models where individual latent trait (θ) is

a real-valued mixing variable and the component densities $f(\mathbf{x}|\theta)$ are defined as the logistic function.

If the mixing variable is discrete, only a finite number of component distributions are produced (i.e., as many as the number of latent classes) and the unconditional probability becomes a weighted sum. The general form is specified as:

$$f(\mathbf{x}) = \sum_{c=1}^{C} \pi_c f(\mathbf{x}|c), \tag{6}$$

where *c* is a discrete mixing variable whose arbitrary quantity $c = \{1,...,C\}$ classifies each respondent's latent class membership, $f(\mathbf{x}|c)$ is the component distribution conditional on latent class membership *c*, and π_c are the relative sizes of latent classes called mixing proportions, which are constrained to be $0 \le \pi_c \le 1$ and $\sum_{c=1}^{C} \pi_c = 1$. In most cases, the component distributions take on the common parametric form but have their own sets of parameters.

When data is analyzed using a discrete mixture distribution, the nature of a mixing variable does not need to be specified a priori. It is a hidden structure, so that the existence of valid latent classes is explored during the estimation process and each respondent is assigned to one of the identified latent classes according to similarity among respondents. This flexible, exploratory capability of discrete mixture distribution models allows for a way to decompose unobserved heterogeneity that would not be detected and modeled within non-mixture models.

2.4.2 Latent class model

The latent class model (LCM: Lazarsfeld & Henry, 1968) is the simplest finite mixture distribution model for item responses. The main purpose of using a LCM is to infer unobserved groups that differ in qualitative sense. Individuals within the same latent class are assumed to behave similarly on relevant behavior while members of different classes are assumed to behave differently.

Before presenting the model formulation of LCM, a brief comparison of the LCMs to the IRT models is useful for a better understanding of both models. First, both IRT and LC models relate a set of item responses and a latent trait variable. Also, the manifest variables, i.e., item responses are treated as discrete variables in both models. The major difference between the two models, however, revolves around the conceptualization of the person trait distribution. The IRT models assume person trait as continuous and provide measures of the trait on a single latent continuum. In addition, respondents in a sample are assumed to come from a qualitatively homogeneous single distribution and, thus, the respondents are different in quantitative sense. On the other hand, in the LCM the respondents are different in qualitative sense. The LCMs treat the person trait as a discrete variable and provide mutually exclusive and exhaustive latent class membership. Within each latent class there is no variation in the item response probability.

The general LCM can be presented by specifying the component distribution with the joint probability function of item responses under the local independence assumption:

$$p(\mathbf{x}) = \sum_{c=1}^{C} \pi_c \prod_{i=1}^{I} p_{ic}^{x} (1 - p_{ic})^{1-x} ,$$

where $p(\mathbf{x})$ is the probability of a response pattern of items $i=\{1,..., I\}$, π_c are the mixing proportions, and p_{ic}^{x} and $(1-p_{ic})^{1-x}$ are the probability of a success and a failure on item i in class c, respectively. Both π_c and p_{ic}^{x} are the model parameters to be estimated.

2.4.3 Mixture IRT models

By integrating a standard IRT model with the LCM, a mixture IRT model is obtained. The integration means that the response probability is now conditional on both respondent's continuous trait distribution (following the IRT models) as well as discrete trait distribution (following the LC models). Therefore, the unconditional probability of an item response pattern **x** for mixture IRT model is:

$$p(\mathbf{x}) = \sum_{c=1}^{C} \pi_c \int_{\theta} \prod_{i=1}^{I} p_i(x_i \mid \theta, c)\ f_c(\theta) d\theta , \qquad (7)$$

where $f_c(\theta)$ is the class-specific trait distribution, of which the items have different parameters.

This integration relaxes both models' assumptions, which can limit the utility of the models in applications. Specifically, the IRT model assumption that respondents in a sample belong to a qualitatively homogeneous distribution is relaxed. Mixture IRT models accommodate heterogeneous subpopulations by allowing item and/or person parameters to vary across latent classes. The differences observed in item and

person parameter estimates across latent classes may provide the ground on which the nature of population heterogeneity can be interpreted. Also, the LCM assumption that the response probability within latent classes is the same is relaxed. In mixture IRT models, each respondent is assigned an estimated latent trait level as well as a latent class membership.

In sum, in mixture IRT models, an IRT model holds within different subpopulations, but in each subpopulation a different set of item and person parameters can be estimated. The mixture IRT models provide a statistical tool to detect and simultaneously model two types of population heterogeneity i.e., quantitative differences on a continuous latent variable as well as qualitative differences on a discrete variable.

Exploration of qualitative individual differences. The major utility of mixture IRT models has been found in their capability to simultaneously model quantitative and qualitative differences among individuals. In previous studies employing different mixture IRT models, researchers identified qualitatively distinguishable latent groups in several realms of study. In cognitive assessments, Rost (1990) applied the mixture Rasch model (MRM) and identified two latent classes in which the members differed in their relative strength in subject contents of a physics test. A random guessing group was detected in a low-stakes achievement test using a mixture 2-PL model (Lau 2009), in a mathematic proficiency test using the MRM (Subedi, 2009), and in a reading proficiency test using a mixture Rasch model (Mislevy & Verhelst, 1990). A latent class in which the members present speededness at the end-items of a test was

separated using a mixture distribution version of the Bock's nominal response model in the study by Bolt, Cohen, and Wollack (2002). Mislevy and Verhelst (1990) suggested a mixture Rasch model with theory-based item parameter structures to detect problem solving strategies. In non-cognitive assessment, Reise and Gomel (1995) applied the MRM to analyze a personality scale data and found a structural difference in the personality factors between two latent classes.

In the analysis of rating scale item responses, the characteristics of latent classes were interpreted in terms of different faking tendencies (Zickar et al., 2004), self-disclosure patterns (Maij-de Meij, et al., 2005), structures of personality factor (Egberink et al., 2010; Rost et al., 1997), and response styles (Austin et al., 2006; Gollwitzer et al., 2005; Meiser & Machunski, 2008; Rost, 1991; Rost et al., 1997; Smith, et al., 2012).

2.5 Mixture Partial Credit Model

2.5.1 Presentation of model

As explained previously, by integrating the LCM and PCM, the MPCM can be derived. The model equation of the PCM defines the probability of an item response pattern $\mathbf{x}$ specified as Equation 7:

$$p(\mathbf{x}) = \sum_{c=1}^{C} \pi_c \prod_{i=1}^{I} \frac{\exp\left(\sum_{k=0}^{x} (\theta_c - \delta_{ikc})\right)}{\sum_{g=0}^{h_i} \left(\exp\left(\sum_{k=0}^{g} (\theta_c - \delta_{ikc})\right)\right)}, \; x = 0, 1, 2, \ldots, h_i, \tag{7}$$

where $p(\mathbf{x})$ is the unconditional probability of an item response pattern **x**, π_c is the mixing proportion with constraints, $0 \le \pi_c \le 1$ and $\sum_{c=1}^{C} \pi_c = 1$, and $\sum_{k=0}^{0} (\theta_c - \delta_{ikc}) \equiv 0$. Note that θ_c and δ_{ikc} are now the class specific person trait and threshold parameters, respectively. The threshold parameters δ_{ikc} are constrained to be $\sum_{i=1}^{I} \sum_{k=0}^{x} \delta_{ikc} = 0$ for all *c* for the model identification purpose.

2.5.2 Parameter estimation

In Section 2.3, the particular feature of Rasch family models i.e., the sufficiency of the total scores for the θ estimation is explained. The total scores (υ_n) obtained from a sample are simply used to eliminate person parameters (θ_n) in estimating item parameters. The property of the sufficient statistic, however, cannot be applied as straightforwardly for the mixture Rasch models as it can for the RM and PCM. That is because latent classes are not known and thus the total scores in each class are not directly observable. As a solution, an estimated quantity for $\pi_{\upsilon|c}$, namely latent score probability, which is the probability of a total score appearing in a class, needs to be introduced. This probability is treated as a model parameter and estimated along with other model parameters. Given that the number of parameters needed to estimate the latent score distribution grows easily as the number of classes and items increases, parsimonious ways to approximate it have been proposed. Software *mdltm* (multidimensional discrete latent traits models: von Davier, 2005a), which is used for the parameter estimation in the current study, uses a 2-parameter log-linear smoothing

approach to parameterize this score distribution. Appling this approach, the distributional model-based score probability ($\hat{\pi}_{\upsilon|c}$) can be obtained:

$$\hat{\pi}_{r|c} = \frac{\exp\left(\frac{\upsilon}{\upsilon_{\max}}\mu_c + \frac{4\upsilon(k-\upsilon)}{\upsilon_{\max}^{\ 2}}\sigma_c\right)}{\sum_{s=0}^{\upsilon_{\max}}\exp\left(\frac{s}{\upsilon_{\max}}\mu_c + \frac{4s(\upsilon_{\max}-s)}{\upsilon_{\max}^{\ 2}}\sigma_c\right)}, \tag{8}$$

where $\upsilon = 0,...,\upsilon_{\max}$, μ_c is the location parameter indicating the average of θ s and σ_c is the variability of that distribution. The obtained score probabilities provide a smoother distribution of expected score frequencies and will be replicated in approximately identical shape in different samples of respondents. This distribution is flexible in terms of the shape that it can take on, so that various shapes of score distributions can be modeled.

One of the benefits of introducing this distributional approximation that uses only two parameters is that it prevents a penalizing factor of the information criteria for model selection from unnecessarily increasing. The details related to this issue of model selection are further addressed in Section 2.5.4. More details about this logistic model for score frequency can be found in Rost and von Davier (1995) and Rost (1997).

In *mdltm*, the Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) is implemented to obtain marginal maximum likelihood (MML) estimates. The MML method makes use of the following factorization:

$$p(\mathbf{x} \mid c) = p(\mathbf{x} \mid c, \upsilon) \cdot p(\upsilon \mid c), \tag{9}$$

where $\upsilon = \sum_{i=1}^{I} x_i$ is the total score and the conditional total score $p(\upsilon \mid c)$ is replaced with the estimated $\hat{\pi}_{\upsilon|c}$ as explained above. By applying the property of the sufficient statistic, the pattern probability conditional on total score instead of estimated θ can be obtained as follows:

$$p(\mathbf{x} \mid c, \upsilon) = \frac{\exp(-\sum_{i=1}^{I} x_i \delta_{ic})}{\gamma_{\upsilon|c}(\exp(-\delta_{\cdot\cdot c}))}, \tag{10}$$

where the denominator $\gamma_{\upsilon|c}(\exp(-\delta_{\cdot\cdot c}))$ is a class-specific symmetric function of the thresholds. It makes the computation of all possible combinations of item parameters that yield a total score and is also required in the E-steps of the item parameter estimation for computing the expected pattern frequencies. An illustration of this computation for the RM difficulty parameters can be found in Baker and Kim (2004, Ch.5). Finally, the full formulation of the pattern probability with person parameter eliminated is as follows:

$$p(\mathbf{x}) = \sum_{c=1}^{C} \pi_c \hat{\pi}_{\upsilon|c} \frac{\exp(-\sum_{i=1}^{I} x_i \delta_{ic})}{\gamma_{\upsilon|c}(\exp(-\delta_{\cdot\cdot c}))}.$$

E-steps. In the expectation steps, the expected pattern frequencies in each latent class are computed on the basis of the observed pattern frequencies and preliminary estimates of the threshold parameters. A randomly selected value can be

used as a starting parameter values for the first iteration. For the subsequent iterations, the estimates of the previous M-step are used. The expected class-specific pattern frequency $\hat{n}(\mathbf{x}|c)$ is a proportion of the ratio of the pattern probability in a class $p(\mathbf{x}|c)$ and the unconditional observed pattern probability $p(\mathbf{x})$:

$$\hat{n}(\mathbf{x}|c) = n(\mathbf{x})\frac{\pi_c p(\mathbf{x}|c)}{p(\mathbf{x})},$$

where $n(\mathbf{x})$ is the observed frequency of response pattern $\mathbf{x}$, the conditional pattern frequency $p(\mathbf{x}|c)$ is defined by Equations 8, 9, and 10. $p(\mathbf{x})$ is the unconditional observed pattern probability i.e., $\sum_{c=1}^{C}\pi_c p(\mathbf{x}|c)$.

M-steps. The expected pattern frequencies for each latent class obtained from the E-step are used in the M-step for the computation of the estimates of the model parameters, π_c, $\pi_{\upsilon|c}$, and δ_{ikc}. These parameters are estimated separately for each class by maximizing the log-likelihood function of class *c*. The log-likelihood function of class *c* may be specified as follows:

$$\ln L_c = \sum_{\mathbf{x}}\hat{n}(\mathbf{x}|c)\left(\ln\pi_{\upsilon|c} - \sum_{i}^{I}\delta_{i x_i c} - \ln(\gamma_{\upsilon|c}(\exp(-\delta_{\cdot\cdot c})))\right).$$

Solving the first derivative to be zero with respect to the threshold parameter yields the (revised) estimate for threshold *k* on item *i* in class *c* as follows:

$$\hat{\delta}_{ikc} = \ln\frac{n_{ikc}}{\sum_{\upsilon=0}^{\upsilon_{\max}} m_{\upsilon|c}\dfrac{\gamma_{\upsilon-1,i|c}}{\gamma_{\upsilon|c}}},$$

where n_{ikc} is preliminary estimates of the number of individual with a response to category *k* on item *i* in class *c*, $m_{\upsilon|c}$ is the number of individuals with score υ in class *c*, and $\gamma_{\upsilon-1,i|c}$ are the symmetric functions of order $\upsilon-1$ of all item parameters except item *i* in class *c*. This symmetric function is iteratively calculated by means of preliminary threshold parameter estimates and revised estimates in each M-step.

The estimates of the mixing proportions ($\hat{\pi}_c$) and conditional score probability ($\hat{\pi}_{\upsilon|c}$) do not need to be calculated iteratively. They can be simply calculated as follows:

$$\hat{\pi}_c = \frac{n_c}{N},$$

$$\hat{\pi}_{\upsilon|c} = \frac{m_{\upsilon|c}}{n_c},$$

where n_c is the number of respondents in class *c*.

θ estimation. During the item parameter estimation, trait parameter θ has been eliminated from the equation. In the final stage of the estimation, the unknown parameter θ can be estimated by solving iteratively the following estimation equation:

$$\upsilon_n = \sum_{i=1}^{I} \frac{\exp(\theta_{nc} - \hat{\delta}_{ikc})}{1+\exp(\theta_{nc} - \hat{\delta}_{ikc})},$$

where $\hat{\delta}_{ikc}$ is the final estimate of *k*th threshold for item *i* in class *c*. Respondent *n* has the trait estimate $\hat{\theta}_{nc}$ under the condition that he or she belongs to class *c* and hence there are as many $\hat{\theta}_{nc}$ as the number of *c* for each respondent. However, these

conditional trait estimates of a single individual usually do not differ much from one class to another because the estimates depend mainly on υ_n, which is the same in all classes (Rost, 1997).

2.5.3 Assigning latent class membership

As the outcomes of the simultaneous modeling of a continuous and a discrete latent variable, each respondent is assigned latent trait estimates as well as probabilities for membership in each latent class. The probability of class membership can be estimated using Bayes' theorem:

$$p(c \mid \mathbf{x}) = \frac{\pi_c \, p(\mathbf{x} \mid c)}{\sum_{c=1}^{C} \pi_c \, p(\mathbf{x} \mid c)}.$$

where $p(c \mid \mathbf{x})$ is the posterior probability of class membership *c* given the item response pattern $\mathbf{x}$. Note that the mixing proportion plays the role of prior probability in the Bayes' theorem and the estimated conditional pattern probability $p(\mathbf{x} \mid c)$ replaces the likelihood and the denominator indicates the total probability. The actual classification is carried out by first using the Bayes' theorem to compute the estimated probability for class membership given each response pattern. Then, respondents may be assigned to the latent class for which the conditional probability of their membership is largest.

2.5.4 Determining the number of latent classes

In the MPCM formulation, the number of latent classes (*C*) is not a model parameter and, thus, must be specified before initiating the parameter estimation

process. Under conditions of uncertainty about the "true" number of unknown subpopulations, the commonly used technique to determine the number of latent classes is to compare the likelihood function of competing models with increasing numbers of latent classes and then choose a model that an information criterion data-model fit indicates as the best- fitting model to the data. Although significance tests are not possible with these indices, comparing the index values for competing models provides some degree of evidence for the nature of trait variable structure.

Information criteria. Many information criterion statistics have been developed under the minimum complexity criteria. Frequently referred information criteria include Akaike's information criterion (AIC: Akaike, 1974), Bayesian information criterion (BIC: Schwarz, 1978) and consistent AIC (CAIC: Bozdogan, 1987). The three statistics are those provided by *mdltm*.

The AIC index can be calculated based on *H* different models being compared:

$$AIC_h = -2\ln(L_h) + 2Par_h,$$

where L_h is the maximum of the likelihood function of the *h*th model and Par_h is the number of independent parameters that are estimated when fitting the *h*th model to the data. In comparing competing models, the model *h* that shows the minimum AIC value is chosen as the model that best fits the data and therefore is considered as the preferred model. It is seen in the equation that when two models have similar maximum likelihood value (L_h), a smaller value of AIC will be associated with the model based on fewer parameters. In this way, AIC prefers a model with less

complexity, in other words, a more parsimonious model. A criticism of the AIC is that it lacks properties of asymptotic consistency because the definition of the AIC does not directly involve the sample size. Consequently, as sample size increases a more complex model would be more likely to be selected based on the AIC.

Schwarz (1978) developed the BIC, which is an asymptotically consistent measure. The computation of the BIC may be specified as follows:

$$BIC_h = -2\ln(L_h) + \ln(N) \times Par_h ,$$

where *N* denotes the sample size. In the same way as is done for AIC, a model *h* that shows the minimum BIC value is chosen as the preferred model. Note that the penalty term for the BIC is larger than for the AIC if the sample size *N* is 8 or greater, which can be seen by the fact that the value of $\ln(8) = 2.08$. Therefore, for reasonable sized samples, the BIC tends to select less complex models (i.e., the solution with a smaller number of classes) than does the AIC.

Bozdogan (1987) extends the AIC to make it asymptotically consistent and to be penalized for over-parameterization more stringently. The CAIC index is computed as follows:

$$CAIC_h = -2\ln(L_h) + (\ln(N) + 1) \times Par_h .$$

Compared to the AIC and BIC, the penalty term for CAIC is even larger, leading to solutions that favor the selection of less complex models than are obtained with the AIC or BIC.

Based on the specific penalty weights, it is expected that different information criterion statistics may lead to different solutions in mixture IRT models. The preference of a more complex model by the AIC may result in over-identification problems under certain conditions whereas the preference of a less complex model by the BIC and CAIC may cause under-identification problems. The relative effectiveness of information criteria has been investigated via simulation studies, where the true conditions are known and hence it is possible to monitor the behavior of information criterion statistics in identifying the correct model.

Model selection in mixture IRT models. There are a limited number of simulation studies on model selection indices in mixture IRT models and all of those studies examined only models for dichotomous responses. No study has thus far investigated the problems of model selection in mixture polytomous IRT models. The following presents the findings from the studies related to dichotomous models.

The first study appearing in literature was one by Li, Cohen, Kim, & Cho (2009), in which a Bayesian estimation approach was used. Their study investigated five different model selection indices including the AIC and BIC, and compared the relative effectiveness of them under 1-, 2-, and 3-PL model with 1-, 2-, 3-, or 4-latent classes. In general, the results showed that the BIC performed the best in terms of detecting correct number of latent classes. For 1-, 2-, and 3-class simulated data, the BIC was accurate in identifying the correct number of classes in every case. However, when the simulated data had 4 classes, it apparently became more difficult for the BIC to distinguish the correct model for the 3PL model. In this case, the BIC tended to

select the simpler model. The result for the AIC showed that the AIC selects more complicated models, particularly when the true model is the 1PL model.

Cho, Jiao, and Macready (2012a, 2012b) investigated the relative effectiveness of AIC and BIC in the context of mixture Rasch and mixture 2-PL model with two classes when marginal maximum likelihood estimation was applied. The studies manipulated qualitative heterogeneity in various ways by setting different sets of item parameter profiles across latent classes and evaluated the correct model selection rates. When more distinctive heterogeneity was generated between two classes causing class separation to be large, the BIC selected the correct model almost perfectly. Under the conditions where the heterogeneity manipulated was small, the BIC under-extracted latent classes while the AIC still tended to over-extract latent classes.

Preinerstorfer and Formann (2012) reported similar results within a conditional maximum likelihood estimation context. They found that the BIC generally performed more accurately than the AIC and that longer test length was positively associated with the correct model selection rate.

2.6 Applications of the MPCM to Study of Response Styles

In Sections 1.2 and 2.4.3, previous empirical studies in which mixture IRT models were employed were briefly introduced. In Section 2.6.1, the findings in the empirical studies related to the differences in response category use and the correction of test score bias are reviewed. Section 2.6.2 summarizes the previous simulation study that investigated the model performance of the MPCM.

2.6.1 Real data analysis

Rost et al. (1997) applied the MPCM to the analysis of NEO-FFI scales and reported the results for the Conscientiousness (*C*) and Extraversion (*E*) scales. For the *C* scale, the item locations across two identified latent classes were not significantly different, which indicated that the items measured the same psychological construct across the latent classes. However, when the thresholds were examined, the larger latent class ($\pi = 0.67$) showed a set of ordered and relatively evenly spaced thresholds for all items while the smaller latent class ($\pi = 0.33$) showed that the first threshold distance was about four times larger than the second threshold distance. The threshold distances in the smaller class indicated that it was very easy to pass the first threshold and very hard to pass the last threshold and, hence, most people in this class responded to the middle categories and avoided the extreme categories. Integrating these findings in item locations and thresholds distances, the authors concluded that the difference characterizing the two latent classes was not in the conscientiousness construct but in the respondent's differential use of response categories. When the *E* scale was analyzed, however, a structural difference in the personality construct as well as the response style difference emerged. The comparison of the item locations based on a two-class model solution revealed that the two identified latent classes reflected a structural difference between sociability and impulsivity. The subsequent MPCM analyses were conducted for these two classes separately and the same pattern of thresholds differences as was presented for the *C* scale was manifested.

Eid and Rauber (2000) applied the MPCM to analyze data from an organizational survey and demonstrated how mixture models could be used to detect measurement invariance caused by response styles. In their analysis, a two-class solution was selected as the best-fitting model based on the BIC. The item location parameters did not differ much between the two latent classes. The differences were observed with respect to the threshold parameters. In the larger latent class (Class 1 with $\pi = 0.71$), all thresholds were ordered indicating that the members of this class used the rating scale in the expected way. Similar to the case depicted in Figure 4, each response category corresponded to an area on the latent continuum for which its response probability was larger than the probabilities of the other categories. In the smaller latent class (Class2 with $\pi = 0.29$), the first two thresholds were disordered for all items and the threshold distances were much smaller than in Class 1. Therefore, the members of Class 2 were characterized as extreme respondents.

Eid and Rauber (2000) also investigated whether latent classes differing in their response styles could be characterized by external variables including age, sex, length of service, length of service on the same position, and leadership level. The results showed that significantly larger proportion of female employees belonged to Class 2. In addition, relatively new employees belonged significantly less frequently to Class 1. People who had been working longer than 10 years in the same position had a higher probability for belonging to Class 2. Finally, employees at different leadership levels showed differences in the probability to belong to each latent class.

Gollwitzer et al. (2005) applied the MPCM to analyze the three anger expression subscales (Anger-in, Anger-out, and Anger-control) of the State-Trait Anger Expression Inventory (STAXI; Spielberger, 1988) obtained from patients hospitalized in a psychosomatic clinic. They observed considerable differences in response styles, which were similar to the differences in non-clinical samples. The largest latent class (Class 1) exhibited ordered and evenly spaced thresholds for both gender group and for all scales, meaning an appropriate use of response categories. It was also shown that respondents who were assigned to Class 1 on one scale were likely to be assigned to Class 1 on the other scales. The second latent class (Class 2) for the female sample presented partly disordered thresholds and narrower threshold distances. The logistic regression analyses were conducted to predict the latent class membership using various personality variables measured by Freiburg Personality Inventory (FPI-R; Fahrenberg, Hampel, & Selg, 1989). The regression analysis results provided some evidence that a social desirable tendency accounted for the response styles identified in Class 2.

Gollwitzer et al. (2005) argued that it was not reasonable to compare all individuals quantitatively with respect to their sum scores, which was the scoring method instructed in the STAXI's handbook (Spielberger, 1988). They suggested a more appropriate scoring strategy that required a two-step procedure. In the first step, individuals would have to be assigned to a latent class in order to qualify differential response styles. They could then be compared with each other within their latent class.

In a second step, class-specific person parameters could be compared across latent classes under the premise that the same trait is being measured in all classes.

Zickar et al. (2004) conducted an experimental study in which respondents were couched to respond honestly or faked positively on a personality inventory. They analyzed the item responses from the experimental sample with the MCPM and found that honestly responding group exhibited the thresholds that were properly ordered and much lower item-level scores than the "faking group". For the faking group, the thresholds were disordered and the difference between the first and second thresholds was much smaller than the difference in the honestly responding group, indicating that few individuals chose the first and second categories in this group.

Zickar et al. (2004) also compared the item responses on the Personal Preferences Inventory (PPI: Personnel Decisions International, 1997) between an applicant group and an incumbent group in an organization. Their MPCM analysis results showed that 27.6% of the applicants were in the extreme faking class whereas 13.7 % of the incumbents belonged to this class. Conversely, 26.5% of the applications were in the honestly responding class. These findings provided some insights that the typical applicant-incumbents comparison assuming that applicants were faking and incumbents were responding honestly had been too restricted.

Smith et al. (2012) analyzed data from the *Beliefs and Attitudes About Memory Survey* (BAMS: Brown, Garry, Silver, & Loftus, 1997) with the mixture Rasch models to investigate the functioning of the "Neutral" category (i.e., middle category) by examining the threshold ordering. Smith et al. (2012) pointed out that disordered

thresholds occur: *i*) when the rating scale includes more categories than the respondents can reliably distinguish, *ii*) when some rating categories are unlabeled, or *iii*) when rating scale includes middle point labeled as undecided or neutral. The analyses of the original 5-point Likert-scale BAMS data showed that disordered thresholds mainly occurred around the "Neutral" category. They treated responses to the "Neutral" category as missing data and reanalyzed the remaining data recoded to an ordered 4-point scale. For each of the three 4-point BAMS subscales, two latent classes were identified based on the CAIC. For the Blending of Memories subscale and the New Born, Womb, and Previous Lives Memories subscale, respondents from each of the latent classes used the items differently, resulting in an item difficulty ordering that was not invariant across latent classes. This indicated that different constructs related to the beliefs about memories might be measured within each latent class. For the Memory Storage subscale, however, the overall item difficulties were approximately the same for both classes except for one item. This led the author to reasonably assume that the same underlying constructs were being measured across the latent classes.

Adjustment of response style effects on test scores by applying the MPCM. Rost et al. (1997) pointed out that the estimated trait parameters of the MCPM are automatically corrected for the effects of a response style and this is the most practical implication of employing the MPCM to the analysis of self-report data. Given that the MPCM provides θ estimates conditional on each response-style class and the sum score is the sufficient statistics for θ estimation, any differences observed in the class-

specific θ estimates for the same raw score can be viewed as an adjustment or correction for the effects of response styles (Rost et al, 1997).

Rost et al. (1997), Gollwitzer et al. (2005), and Smith et al. (2012) graphically showed the relation between sum scores and θ estimates in each latent class to demonstrate how the class-specific person traits estimated for the same sum score differ across the classes. The results of those studies commonly showed that respondents who responded to more extreme categories earned less extreme theta estimates than the respondent with the same sum score but moderate response styles. These results implied that interpreting sum score difference among individuals without considering their response styles may lead to false inferences concerning individual differences in their latent trait level.

Although the potential of rectifying score bias by employing the MPCM was demonstrated in those empirical data analytic studies, it has not been investigated how the correction would operate for different types of response styles when multiple kinds of response styles are present.

Related to the correction of sum score bias, an important psychometric issue of interest is whether theta estimates obtained with a mixture IRT model may provide a better prediction of an external criterion, compared to the theta estimates obtained with its non-mixture counterpart. Maij-de-Meij et al. (2008) applied the mixture nominal response model and the MPCM to personality inventory scales, Extraversion (*E*) and Neuroticism (*N*), and investigated whether theta estimates provided by the mixture models resulted in a better prediction of relevant external criteria. The results of this

study showed that for *N* scale, the correlations between theta estimates and criterion measures were higher for the mixture models than for the non-mixture model. However, this improvement was not observed for the *E* scale.

2.6.2. Simulated data analysis

As reviewed in previous sections, there have been increasing applications of the MPCM. Unfortunately, however, little is known about model performance of the MPCM in accurately estimating the model parameters. Only one simulation study conducted by Rost (1991) demonstrated the capability of the MPCM to "unmix" heterogeneous item responses data. Rost (1991) created three sets of data, each of which was comparable with the PCM, and selectively combined two of the three data sets to generate several mixtures of two latent classes. In generating the mixture data sets, he manipulated sample size, threshold distance, and the ranges of θ, so that the mixtures differed with respect to "degree of heterogeneity". Specifically, the largest first data set (N = 1000) had a wide range of item locations (-2.7 to +2.7), equal threshold distances ($\delta_{i1} - \delta_{i2}$ = 0.5 and $\delta_{i2} - \delta_{i3}$ = 0.5), and a wide-range of θ values (-2.5 to +1.0). The second data set (N = 600) had a smaller range of item locations (-1.8 to +1.8), reversed thresholds with extremely unequal threshold distances ($\delta_{i1} - \delta_{i2}$ = 1.4 and $\delta_{i2} - \delta_{i3}$ = 0.2), and a narrow-range of θ values (-1.0 to +1.0). The third data set (N=800) had no variation of item locations (0 for all items), large and equal threshold distances ($\delta_{i1} - \delta_{i2}$ = 1.0 and $\delta_{i2} - \delta_{i3}$ = 1.0), and a narrow-range of θ values (0 to1.5). In this study, depending on the manipulated degree of heterogeneity,

the difficulty in detecting latent classes in mixture distributions was anticipated. The mixture of the first and second data sets was expected to be easiest to unmix because the item parameters and threshold distances differ strongly while the mixture of the second and third data sets was expected to be most difficult to unmix.

The accuracy of thresholds recovery, mixing proportion recovery, and class-specific mean score recovery from a single replication result was evaluated by comparing the results for mixture data with those for non-mixture data. Results showed that the mean threshold distances and the class-specific score distributions were recovered fairly well. Some large deviations from the simulated condition were observed for the mixing proportions under certain conditions. These deviations, however, were interpreted as effects of the particular threshold sets manipulated not as a bias of the estimation procedure. Rost (1991) also evaluated the quality of estimates for the mixture with three-classes and found that the accuracy of the parameter recovery for the three-class model was comparable with the estimates in the two-class model. Regarding the model selection procedure, the AIC correctly identified the generated number of latent classes.

Chapter 3: Methodology

3.1 Objectives and Research Questions

The major objective of the current study is twofold: (*i*) to evaluate the quality of the respondent classification as well as item and person trait parameter recovery of the MPCM when the population is a mixture of different response-style respondents, and (*ii*) to investigate how the MPCM makes an adjustment of the latent trait estimates to compensate for the confounding effects of different response styles on test scores. In addition to the major goals, the current study also explores the effectiveness of the information criterion statistics in identifying the correct number of latent classes in the MPCM. These objectives were addressed via a simulation study. The manipulated factors for which the effects were assessed were type of mixture of response styles, mixing proportions, sample size, and test length. The specific research questions that were addressed in this study are as follows:

1. What percentage of respondents does the MPCM correctly classify within their true response-style class under various conditions?
2. What percentage of replications does the information criterion statistics identify the correct number of latent classes?
3. What degree of the accuracy of thresholds parameter recovery does the MPCM provide under various simulation conditions when the accuracy is assessed by Pearson *r*, root mean square error, and standard error of estimates?

4. What degree of the accuracy of person trait parameter recovery does the MPCM provide under various simulation conditions when the accuracy is assessed by bias, Pearson *r*, and root mean square error?
5. How are sum (total) scores and class-specific person trait parameters estimated with the MPCM related to each other under the simulated types of mixture distribution?

3.2 Overview of Simulation Study

3.2.1 Manipulated factors

The current simulation study selectively considered the five different types of response-style mixture distribution: (*i*) ORS and ERS, (*ii*) ORS and MRS, (*iii*) ORS and ARS, (*iv*) ORS, ERS, and MRS, as well as (*v*) ORS, ERS, MRS, and ARS.

The mixing proportions were manipulated to be equal or unequal. The "equal" condition represents the population where different response-style respondents are mixed with equal proportions and the "unequal" condition represents the population where majority of the respondents are ORS respondents and very small proportion of respondents presents distorted response styles. Table 1 provides a summary of the types of mixture and mixing proportions manipulated in the current study.

Table 1. *Manipulated Simulation Conditions of Population Heterogeneity*

	Mixing proportions	Class1(π_1)	Class2(π_2)	Class3(π_3)	Class4(π_4)
1	Equal	ORS(1/2)	ERS(1/2)		
2		ORS(1/2)	MRS(1/2)		
3		ORS(1/2)	ARS(1/2)		
4		ORS(1/3)	ERS(1/3)	MRS(1/3)	
5		ORS(1/4)	ERS(1/4)	MRS(1/4)	ARS(1/4)
6	Unequal	ORS(9/10)	ERS(1/10)		
7		ORS(9/10)	MRS(1/10)		
8		ORS(9/10)	ARS(1/10)		
9		ORS(8/10)	ERS(1/10)	MRS(1/10)	
10		ORS(7/10)	ERS(1/10)	MRS(1/10)	ARS(1/10)

Note: π_c = mixing proportion for class *c*, ORS = ordinary response style, ERS = extreme response style, MRS = middle-category style, ARS = acquiescent response style

Two other manipulated factors were sample size and test length. Sample sizes were chosen at three levels, medium (*N*=1200), moderately large (*N*=3000), and large (*N*=6000). As for test length, since it is common that a psychological instrument has a small number of items per subscale, as small as 4-item (*I*=4) was explored as well as moderate number of items (*I*=10) and large number of items (*I*=20). These four manipulated factors were completely crossed resulting in the total number of ninety simulation conditions.

3.2.2 Fixed factors

Three factors, i.e., the number of response categories, latent trait distribution within latent class, and item locations were fixed in the current study. First, the number of response categories was fixed at five. Second, the latent trait distribution was generated to be a normal distribution with the mean of 0 and the standard

deviation of 1 for each latent class. Third, the item location of item *i* held invariant across the ORS, ERS, and MRS class in order for the latent classes to differ only with respect to the dispersion of item responses (Rost et al, 1997). For the ARS class, however, the generated item location for item *i* was not the same as that for the other response-style classes. The high response probability for the category 3 and 4 of a positively worded item *i* resulted in a very low item location for that item. Similarly, very high item locations for the negatively worded items were resulted. As this simulated condition for the item parameters in the ARS class indicates, if there is a group of ARS respondents in a sample, non-invariant item locations are likely to be manifested in a latent class.

3.2.3 Response scale

The current study assumed that item responses were obtained with a five-category Likert-scale that had a built-in balanced scale. In the balanced scale, a pair of items asked an equivalent construct in a positive as well as a negative statement. In scoring the category responses, responses to negatively worded items were reversely coded before being analyzed. For example, an endorsement of the category 4, 'strongly agree' on these items was scored as 0 and an endorsement of the category 1, 'disagree' as 3. Using reversely coded category responses, instead of raw responses, affected the marginal distribution of category responses for ARS class. The raw response frequency distributions for the ARS class would be negatively skewed for all items before recoding responses. After the recoding process, however, the category

response frequency distributions for negatively worded items were positively skewed as can be seen in Figure 9.

3.3 Data Generation

The rating scale item responses that were confounded by the effects of response styles were generated based on the relation between threshold values and response category probabilities defined in the PCM. The common method of generating thresholds such as randomly selecting threshold values within certain range of θ distribution, would not produce the item responses that characterize ERS, MRS, or ARS. The subsequent section presents the details of how to determine the population generating thresholds for each response-style subpopulation. The generation of item responses is then followed.

3.3.1 Population generating thresholds

The first step was to clearly delineate distinguishing features of the four response-style classes by presuming marginal frequency distributions of category responses for each response-style class. Figure 9 presents the expected frequency distributions of category responses marginalized over all items administered in four different response-style classes. The specific probability values are presented in Table 2. For example, assuming that theta distribution is a normal distribution, 14% of ORS respondents would choose ‘strongly disagree’, 22% ‘disagree’, 28% ‘neither disagree nor agree’, 22% ‘agree’, and 14% ‘strongly agree’ on average over all items. If a group of people has ERS tendency, about 81% of them would select ‘strongly disagree’ or ‘strongly agree’. In determining these marginal probabilities, a rather

arbitrary decision was made because there was neither theoretical grounded nor empirically reported category response frequencies related to the response styles. Since too sparse category response frequencies cause problems in estimation, extremely small category response frequency (i.e., near zero percent) for any item was avoided.

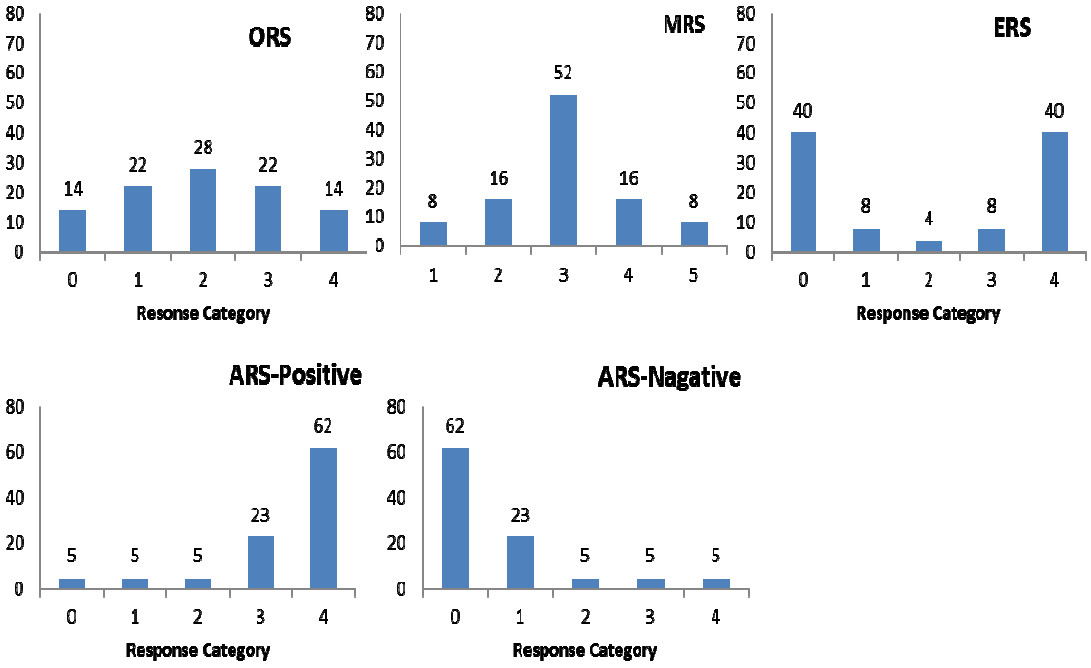

Figure 9. Expected marginal frequency distributions of category responses for different response styles (%)

Table 2. *Expected Marginal Category Probabilities for Different Response-style Classes*

	Category0	Category1	Category2	Category3	Category4
ORS	0.14	0.22	0.28	0.22	0.14
ERS	0.40	0.08	0.04	0.08	0.40
MRS	0.08	0.16	0.52	0.16	0.08
ARS(positive)	0.05	0.05	0.05	0.23	0.62
ARS(negative)	0.62	0.23	0.05	0.05	0.05

Note: ORS = ordinary response style, ERS = extreme response style, MRS = middle-category response style, ARS = acquiescent response style.

The second step was to make variation of the category probabilities among items. As shown in Table 3, while ensuring the marginal category probabilities approximate the values initially specified in Table 2, the category probabilities for each item were manipulated to be different among items. Table 3 shows the variations created for ten items for the ORS class. The category probabilities for individual items for the other response styles are presented in the Appendix A.

Table 3. *Category Probabilities for Individual Items for ORS Class*

Item	Category1	Category2	Category3	Category 4	Category 5
1	0.1478	0.2245	0.2556	0.2245	0.1478
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

Note that the means of the category probabilities of the ten items remain almost the same as the marginal category probabilities specified in Table 2. These variations among items were manipulated to generate item responses that fit the PCM instead of the rating scale model (RSM, Andrich, 1978), The RSM is restricted to have a common set of thresholds across all items.

The next step was to compute threshold probability for each step by applying the simple Rasch logistic model to the series of adjacent categories. The computations are demonstrated using an example of the first item in Table 3. As presented in Equation 1, the *k*th threshold for item *i* (δ_{ik}) can be obtained by computing the natural logarithm of the odds ratio and subtracting it from the person trait density:

$$\theta_n - \delta_{ik} = \ln\left(\frac{P_{nik}}{P_{ni,k-1}}\right),$$

$$\delta_{ik} = -\ln\left(\frac{P_{nik}}{P_{ni,k-1}}\right) + \theta_n . \qquad (11)$$

Ignoring the trait density (or $\theta_n = 0$) in Equation 11, δ_{ik} can be computed as follows:

	Category0	Category1	Category2	Category3	Category4
Category probability (ϕ_{ik})	0.147	0.225	0.256	0.225	0.147

	Step1	Step2	Step3	Step4
Threshold probability (p_{nik})	$\left(\begin{array}{l}\frac{\phi_{ni1}}{\phi_{n10}+\phi_{n11}} \\ =\frac{0.225}{0.147+0.225} \\ =0.60\end{array}\right)$	$\left(\begin{array}{l}\frac{\phi_{ni2}}{\phi_{n11}+\phi_{n12}} \\ =\frac{0.256}{0..225+0.256} \\ =0.53\end{array}\right)$	$\left(\begin{array}{l}\frac{\phi_{ni3}}{\phi_{n12}+\phi_{n13}} \\ =\frac{0.225}{0.256+0.225} \\ =0.47\end{array}\right)$	$\left(\begin{array}{l}\frac{\phi_{ni4}}{\phi_{n13}+\phi_{n14}} \\ =\frac{0.147}{0.225+0.147} \\ =0.40\end{array}\right)$
Odds $\left(\frac{P_{nik}}{1-p_{nik}}\right)$	$\left(\frac{0.60}{1-0.60}\right)$ $=1.53$	$\left(\frac{0.53}{1-0.53}\right)$ $=1.14$	$\left(\frac{0.47}{1-0.47}\right)$ $=0.88$	$\left(\frac{0.40}{1-0.40}\right)$ $=0.65$
Ln(Odds)	0.426	0.129	-0.129	-0.426
Threshold (δ_{ik})	-0.426	-0.129	0.129	0.426

During this thresholds computation, item locations were fixed to zero. For the items to have different levels of difficulty, a positive or negative constant was added to each threshold. The varying item difficulties manipulated are presented in Tables 4 to Table 7.

In the computation presented above, the item threshold values were computed without considering θ distribution. In IRT models, the probability of an item response is determined as a function of both item and person parameters. Therefore, the person trait density needed to be combined with the computed threshold values (Equation 11). In order to achieve this combination, a histogram that follows the normal distribution was constructed under which the determined thresholds (i.e., cut points on theta continuum) were adjusted. The procedures of this adjustment were the following: theta range from -2.5 to 2.5 was divided into nine intervals with 0.5 increments and then a

sample of 10000 respondents was allotted to each interval based on the cumulative normal density function. Using this sample of respondents and the initially computed threshold values for ten items, PCM item responses were generated. The generated item responses were analyzed to check the marginal category probabilities. While monitoring the resulting category probabilities, several sets of four constants were alternatively added to the initial threshold values until a set of thresholds that produced the expected marginal category probabilities as close as possible. Tables 4 to 7 present the threshold parameters that were obtained based on these adjustments for the ORS, ERS, MRS, and ARS class, respectively. Thresholds for ten items were first determined and those ten items were used twice to create 20-item test. Four items among the ten, which are indicated in the Tables 4 to 7, were selected to create 4-items test. The corresponding plots for the determined thresholds for ten items are presented in Figures 10 to 13. These threshold plots represent the locations of each threshold on the latent trait continuum on the y-axis. The characteristics of the sets of threshold parameters for each class are described in the subsequent sections.

Thresholds for ORS class. The population generating thresholds for the ORS class are presented in Figure 10. As seen in Figure 4 in chapter 2, which presents ordered and evenly spaced thresholds for a single item, the threshold plot in Figure 10 shows those properties across all items. In this group, it is seen that passing a higher threshold requires more of the latent trait θ.

Table 4. *Threshold Values Used for the Generation of the ORS Class*

Item	Threshold1	Threshold2	Threshold3	Threshold4	Location
1[a]	-1.5181	-0.5998	0.4998	1.4181	-0.05
2[a]	-1.2924	-0.6768	0.7768	1.3924	0.05
3	-1.8123	-0.6265	0.4265	1.6123	-0.10
4	-1.7632	-0.5510	0.7510	1.9632	0.10
5	-1.8469	-0.7522	0.4522	1.5469	-0.15
6	-1.1232	-0.4750	0.7750	1.4232	0.15
7	-1.7999	-0.7849	0.3849	1.3999	-0.20
8	-1.1654	-0.4421	0.8421	1.5654	0.20
9[a]	-1.6191	-0.9980	0.4980	1.1191	-0.25
10[a]	-1.0966	-0.7379	1.2379	1.5966	0.25
Mean	-1.5037	-0.6644	0.6644	1.5037	0

Note: [a] Selected item for 4-item test length condition

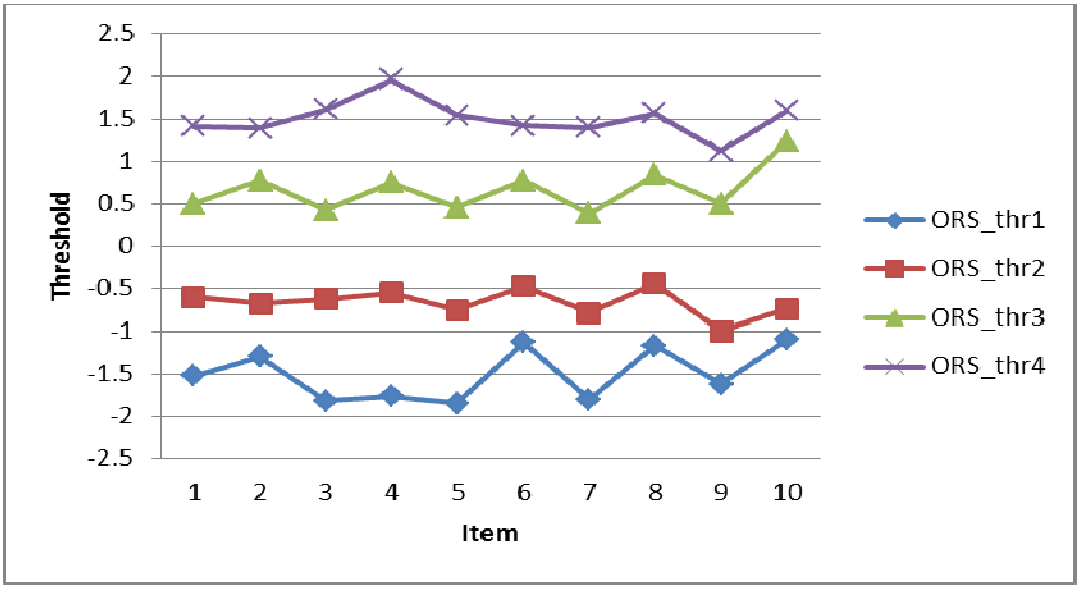

Figure 10. Thresholds plot for 10 items for ORS class

Thresholds for MRS class. The population generating thresholds for the MRS class are presented in Figure 11. As can be seen in Figure 5, the distances between

second and third thresholds are large and there are reversals between δ_1 and δ_2 as well as thresholds δ_3 and δ_4.

Table 5.*Threshold Values Used for the Generation of the MRS Class*

Item	Threshold1	Threshold2	Threshold3	Threshold4	Location
1[a]	-2.1328	-2.8106	2.7106	2.0328	-0.05
2[a]	-1.9616	-2.3956	2.4956	2.0616	0.05
3	-1.1515	-3.2403	3.0403	0.9515	-0.10
4	-0.8654	-2.3722	2.5722	1.0654	0.10
5	-1.2218	-3.1974	2.8974	0.9218	-0.15
6	-1.0001	-2.7372	3.0372	1.3001	0.15
7	-2.0917	-2.7031	2.3031	1.6917	-0.20
8	-1.3698	-2.2797	2.6797	1.7698	0.20
9[a]	-1.9528	-2.2618	1.7618	1.4528	-0.25
10[a]	-1.1903	-2.1591	2.6591	1.6903	0.25
Mean	-1.4938	-2.6157	2.6157	1.4938	0

Note: [a] Selected item for 4-item test length condition

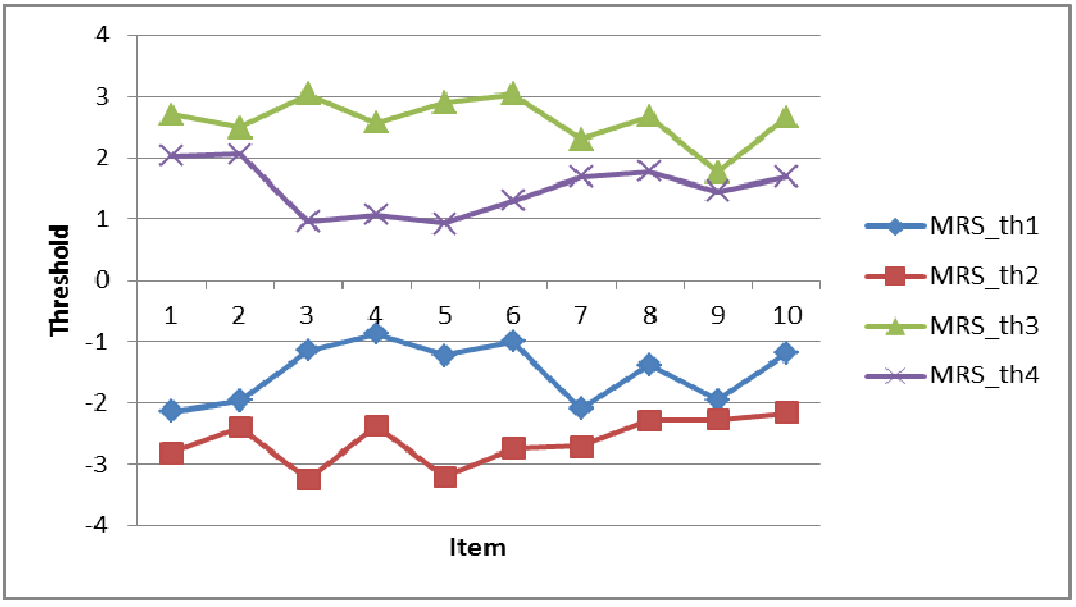

Figure 11. Thresholds plot for 10 items for the MRS class

Thresholds for ERS class. The population generating thresholds for the ERS are presented in Figure 12. As can be seen in Figure 7, the reversals occur and there are items that have no area between thresholds, indicating very sparse expected responses for some categories. Generally, the first threshold value is the greatest in this class. It indicates that it is hard for people in this class to pass the first threshold and, therefore, they end up with selecting the first category ($k = 0$) rather than the second category ($k = 1$). On the other hand, the last threshold is the easiest to pass, indicating that respondents tend to pass the last threshold easily and select the last category ($k = 4$).

Table 6. *Threshold Values Used for the Generation of the ERS Class*

Item	Threshold1	Threshold2	Threshold3	Threshold4	Location
1[a]	0.4043	0.6851	-0.7851	-0.5043	-0.05
2[a]	0.7207	0.2235	-0.1235	-0.6207	0.05
3	1.0029	-0.1963	-0.0037	-1.2029	-0.10
4	1.1222	0.6516	-0.4516	-0.9222	0.10
5	0.6489	0.1037	-0.4037	-0.9489	-0.15
6	0.8159	1.1774	-0.8774	-0.5159	0.15
7	1.1394	0.2912	-0.6912	-1.5394	-0.20
8	1.0474	0.7020	-0.3020	-0.6474	0.20
9[a]	0.246	0.2106	-0.7106	-0.7460	-0.25
10[a]	1.0952	0.9962	-0.4962	-0.5952	0.25
Mean	0.8243	0.4845	-0.4845	-0.8243	0

Note: [a] Selected item for 4-item test length condition

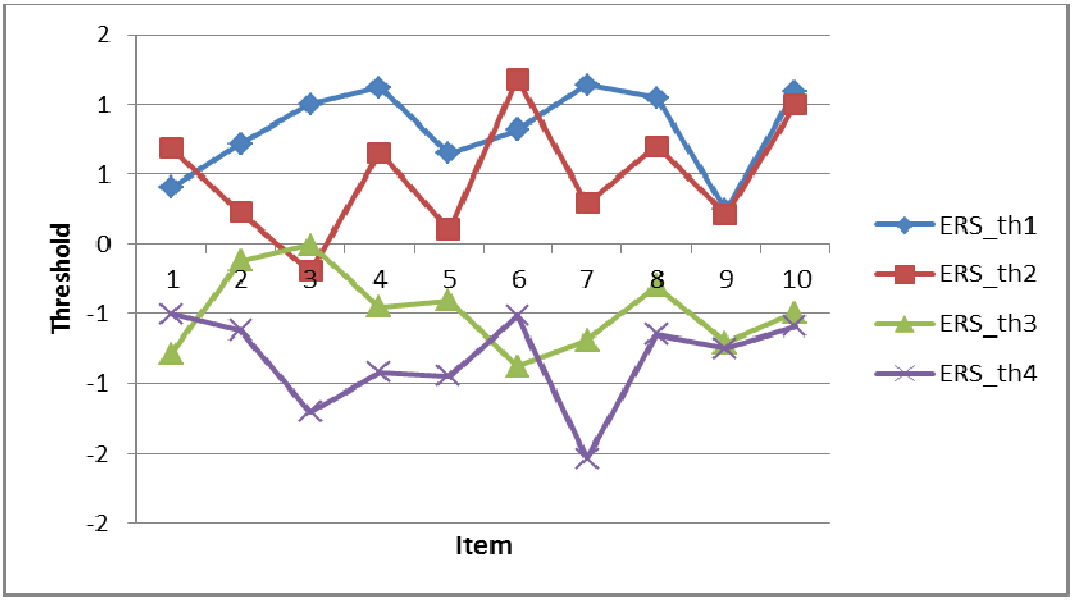

Figure 12. Thresholds plot for 10 items for the ERS class

Thresholds for ARS class. The population generating thresholds for the ARS class are presented in Figure 13. The first five items are those that are written in positive statements whereas Item 6 to Item 10 are those that are written in negative statement. The negatively stated items' thresholds profile locates upper range of theta continuum whereas the positively stated items' thresholds profile locates lower range of theta continuum.

Table 7. *Threshold Values Used for the Generation of the ARS Class*

Item	Threshold1	Threshold2	Threshold3	Threshold4	location
1[a]	-1.5092	-1.6300	-2.4202	-0.9255	-1.6210
2[a]	-1.6323	-1.0445	-2.5418	-0.8828	-1.5250
3	-1.7054	-1.6953	-2.1910	-1.2787	-1.7170
4	-1.3810	-1.2554	-2.4918	-0.6646	-1.4480
5	-1.8507	-1.5547	-2.3851	-1.2644	-1.7630
Mean	-1.6157	-1.4360	-2.4060	-1.0032	-1.6148
6[a]	0.8255	2.3202	1.5300	1.4092	1.5225
7[a]	0.9828	2.6418	1.1445	1.7323	1.6235
8	1.0787	1.9910	1.4953	1.5054	1.5176
9	0.8646	2.6918	1.4554	1.5810	1.6482
10	0.9644	2.0851	1.2547	1.5507	1.4625
Mean	0.9432	2.3460	1.3760	1.5557	1.5549

Note: [a] Selected item for 4-item test length condition

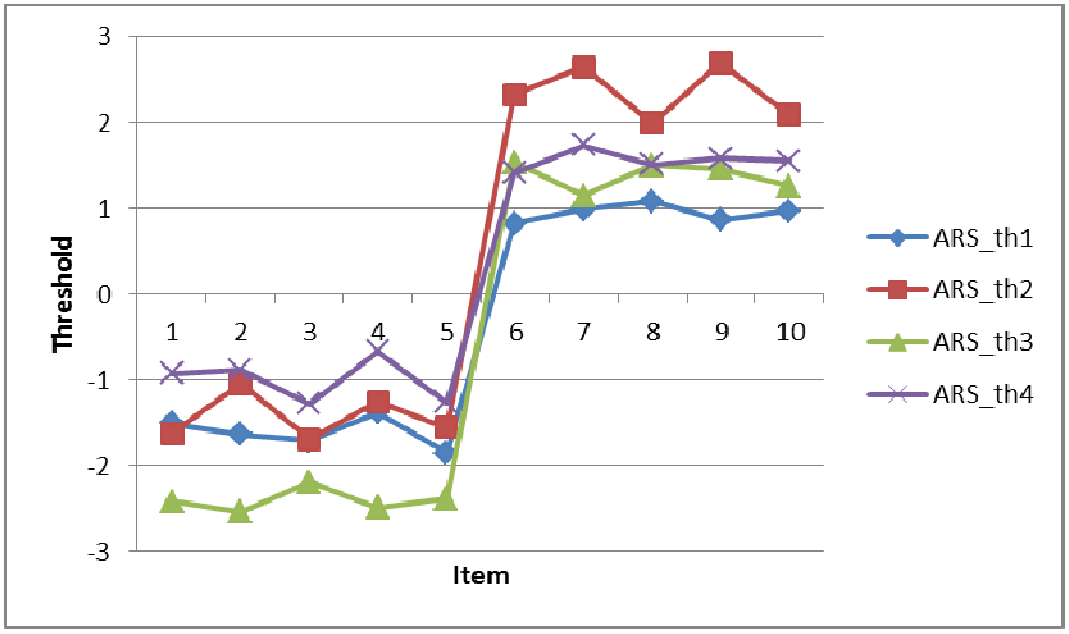

Figure 13. Thresholds plot for 10 items for the ARS class

(Items 1 to 5 are positively stated, Items 6 to 10 are negatively stated)

3.3.2 Item responses generation.

To generate item responses, person trait parameters θ_n were randomly drawn for each replication from a standard normal distribution $N \sim (0,1)$. The true θ_n and population generating threshold parameters determined for each response style were substituted in the MPCM formula. Five category probabilities (ϕ_{ik}) were computed for each respondent as demonstrated in Equation 5. These obtained category probabilities became the success probability of a multinomial distribution. Assuming that one experiment was performed that yielded $k = 5$ possible outcomes with probabilities ϕ_{i1} …, ϕ_{ik}, if the kth outcome was obtained, the kth entry of the multinomial random vector took on a value of 1, while all other entries took on values of 0. The value 1 was scored as k-1, and finally category scores from 0 to 4 were assigned. The item responses data used in the simulation was generated with R 2.14.1 (R Development Core Team, 2011).

The following Figures 14 and 15 present the conditional frequency distributions of category responses obtained for a single simulated data set. In the plots, the data set is divided into four groups according to the respondents' θ level, i.e., below 25th percentile, from 25th to 50th, from 50th to 75th, and above 75th percentile. Within each group, the frequency of category responses was counted. Figure 14 is based on an item with lower item location whereas Figure 15 is based on an item with higher item location. It is clearly seen from Figures 14 and 15 that the category response probabilities are jointly influenced by the respondent's θ level and

a response style. For the ORS class with no response style bias involved, high response category frequencies gradually shift from the lower categories to higher categories as the percentile becomes higher. This pattern of category probability shift conditional on θ level is commonly observed across all response-style classes. If ERS, MRS, or ARS is involved, however, particular response categories tend to produce the largest frequency within across all levels of θ while the gradual shift of the category probability conditional on θ levels remains.

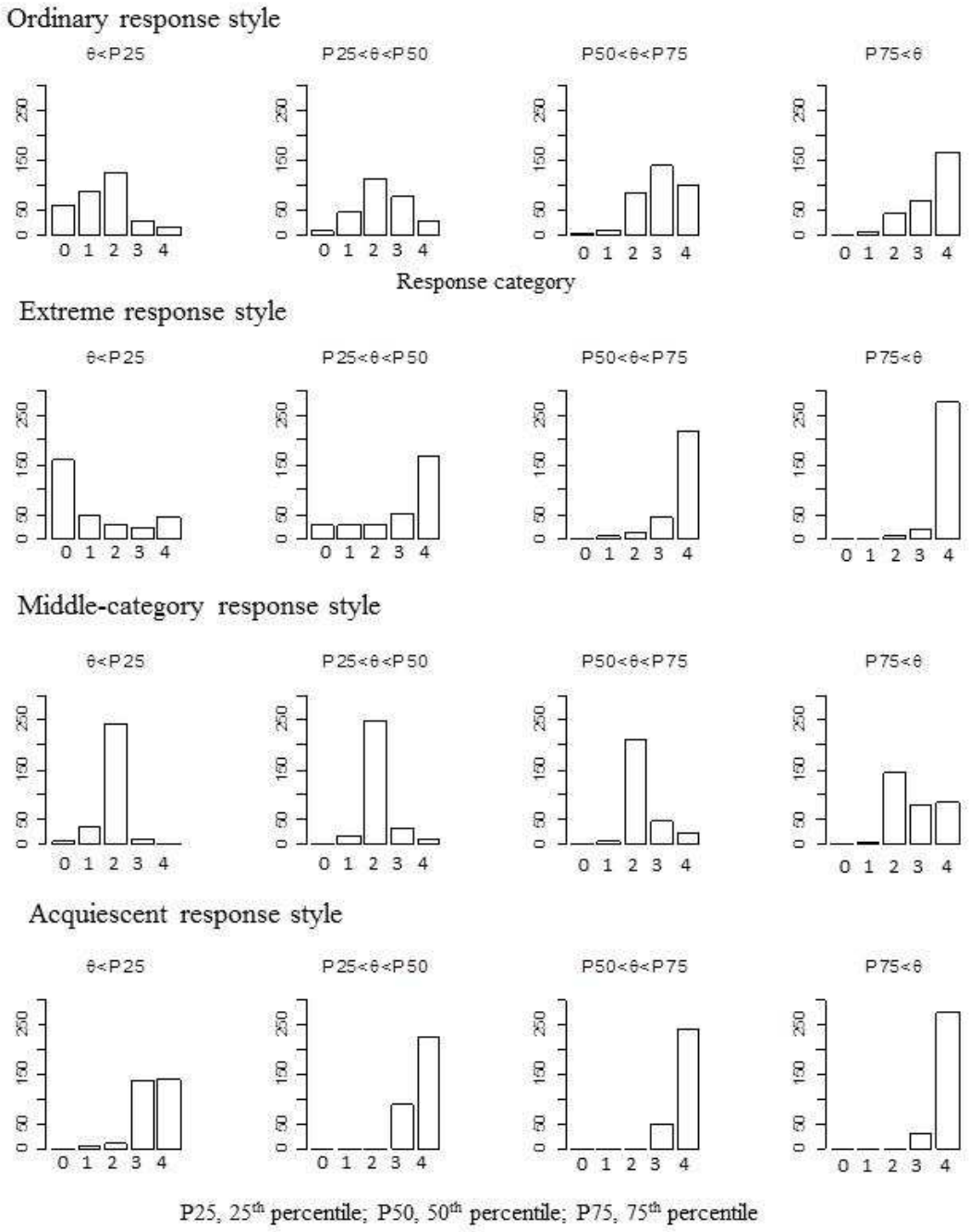

Figure 14. Conditional frequency distributions of category responses for an item with lower item location

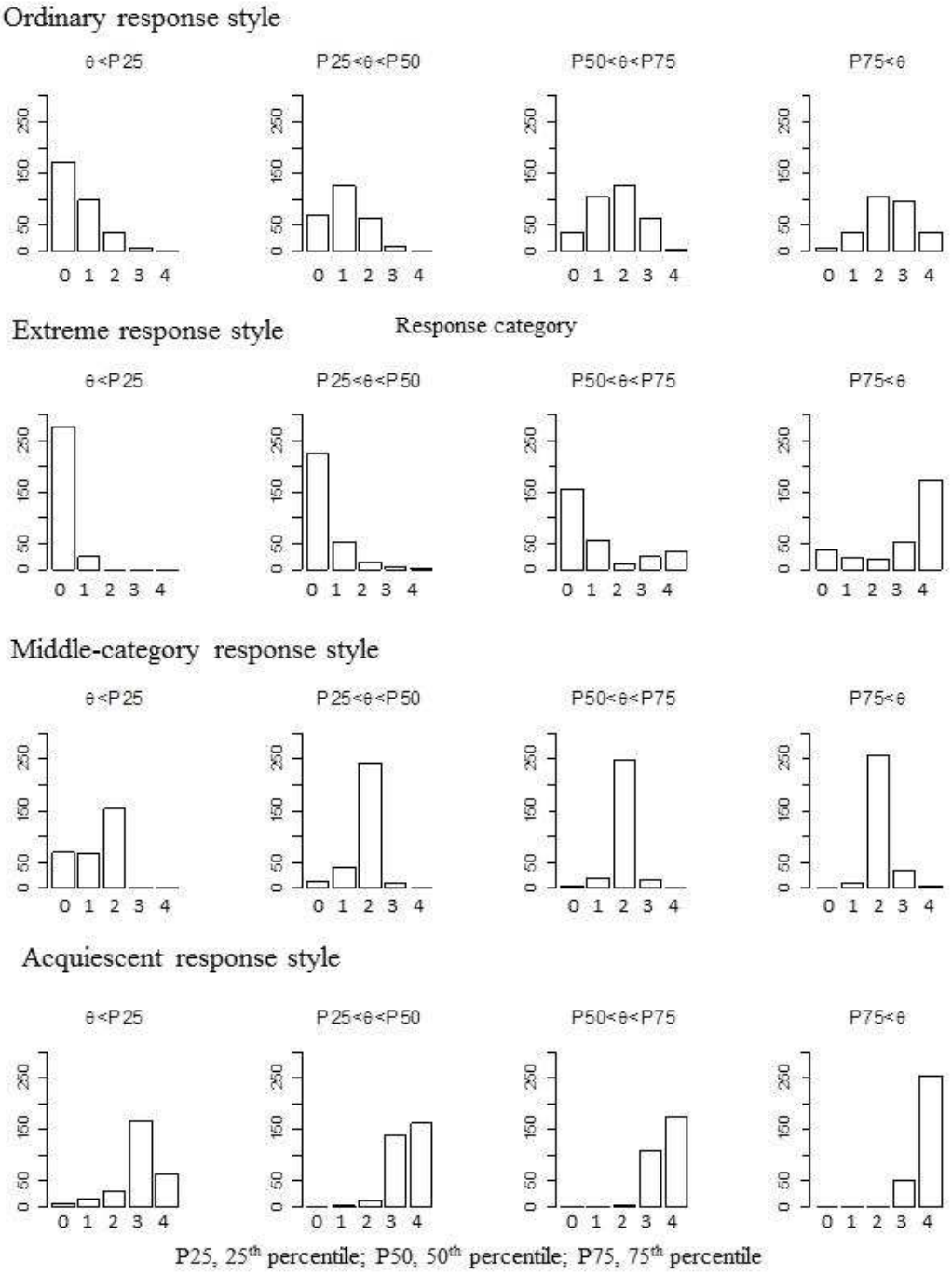

Figure 15. Conditional frequency distributions of category responses for an item with higher item location

3.4 Analysis and Evaluation Criteria

The simulated data sets that represent different mixtures of response-style respondents were estimated with the MCPM using *mdltm* software. *mdltm* allows the analyses with a wide range of latent variable models such as uni-dimensional and multi-dimensional IRT models, latent class models, mixture IRT models and diagnostic models (e.g., von Davier, 2005b). It implements the EM algorithm (Dempster, Laird, & Rubin, 1977) to obtain marginal maximum likelihood estimates of parameters. The parameter estimates provided by *mdltm* were collated and the evaluation criteria were calculated using R 2.14.1.

3.4.1 Fitting competing models

Assuming that the true model was known as the MPCM but the number of latent classes in population was unknown, the current study fit simulated data with three MPCMs with increasing numbers of latent classes. For 2-class generated data sets, 1-, 2-, and 3-class MPCM were fit to the data. For 3-class generated data sets, 2-, 3-, and 4-class MPCM were fit. Finally, for 4-class generated data sets, 3-, 4-, and 5-class MPCM were fit. These three competing estimation models: *i*) under-fitting model, which had one class less than the data generation model, *ii*) correct-fitting model, which had the same number of classes as the data generation model, and *iii*) over-fitting model, which had one class more than the data generation model, were compared with respective to their information criterion statistics, AIC, BIC, and CAIC.

3.4.2 Convergence check

To ensure that the results of each simulation analysis were grounded only on well-estimated solutions, convergence checks were conducted for each of the three competing solutions for each simulated data set. If non-convergence occurred for the correct-fitting model, all three competing solutions from that replication were discarded. To make up for the simulation data sets that were discarded as a result of non-convergent solutions, additional data set were generated. This allowed for a total of one hundred converged replication results for each simulation condition.

3.4.3 Model selection

To assess the relative effectiveness of the performance of the information criterion statistics, AIC, BIC, and CAIC in identifying the correct number of latent classes in the MPCM, the index values were obtained for each of the three estimation models. One among the three estimation models that provided the smallest index value was selected as being associated with the best-fitting model. For each index, the proportions of replications in which the true model was identified as the best-fitting model were computed. In addition, the proportions of under-identification and over-identification of latent classes were also examined. The results of the three indices were compared to find their relative effectiveness in identifying the correct number of latent classes under the various simulation conditions.

3.4.4 Problem of label switching

Label switching refers to the arbitrary mismatch between generated class membership and estimated class membership in a simulation study of mixture

modeling. In the current study, for a mixture data of ORS and ERS, for example, there are two possible ways that the estimated latent classes are labeled: ORS for the first estimated class and ERS for the second estimated class or conversely, ERS for the first and ORS for the second estimated class. In a general formulation, there are up to $C!$ ($C \times C\text{-}1 \times \ldots \times 2 \times 1$, where C is the number of latent classes) possible permutations of latent class membership assignments. Only one of the possible permutations is the correct match and others indicate the occurrence of various patterns of label switching.

In order to obtain correct measures for parameter recovery evaluation, switched labels must be detected and mismatched class membership must be corrected before aggregating estimates across multiple replications. In a simulation study where a large number of replication results need to be aggregated, it is practically impossible to manually inspect individual output for each data set to identify the occurrence of label switching. The process of correcting latent class labels needs to be automatized in the course of analysis.

In the current study, a post-hoc technique was devised by the author to detect and correct switched latent class membership based on the information from the threshold estimates. This algorithm takes advantage of the distinctive order of thresholds that characterize each response-style class. As presented in Table 3, the mean values of population generating thresholds across all items for each response-style class show particular orders in terms of their magnitude. If the means of estimated thresholds (δ_1, δ_2, δ_3, and δ_4) for an estimated class satisfies the order of $\{\delta_1 < \delta_2 < \delta_3 < \delta_4\}$, that class is identified as an ORS class. If the set of means

satisfies the condition of { $\delta_1 > 0$ and $\delta_4 < 0$} in a class, that class is identified as an ERS class. For MRS and ARS class, the conditions of { $\delta_1 < 0$ and $\delta_2 < 0$ and $\delta_1 > \delta_2$ } and { $\delta_1 < 0$ and $\delta_2 > 0$ and $\delta_3 < 0$ and $\delta_4 > 0$} are applied, respectively.

Table 8. *Means of Generated Threshold Parameters for Each Response-style Class*

Class	Threshold1	Threshold2	Threshold3	Threshold4
ORS	-1.5037	-0.6644	0.6644	1.5037
ERS	0.8243	0.4845	-0.4845	-0.8243
MRS	-1.4938	-2.6157	2.6157	1.4938
ARS	-0.3363	0.4550	-0.5150	0.2763

In addition to employing this algorithm using thresholds characteristics, a different algorithm that is based on the information from respondent classification developed by Tueller, Drotar, and Lubke (2011) was implemented. The results of employing these two different algorithms were compared.

3.4.5 Classification accuracy

The classification accuracy was evaluated for the correct-fitting model solutions. The classification accuracy was computed as the proportion of respondents who were assigned to their generated class membership based on the magnitudes of the posterior probabilities for the various class memberships. Not only the correct classification rate but also the nature of misclassifications was closely examined. Misclassified individuals were cross-tabulated for all possible combinations of misclassification to explore whether there was any particular misclassification pattern.

3.4.6 Threshold parameter recovery

The accuracy of threshold parameter recovery was evaluated in terms of Pearson *r*, root mean square error (RMSE), and standard error of estimates (SE). Correlation and RMSE provide the measures of overall accuracy of parameter estimates. The closer the generated and estimated parameters are to each other, the higher positive correlation and the smaller RMSE are expected. For threshold parameter recovery, SE was computed based on the standard deviation of sample estimates from their average value. This indicates the stability of parameter estimates. A great fluctuation of estimated parameter values from replication to replication increases the SE. For item parameter recovery, the four evaluation criteria were calculated for each of four thresholds. They are computed as follows:

$$Corr_{\hat{\delta}_k \delta_k} = \frac{1}{W} \sum_{w=1}^{W} \sum_{i=1}^{I} r_{\hat{\delta}_{wik} \delta_{ik}} , \; k = 1, .., 4.$$

$$RMSE(\hat{\delta}_k) = \sqrt{\frac{\sum_{w=1}^{W} \sum_{i=1}^{I} (\hat{\delta}_{wik} - \delta_{ik})^2}{I \times W}} ,$$

$$SE(\hat{\delta}_k) = \sqrt{\frac{1}{I \times W} \sum_{w=1}^{W} \sum_{i=1}^{I} \left(\hat{\delta}_{wik} - \frac{\sum_{w=1}^{W} \sum_{i=1}^{I} \hat{\delta}_{wik}}{I \times W} \right)^2} .$$

where $r_{\hat{\delta}_k \delta_k}$ is the Pearson *r* between *k*th true threshold (δ_{ik})and its estimate ($\hat{\delta}_{ik}$). *i* indicates *i*th item (*i* = 1,…*I*), *w* is *w*th replication (*w* = 1,…*W*).

The mean bias, which is the measure of discrepancy between generated and estimated parameters, was not considered as an evaluation criterion for threshold

parameter recovery in the current study. During the parameter estimation in the current study, the item constrain method was used for the purpose of model identification. As introduced briefly in Section 2.5.1, either item parameter or person trait parameter needs to be constrained to solve the indeterminacy problem in IRT models. The software *mdltm* allows user to choose either of the two constrain methods. If item constraints are used, the sum of the estimated thresholds will be zero in each latent class while if person constraints are used, the sum of the estimated θs will be zero in each latent class. The current study used the former method and, consequently, the mean bias across thresholds and items turned out to be zero for all simulation conditions, which was illegitimate to be used as an evaluation criterion as was originally proposed.

3.4.7 Person trait parameter recovery

The accuracy of person trait parameter (θ) recovery was evaluated in terms of Pearson *r*, bias, and root mean square error (RMSE). For theta recovery, the evaluation criteria were calculated for each class as follows:

$$Corr_{\hat{\theta}\theta} = \frac{1}{W}\sum_{w=1}^{W}\sum_{n=1}^{N} r_{\hat{\theta}_{wn}\theta_n} ,$$

$$Bias(\hat{\theta}) = \frac{1}{N \times W}\sum_{w=1}^{W}\sum_{n=1}^{N}(\hat{\theta}_n - \theta_n) ,$$

$$RMSE(\hat{\theta}) = \sqrt{\frac{\sum_{w=1}^{W}\sum_{n=1}^{N}(\hat{\theta}_n - \theta_n)^2}{N \times W}} ,$$

where θ_n is person *n*th true trait , $\hat{\theta}_n$ is its estimate and *N* is the sample size or total number of respondents.

3.4.8 Model-based correction of score bias due to response styles

The relation between sum scores and the MPCM θ estimates was investigated. To explore how the relation differ across latent classes when two, three, or four different types of response style were mixed, plots in which the MPCM θ estimates were depicted as a function of sum scores were created.

3.4.9 Evaluation of effects of manipulated factors

One of the main interests of the current study was to investigate the influence of the four factors on the MPCM performance: *i*) type of mixture at five levels, *ii*) mixing proportions at two levels, *iii*) sample size at three levels, and *ii*) test length at three levels.

Using the evaluation criteria measures (i.e., percentages, biases, RMSEs, correlations, and SEs) as the dependent variables, several factorial ANOVAs were conducted. Four main effects of the manipulated factors and all two-way interaction effects were included in the ANOVA model. The higher order interaction effects were folded into the error term. In the current study, many cell means were unavailable because of the exclusions of the simulation conditions in which the problems of estimation and label switching occurred. Under this incomplete design where some estimated cell means were missing, the interpretation of higher order interaction effects was seen as being quite difficulty to properly interpret and quite limited and,

thus, would provide limited (possibly misleading) information about the manipulated factors in this study.

The influence of manipulated factors was determined to be statistically significant if the associated *p*-value < .05. Practical significance was measured by the effect size index, $\eta^2 = \frac{SS_{effect}}{SS_{total}}$, defined as the variance accounted for by the manipulated effect. According to Cohen (1988), η^2 of 0.06 and 0.14 represent medium and large effect sizes for factorial ANOVA analysis, respectively. In the current study, the importance of the effects of the manipulated factors was evaluated based on the combination of statistical significance and practical significance. Only those manipulated factors for which their *p*-value was smaller than 0.05 and, at the same time, η^2 was greater than 0.06 for medium effect or 0.14 for large effect was interpreted for its importance.

Chapter 4: Results

Chapter 4 presents results of the current simulation study in six sections. Before presenting the results to answer the main research questions, the first section 4.1 addresses how the current study treated problems related to the convergence of the program to provide reasonable model parameter estimates as well as issues surrounding label switching. Section 4.2 provides the results of model selection under the MPCM based on information criterion statistics. Assessment of the results of model performance in the recovery of latent class membership, item threshold parameters, and person trait parameters are provided in Section 4.3, 4.4, and 4.5, respectively. Finally, findings regarding the model-based correction of person trait estimates are discussed in Section 4.6.

4.1. Initial Treatment of Estimation Problems and Label Switching Problems

4.1.1 Non-convergence and boundary estimates

The population models used to generate item response data for this simulation study were five different MPCMs: *i*) three 2-class MPCMs representing mixtures of the ORS-ERS, ORS-MRS, and ORS-ARS, *ii*) a 3-class MPCM representing a mixture of the ORS-ERS-MRS, and *iii*) a 4-class MPCM representing a mixture of the ORS-ERS-MRS-ARS. These five data generation models were estimated under not only the same MPCM model (i.e., correct-fitting), but also an under-fitting model (i.e., estimation with the MPCM that has one class fewer than the data generation model) as

well as over-fitting (i.e., estimation with the MCPM that has one more latent classes than the population generating model).

Two situations that may indicate problems in achieving convergence of parameter estimates were checked for these three estimation solutions. The first situation could be characterized when estimation terminated without convergence. The second situation that prompted monitoring occurred when maximum likelihood estimates of item thresholds skirted the boundary of permissible parameter values. These two problems were reported separately. The software *mdltm* provides an explicit warning message that indicates the occurrence of the first of these situations. The percentage of replications in which this warning message appeared is reported in Table 9. For the second condition, threshold estimates that were more extreme than 9.0 or -9.0 were flagged and the percentage of the replications in which one or more boundary estimates were flagged is reported in parentheses in Table 9.

Correct-parameterization. Under the correct-fitting, non-convergence as well as boundary estimates did not occur across all levels of the ORS-ERS mixtures. However, for the other types of mixtures, significant numbers of boundary estimates appeared when the sample size was relatively small (N = 1200). Specifically, boundary estimates occurred for the MRS or ARS thresholds when the expected response probabilities for the corresponding response categories were essentially zero. When the sample size was N = 1200 and the mixing proportions were $\pi = 0.9$ versus $\pi = 0.1$, there were only 120 responses in the MRS or ARS class. Recall that the expected category probability for the 1st and 5th response categories for the MRS class

was set up to be approximately 6% while that for the 1st, 2nd, and 3rd categories for the ARS class was approximately 5%. That means that as small as 72 or 60 responses were assigned for those response categories. This data generation condition resulted in essentially zero expected frequencies in some randomly generated samples and may very well explain why the software converged to such extreme boundary values. It appears that the sample size of $N = 1200$ was not large enough to provide sufficient information and subsequent maximum likelihood estimates often fell at the boundary.

Under-parameterization. Under the under-fitting, neither non-convergence nor boundary estimates occurred for any of the 2-response-style mixtures as well as for the 3-response-style mixtures. However, the 4-response-style mixture with 4-items and a sample size of $N = 6000$ produced a non-convergence rate of 0.49 when it was fit with an under-fitting model.

Over-parameterization. Expectedly, under the over-fitting, estimation problems increased and almost all simulation conditions produced boundary threshold estimates. The average rate of the occurrence of boundary estimates problems was 0.46. The higher rate of boundary estimates were observed when *i*) the data generation model had three or four latent classes, *ii*) the sample size was $N = 1200$, or *iii*) the mixing proportions were unequal. These findings may contain real implications for practitioners using these methods in real data analytic situations. That is, the occurrence of infeasible extreme threshold values may be an indication of over-parameterization (estimating a model with too many latent classes) or an insufficient sample size to estimate parameters of a given data set, or a combination of the two.

Table 9. *Percentages of the Occurrence of Non-convergence and Boundary Threshold Estimates*

	Type of Mixture		ORS ERS			ORS MRS			ORS ARS		
	Estimation model		1class	2class	3class	1class	2class	3class	1class	2class	3class
Mixing Proportions	Item	Sample									
50:50	4	1200	0 (0)†	0 (0)	0 (6)	0 (0)	0 (9)	0 (67)	0 (0)	0 (0)	1 (6)
		3000	0 (0)	0 (0)	0 (0)	0 (0)	0 (3)	0 (10)	0 (0)	0 (0)	0 (1)
		6000	0 (0)	0 (0)	0 (1)	0 (0)	0 (0)	0 (16)	0 (0)	9 (0)	8 (1)
	10	1200	0 (0)	0 (0)	5 (51)	0 (0)	0 (0)	0 (96)	0 (0)	0 (0)	0 (56)
		3000	0 (0)	0 (0)	1 (35)	0 (0)	0 (0)	0 (82)	0 (0)	0 (0)	0 (92)
		6000	0 (0)	0 (0)	2 (27)	0 (0)	0 (0)	0 (84)	0 (0)	0 (0)	0 (87)
	20	1200	0 (0)	0 (0)	2 (44)	0 (0)	0 (0)	2 (99)	0 (0)	0 (0)	2 (33)
		3000	0 (0)	0 (0)	2 (41)	0 (0)	0 (0)	1 (87)	0 (0)	0 (0)	0 (26)
		6000	0 (0)	0 (0)	2 (32)	0 (0)	0 (0)	16(50)	0 (0)	0 (0)	0 (60)
90:10	4	1200	0 (0)	0 (0)	0 (19)	0 (0)	3 (1)	7 (24)	0 (0)	3(25)	13(42)
		3000	0 (0)	0 (0)	0 (6)	0 (0)	1 (3)	20(13)	0 (0)	0 (2)	15 (8)
		6000	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (21)	0 (0)	8 (0)	7 (5)
	10	1200	0 (0)	0 (0)	1 (48)	0 (0)	0(65)‡	0 (79)	0 (0)	1(48)‡	0 (15)
		3000	0 (0)	0 (0)	0 (34)	0 (0)	0 (16)	0 (54)	0 (0)	0 (6)	0 (16)
		6000	0 (0)	0 (0)	0 (31)	0 (0)	0 (0)	1 (17)	0 (0)	0 (0)	0 (0)
	20	1200	0 (0)	0 (0)	0 (25)	0 (0)	0(75)‡	0 (93)	0 (0)	0(58)‡	0 (64)
		3000	0 (0)	0 (0)	1 (22)	0 (0)	0 (7)	0 (37)	0 (0)	0 (0)	0 (23)
		6000	0 (0)	0 (0)	3 (77)	0 (0)	0 (0)	7 (45)	0 (0)	0 (0)	3 (4)

Table 9_continued

	Type of Mixture		ORS ERS MRS			ORS ERS MRS ARS		
	Estimation model		2class	3class	4class	3class	4class	5class
Mixing Proportions	Item	Sample						
50:50	4	1200	0 (0)	0 (8)	0 (41)	0 (5)	0 (48) ‡	0 (67)
		3000	0 (0)	0 (5)	2 (21)	0 (0)	0 (12)	0 (41)
		6000	0 (0)	1 (1)	0 (12)	0 (0)	0 (0)	0 (21)
	10	1200	0 (0)	0 (5)	4 (79)	0 (13)	0 (27)	3 (89)
		3000	0 (0)	0 (0)	0 (32)	0 (0)	0 (1)	0 (44)
		6000	0 (0)	0 (0)	2 (51)	0 (0)	0 (0)	2 (44)
	20	1200	0 (0)	0 (1)	0 (29)	0 (6)	0 (10)	1 (74)
		3000	0 (0)	0 (0)	0 (82)	0 (0)	0 (0)	0 (51)
		6000	0 (0)	0 (0)	0 (92)	0 (0)	0 (4)	0 (77)
90:10	4	1200	0 (0)	0 (15)	0 (61)	0 (29)	0 (46) ‡	0 (71)
		3000	0 (0)	0 (7)	0 (42)	2 (11)	0 (11)	1 (30)
		6000	0 (0)	2 (3)	3 (14)	49 (0)	5 (3)	5 (66)
	10	1200	0 (0)	0 (69) ‡	1 (93)	1 (51)	0 (96) ‡	2 (99)
		3000	0 (0)	0 (19)	1 (90)	0 (18)	0 (32)	0 (77)
		6000	0 (0)	0 (0)	0 (44)	0 (1)	0 (3)	0 (50)
	20	1200	0 (0)	0 (79) ‡	1 (98)	0 (78)	0 (93) ‡	0 (99)
		3000	0 (0)	0 (8)	0 (79)	0 (8)	0 (19)	0 (83)
		6000	0 (0)	1 (9)	0 (32)	0 (0)	0 (9)	0 (43)

Note. † Percentage of the occurrences of boundary estimates is presented in parentheses
‡ Excluded from simulation summary due to high occurrence rate of boundary estimates

Exclusion of estimation solutions with estimation problems. Ten conditions out of ninety in the current simulation design presented boundary thresholds estimates

in more than approximately half of the replications when the generated data sets were parameterized with the correct model. These problematic conditions with a high level of estimation problems were excluded from the simulation summary and are listed in Table 10. For other simulation conditions with a moderate level of estimation problems, (i.e., either non-convergence or estimates at boundary values between 1 % and 30 %), the problematic results were discarded and new replications that did not present these problems replaced the discarded replications.

Table 10.*Specifications of Simulation Conditions Excluded from Simulation Summary Due to Estimation Problems*

Type of mixture	Mixing proportions	Number of items	Sample size	Occurrence rate of boundary estimates (%)
ORS-MRS	0.9 : 0.1	10	1200	65
ORS-MRS	0.9 : 0.1	20	1200	75
ORS-ARS	0.9 : 0.1	10	1200	48
ORS-ARS	0.9 : 0.1	20	1200	58
ORS-ERS-MRS	0.9 : 0.1	10	1200	69
ORS-ERS-MRS	0.9 : 0.1	20	1200	79
ORS-ERS-MRS-ARS	0.5 : 0.5	4	1200	48
ORS-ERS-MRS-ARS	0.9 : 0.1	4	1200	46
ORS-ERS-MRS-ARS	0.9 : 0.1	10	1200	96
ORS-ERS-MRS-ARS	0.9 : 0.1	20	1200	93

In general, parameter estimation in the MPCM achieved fairly high convergence rates across various simulation conditions. However, the sample size of

N = 1200 appeared to be insufficient to provide well-estimated parameters especially when a small proportion of the respondents in a sample presented ERS, MRS or ARS.

4.1.2 Label switching problems

As is usual in any mixture modeling simulation study, label switching occurred. In the current study, label switching was detected using two different algorithmic approaches. The first algorithm was based on the information from the threshold estimates developed by the author while the second algorithm was based on the information from respondent classifications developed by Tueller, Drotar, and Lubke (2011).

Label switching correction algorithm based on thresholds information. As explained in Section 3.4.4, to automate the correction of switched class membership, an algorithm was developed that exploited the distinctive order of the thresholds that characterized each response style. To demonstrate how the algorithm works, an illustrative example in which threshold estimates from the 4-response-style mixture with 10-itmes and a sample size of N = 6000 was used in the following.

First, the mean thresholds for ten items were calculated for each replication. Instead of using individual item threshold estimates, the mean values over all items were used because mean values were more consistent from replication to replication than individual item threshold estimates. The following matrix shows the mean thresholds for the first five replications.

	Class 1				Class 2				Class 3				Class 4			
	δ_1	δ_2	δ_3	δ_4	δ_1	δ_2	δ_3	δ_4	δ_1	δ_2	δ_3	δ_4	δ_1	δ_2	δ_3	δ_4
Rep1	0.83	0.45	-0.47	-0.81	-1.46	-2.69	2.63	1.52	-0.25	0.39	-0.46	0.32	-1.47	-0.73	0.71	1.50
Rep2	0.86	0.39	-0.41	-0.85	-1.40	-2.71	2.66	1.46	-0.39	0.44	-0.42	0.38	-1.48	-0.70	0.78	1.41
Rep3	0.83	0.54	-0.52	-0.86	-1.45	-0.65	0.65	1.44	-0.19	0.45	-0.54	0.29	-1.44	-2.58	2.63	1.40
Rep4	-1.44	-0.68	0.62	1.50	0.83	0.51	-0.41	-0.93	-0.27	0.45	-0.48	0.30	-1.41	-2.59	2.55	1.45
Rep5	-1.48	-0.61	0.65	1.44	0.84	0.42	-0.48	-0.78	-0.38	0.70	-0.54	0.21	-1.47	-2.55	2.61	1.42

The first set of four thresholds from Replication 1 satisfies the condition of $\{\delta_1 > 0$ and $\delta_4 < 0\}$, which characterizes the ERS class. Note that any of the remaining sets do not meet this condition. The second set satisfies the condition of $\{\delta_1 < 0$ and $\delta_2 < 0$ and $\delta_1 > \delta_2\}$, which characterizes the MRS class. The third set satisfies the condition of $\{\delta_1 < 0$ and $\delta_2 > 0$ and $\delta_3 < 0$ and $\delta_4 > 0\}$, which characterizes the ARS class and finally, the fourth set satisfies the condition of $\{\delta_1 < \delta_2 < \delta_3 < \delta_4\}$, which characterizes the ORS class. Originally, the generated latent class labels were ORS, ERS, MRS, and ARS for class 1, class 2, class 3, and class 4, respectively. Thus, the estimated class labels for Replication 1, i.e., ERS, MRS, ARS, and ORS were identified as switched labels.

There are 4! = 24 possible ways that four class labels can be switched. Each replication was checked for all twenty-four possible mismatches and the proper label was labeled for each latent class. The switched class labels that were identified for the five replications in the illustration are as follows:

	Class 1	Class 2	Class 3	Class 4
Rep1	ERS	MRS	ARS	ORS
Rep2	ERS	MRS	ARS	ORS
Rep3	ERS	ORS	ARS	MRS
Rep4	ORS	ERS	ARS	MRS
Rep5	ORS	ERS	ARS	MRS

Based on these identified class labels, the thresholds matrix was reorganized as presented below. Likewise, matrices of class membership assignment as well as person trait estimates (not presented in this document) were also rearranged for use in the subsequent analyses in the study.

	ORS				ERS				MRS				ARS			
	δ_1	δ_2	δ_3	δ_4	δ_1	δ_2	δ_3	δ_4	δ_1	δ_2	δ_3	δ_4	δ_1	δ_2	δ_3	δ_4
Rep1	-1.47	-0.73	0.71	1.50	0.83	0.45	-0.47	-0.81	-1.46	-2.69	2.63	1.52	-0.25	0.39	-0.46	0.32
Rep2	-1.48	-0.70	0.78	1.41	0.86	0.39	-0.41	-0.85	-1.40	-2.71	2.66	1.46	-0.39	0.44	-0.42	0.38
Rep3	-1.45	-0.65	0.65	1.44	0.83	0.54	-0.52	-0.86	-1.44	-2.58	2.63	1.40	-0.19	0.45	-0.54	0.29
Rep4	-1.44	-0.68	0.62	1.50	0.83	0.51	-0.41	-0.93	-1.41	-2.59	2.55	1.45	-0.27	0.45	-0.48	0.30
Rep5	-1.48	-0.61	0.65	1.44	0.84	0.42	-0.48	-0.78	-1.47	-2.55	2.61	1.42	-0.38	0.70	-0.54	0.21

This label switching correction algorithm successfully identified switched labels when the quality of thresholds recovery was fairly good. However, this algorithm seemed to be rather strict, so that some switched labels were not automatically detected although they were discernible if inspected individually by looking at the whole picture of all items' threshold estimates in all classes.

Label switching correction algorithm based on classification information. Tueller, Drotar, and Lubke (2011) developed a switched label detection algorithm that

utilized respondent classification results after estimation was completed. Their algorithm assumed that the frequency of correctly classified cases must be greater than the frequencies of misclassified cases. Therefore, each column of the class assignment matrix must have one column maxima. To help in understanding the algorithm developed by Tueller and his colleagues, three exemplar matrices of the frequencies of class membership assignment are presented below. The columns of the matrices represent true class membership and the rows represent assigned class membership. The first matrix shows a case where labels were not switched. The second matrix shows a case where the labels were switched and can be corrected. The third matrix shows a case where the labels were switched but cannot be corrected via their algorithm because its column has more than one column maxima.

Labels not switched			
	True 1	True 2	True 3
Assign1	96	6	2
Assign2	1	91	5
Assign3	3	7	89

Labels Switched			
	True 1	True 2	True 3
Assign1	9	60	9
Assign2	80	1	14
Assign3	11	39	77

Cannot be corrected			
	True 1	True 2	True 3
Assign1	38	33	36
Assign2	38	31	35
Assign3	24	36	34

Tueller et al. (2011) pointed out that reliable use of this algorithm requires reasonably high classification accuracy. They provided guidelines to prevent spurious correction by setting up a level of class assignment criterion that allows the researcher to decide how much more respondents are required to be correctly assigned than expected by chance.

Although drastic improvement was not anticipated from an additional application of the Tueller's algorithm, it seemed to be a potential alternative to

maximize the efficiency of automatic procedure to resolve the label switching dilemma. Since Tueller's algorithm uses different sources of information, some replications for which the algorithm based on thresholds was not able to detect switched labels may find a solution via Tueller's algorithm.

Results of detecting and correcting switched labels. When the two algorithms were both able to solve switched labels, they yielded identical results. Interestingly, switched labels in some replications were detected by only one of the algorithms, but not both. The two algorithms, therefore, were incorporated in the course of the analysis and, as a result, switched labels in more replications were solvable in an automated manner than when either of the two algorithms was used alone.

There were thirteen simulation conditions in which label switching could not be detected for some of the replications despite applying the two algorithms as well as a more in-depth manual inspection carried out for individual outputs. The following illustration presents a case of switched labels, which was not able to be solved by any of the three methods: *i*) estimated thresholds did not hold the particular conditions of the order of thresholds, *ii*) the class assignment matrix presented more than one column maxima, and *iii*) the manual inspection of the thresholds of all four items was not informative to separate three classes.

Class 1				Class 2				Class 3			
δ_1	δ_2	δ_3	δ_4	δ_1	δ_2	δ_3	δ_4	δ_1	δ_2	δ_3	δ_4
-1.87	-0.97	1.44	1.41	-0.19	-0.82	0.17	0.84	-1.23	-0.53	0.72	1.04

Labels not switched

	True 1	True 2	True 3
Assign1	381	0	95
Assign2	347	56	18
Assign3	232	64	7

	Class 1				Class 2				Class 3			
	δ_1	δ_2	δ_3	δ_4	δ_1	δ_2	δ_3	δ_4	δ_1	δ_2	δ_3	δ_4
Item1	-2.28	-0.95	1.37	1.39	-0.77	-0.28	-0.21	0.89	-1.39	-0.47	-0.33	-0.57
Item2	-2.48	-0.91	1.57	1.74	-0.10	-0.52	0.04	0.95	-1.17	-0.27	2.22	3.26
Item3	-1.64	-1.05	0.81	1.12	-0.61	-0.85	0.07	0.53	-1.43	-0.90	-0.34	-0.34
Item4	-1.08	-0.99	1.99	1.38	0.71	-1.62	0.80	0.98	-0.94	-0.46	1.32	1.82

As implied in the above example, the fact that there were unsolvable switched labels should not be regarded as an indication of any flaw or ineffectiveness of the algorithms. Instead, it seemed to be a reflection of the nature of the generated data sets and/or quality of the estimation. These thirteen conditions were also the ones for which the model selection based on the information criteria failed to identify the correct data generation model (The related results of model selection are presented subsequently in Section 4. 2). The specifications of the simulation conditions in which unsolvable switched labels were observed and the occurrence rates are summarized in Table 11. For these thirteen conditions, unsolvable replications were discarded and only the remaining solvable solutions were used to compute the evaluation criteria.

Table 11. *Specifications of Simulation Conditions in which Switched Labels are unsolvable*

Type of mixture	Mixing proportions	Number of items	Sample size	Percentage of the occurrences of unsolvable switched labels
ORS-ERS	0.9 : 0.1	4	1200	41
ORS-MRS	0.9 : 0.1	4	1200	41
ORS-MRS	0.9 : 0.1	4	3000	40
ORS-ERS-MRS	0.5 : 0.5	4	1200	14
ORS-ERS-MRS	0.9 : 0.1	4	1200	45
ORS-ERS-MRS	0.9 : 0.1	4	3000	45
ORS-ERS-MRS	0.9 : 0.1	4	6000	42
ORS-ERS-MRS-ARS	0.5 : 0.5	4	3000	46
ORS-ERS-MRS-ARS	0.5 : 0.5	4	6000	56
ORS-ERS-MRS-ARS	0.5 : 0.5	10	1200	43
ORS-ERS-MRS-ARS	0.9 : 0.1	4	3000	67
ORS-ERS-MRS-ARS	0.9 : 0.1	4	6000	63
ORS-ERS-MRS-ARS	0.9 : 0.1	10	3000	54

4.2. Model selection

Once the replications that did not converge had been replaced, the AIC, BIC, and CAIC values were collated from each of the three competing estimation solutions for each replication. The percentage of the replications in which one of the competing models being identified as the best-fitting model by each information criterion index was recorded. The following Tables 12-16 presented the results.

Generally, the BIC and CAIC performed nearly equally well with a slightly higher accuracy rate for the BIC across many conditions. On the other hand, the AIC resulted in over-identification problem (choosing a model with more classes) across all

of the simulation conditions. In the current study, the BIC was found to be the most effective information criterion statistic to use for the identification of the correct number of latent classes of the MPCM. The model selection results for each type of mixture are presented in the following sections in detail.

Model selection under the ORS-ERS mixtures. Table 12 presents the selection results for the ORS-ERS mixtures. The ORS-ERS mixtures were well recognized as 2-response-style mixtures based on the BIC and CAIC across all simulation conditions. An exception was the condition of the 4-items and a sample of $N = 1200$ with unequal mixing proportions, which resulted in 97% of under-identification problem (choosing a model with fewer classes). Note that this condition presented 41% of unsolvable label switching problem as well. Table 13 presents the results of the model selection under the ORS-MRS mixtures. Generally, the ORS-MRS mixtures were not identified as correctly as other types of 2-response-style mixtures.

As introduced in Section 2.6.2, “degree of heterogeneity” is related to the difficulty of detecting component distributions in the MCPM. It was predicted that when the item parameters and threshold distances differ strongly, “unmix” the mixture distribution will be easier in Rost (1991). Looking back at the category characteristic curves (CCCs) illustrated in Figure 4 - Figure 8, the differences between the ORS and MRS thresholds may be seen as being less distinctive than those between the ORS and ERS thresholds as well as the ORS and ARS thresholds. Consequently, the ORS-MRS mixtures were relatively more difficult to be identified as a mixture distribution.

Table 12.*Model Selection under the ORS-ERS Mixtures*

Information Criterion				AIC			BIC			CAIC		
Number of classes of the estimation model				1	2	3	1	2	3	1	2	3
Type of Mixture	Mixing Proportions	Item	Sample									
ORS	50:50	4	1200	0	89	11	0	100	0	0	100	0
			3000	0	97	3	0	100	0	0	100	0
			6000	0	92	8	0	100	0	0	100	0
		10	1200	0	74	26	0	100	0	0	100	0
			3000	0	85	15	0	100	0	0	100	0
			6000	0	72	28	0	100	0	0	100	0
		20	1200	0	78	22	0	100	0	0	100	0
			3000	0	55	45	0	100	0	0	100	0
			6000	0	52	48	0	100	0	0	100	0
ERS	90:10	4	1200	0	92	8	97	3	0	100	0	0
			3000	0	87	13	1	99	0	6	94	0
			6000	0	89	11	0	100	0	0	100	0
		10	1200	0	71	29	0	100	0	0	100	0
			3000	0	70	30	0	100	0	0	100	0
			6000	0	38	62	0	100	0	0	100	0
		20	1200	0	57	43	0	100	0	0	100	0
			3000	0	52	48	0	100	0	0	100	0
			6000	0	96	4	0	100	0	0	100	0

When the condition was the 4-items and a sample size of N = 1200 with equal mixing proportions, only 48% of the ORS-MRS data sets were correctly identified. When the mixing proportions were unequal, the correct model selection rates based on the BIC or CAIC became even lower and an increase in the sample size from N = 1200 to N = 6000 did not improve the rates significantly. Despite the increase in the number of items up to ten, the correct selection rates was still very low (5%) with a sample size of N = 1200.

Table 13.*Model Selection under the ORS-MRS Mixtures*

Information Criterion				AIC			BIC			CAIC		
Number of classes of the estimation model				1	2	3	1	2	3	1	2	3
Type of Mixture	Mixing Proportions	Item	Sample									
ORS	50:50	4	1200	0	83	17	52	48	0	77	23	0
			3000	0	34	66	0	100	0	0	100	0
			6000	0	11	89	0	100	0	0	100	0
		10	1200	0	85	15	0	100	0	0	100	0
			3000	0	86	14	0	100	0	0	100	0
			6000	0	88	12	0	100	0	0	100	0
		20	1200	0	59	41	0	100	0	0	100	0
			3000	0	78	22	0	100	0	0	100	0
			6000	0	81	19	0	100	0	0	100	0
MRS	90:10	4	1200	34	58	8	100	0	0	100	0	0
			3000	1	94	5	100	0	0	100	0	0
			6000	3	64	33	90	10	0	92	8	0
		10	1200	0	65	35	95	5	0	100	0	0
			3000	0	40	60	9	91	0	12	88	0
			6000	0	5	95	0	100	0	0	100	0
		20	1200	0	38	62	0	100	0	5	95	0
			3000	0	8	92	0	100	0	0	100	0
			6000	0	52	48	0	100	0	0	100	0

Table 14 presents the results of model selection under the ORS-ARS mixtures. All levels of ORS-ARS data sets were identified correctly as a 2-class mixture based on the BIC and the CAIC. It appeared that the highly pronounced thresholds characteristics in the ARS class i.e., all thresholds are positive for half of items and all thresholds are negative for the other half of items, made the identification of this class easier than the identification of either the ERS or MRS class.

Table 14. *Model Selection under the ORS-ARS Mixtures*

Information Criterion				AIC			BIC			CAIC		
Number of classes of the estimation model				1	2	3	1	2	3	1	2	3
Type of Mixture	Mixing Proportions	Item	Sample									
ORS ARS	50:50	4	1200	0	79	21	0	100	0	0	100	0
			3000	0	56	44	0	100	0	0	100	0
			6000	0	24	76	0	100	0	0	100	0
		10	1200	0	50	50	0	100	0	0	100	0
			3000	0	27	73	0	100	0	0	100	0
			6000	0	11	89	0	100	0	0	100	0
		20	1200	0	46	54	0	100	0	0	100	0
			3000	0	10	90	0	100	0	0	100	0
			6000	0	0	100	0	100	0	0	100	0
	90:10	4	1200	0	86	14	0	100	0	0	100	0
			3000	0	86	14	0	100	0	0	100	0
			6000	0	77	23	0	100	0	0	100	0
		10	1200	0	43	57	0	100	0	0	100	0
			3000	0	20	80	0	100	0	0	100	0
			6000	0	1	99	0	100	0	0	100	0
		20	1200	0	24	76	0	100	0	0	100	0
			3000	0	52	48	0	100	0	0	100	0
			6000	0	3	97	0	100	0	0	100	0

Table 15 and Table 16 present the results of the model selection for the 3-response-style and 4-response-style mixtures. Given the results of the 2-response style mixtures, it was foreseen that the data generation model with three or four response styles would have difficulties to be identified under the 4-items conditions. The results showed that if each response style constitutes an equal proportion of population a sample size of N = 1200 with 10-items seemed to be minimum condition in which 3-response-style or 4-response-style mixtures can be correctly identified based on the BIC or the CAIC. When the mixing proportions were unequal, a sample size of N = 3000 with 10-items seemed to be necessary for the correct model selection.

Table 15. *Model Selection under the ORS-ERS-MRS Mixtures*

Information Criterion				AIC			BIC			CAIC		
Number of classes of the estimation model				2	3	4	2	3	4	2	3	4
Type of Mixture	Mixing Proportions	Item	Sample									
ORS ERS MRS	33:33:33	4	1200	12	74	14	99	1	0	100	0	0
			3000	0	93	7	99	1	0	99	1	0
			6000	0	87	13	38	62	0	57	43	0
		10	1200	0	85	15	0	100	0	0	100	0
			3000	0	96	4	0	100	0	0	100	0
			6000	0	87	13	0	100	0	0	100	0
		20	1200	0	91	9	0	100	0	0	100	0
			3000	0	84	16	0	100	0	0	100	0
			6000	0	88	12	0	100	0	0	100	0
	80:10:10	4	1200	67	29	4	100	0	0	100	0	0
			3000	14	61	25	84	16	0	93	7	0
			6000	3	57	40	96	4	0	97	3	0
		10	1200	0	75	25	94	6	0	100	0	0
			3000	0	75	25	0	100	0	0	100	0
			6000	0	69	31	0	100	0	0	100	0
		20	1200	0	91	9	0	100	0	5	95	0
			3000	0	51	49	0	100	0	0	100	0
			6000	0	71	29	0	100	0	0	100	0

Table 16. *Model Selection under the ORS-ERS-MRS-ARS Mixtures*

Information Criterion				AIC			BIC			CAIC		
Number of classes of the estimation model				3	4	5	3	4	5	3	4	5
Type of Mixture	Mixing Proportions	Item	Sample									
ORS ERS MRS ARS	25:25:25:25	4	1200	33	65	2	99	1	0	99	1	0
			3000	16	16	68	99	1	0	99	1	0
			6000	0	88	12	94	6	0	99	1	0
		10	1200	0	89	11	4	96	0	23	77	0
			3000	0	84	16	0	100	0	0	100	0
			6000	0	91	9	0	100	0	0	100	0
		20	1200	0	96	4	0	100	0	0	100	0
			3000	0	86	14	0	100	0	0	100	0
			6000	0	71	29	0	100	0	0	100	0
	70:10:10:10	4	1200	62	26	12	100	0	0	100	0	0
			3000	24	41	35	96	4	0	99	1	0
			6000	1	67	32	46	54	0	48	52	0
		10	1200	0	80	20	44	56	0	45	55	0
			3000	0	78	22	7	93	0	7	93	0
			6000	0	59	41	0	100	0	0	100	0
		20	1200	0	92	8	0	100	0	6	94	0
			3000	0	72	28	0	100	0	0	100	0
			6000	0	37	63	0	100	0	0	100	0

4.3 Classification of Respondents

The simulation results regarding classification of respondents with respect to their response style are presented in two parts separately: *i*) for correct classifications and *ii*) misclassifications. The mean percentage of respondents who were correctly assigned to their true (generated) class membership was computed over one hundred replications as an index of classification accuracy. Likewise, the mean percentage of respondents who were incorrectly assigned to a class other than their true class was computed as an index of misclassification rate. In addition, the standard error (SE) of

the classification accuracy as well as the SE of the misclassification were obtained by computing the standard deviation of the one-hundred percentage values.

4.3.1. Classification accuracy

Classification accuracy for each response class is presented in Table 17 along with the SE of the classification accuracy in parentheses. The blank cells in the table represent the conditions for which a high proportion of replications presented estimation problems and thus, the classification accuracy was not computed. The cells marked with asterisks in the table are the conditions in which a high percentage of unsolvable label switching problems occurred. For those conditions, the classification accuracy was computed with a fewer number of solutions, the ones excluding unsolvable replications.

The conditions marked with asterisks, however, presented an unexpected trend in the simulation results. In these conditions, although the simulated testing circumstances were relatively "poor" (e.g. smaller number of test items and small sample size) the classification accuracy turned out to be better. One explanation for this aberrant trend could be that because the solutions that achieved relatively more accurate estimates were selectively retained. It was also clearly shown that the classification accuracies were accompanied with very high SE under those conditions. Taking all of this information into account, the conditions marked with asterisks were excluded from the ANOVA analysis along with the conditions with estimation problems.

Table 17. *Percentages of Correct Classification and Standard Errors of Classification Accuracy*

Type of mixture and mixing proportions		ORS 0.5 ERS 0.5		ORS 0.9 ERS 0.1		ORS 0.5 MRS 0.5		ORS 0.9 MRS 0.1		ORS 0.5 ARS 0.5		ORS 0.9 ARS 0.1	
Assigned class		ORS	ERS	ORS	ERS	ORS	MRS	ORS	MRS	ORS	ARS	ORS	ARS
Item	Sample size												
4	1200	80.78	86.88	90.14*	66.14*	80.60*	72.60*	65.32*	71.69*	94.07	94.02	98.45	86.27
		(4.1)	(5.3)	(6.0)	(11.5)	(8.0)	(7.2)	(14.0)	(22.4)	(2.2)	(2.1)	(1.1)	(4.1)
	3000	81.46	87.77	93.73	61.43	90.70	58.32	91.72*	50.02*	94.50	93.77	98.80	86.41
		(2.5)	(2.3)	(2.8)	(10.2)	(2.4)	(5.5)	(3.4)	(10.3)	(1.8)	(1.8)	(0.7)	(2.4)
	6000	81.35	88.26	95.19	58.04	91.09	57.97	69.27	58.22	95.05	93.41	98.81	86.40
		(2.2)	(1.9)	(1.3)	(5.7)	(1.8)	(4.9)	(3.0)	(4.8)	(1.2)	(1.3)	(0.6)	(1.8)
10	1200	93.15	96.30	97.36	86.58	90.85	87.94			98.00	98.64		
		(1.2)	(1.0)	(0.7)	(4.1)	(1.9)	(2.2)			(0.5)	(0.7)		
	3000	93.32	96.27	97.39	87.18	91.05	88.24	97.45	69.15	98.91	98.79	99.76	96.27
		(0.9)	(0.6)	(0.4)	(2.7)	(1.1)	(1.3)	(0.9)	(4.2)	(0.4)	(0.6)	(0.2)	(1.3)
	6000	93.19	96.39	97.48	87.50	91.38	88.06	98.04	67.88	98.93	98.88	99.84	96.18
		(0.6)	(0.5)	(0.3)	(1.5)	(0.8)	(0.8)	(0.4)	(2.3)	(0.2)	(0.4)	(0.1)	(0.9)
20	1200	96.69	98.93	98.29	96.22	96.44	94.19			99.69	99.61		
		(0.6)	(0.4)	(0.5)	(1.8)	(1.0)	(1.5)			(0.3)	(0.3)		
	3000	96.73	98.88	98.38	96.31	96.49	94.38	99.04	86.17	99.74	99.63	99.90	98.84
		(0.5)	(0.3)	(0.3)	(1.2)	(0.5)	(0.7)	(0.2)	(2.4)	(0.2)	(0.2)	(0.1)	(0.7)
	6000	96.77	98.94	98.39	96.45	96.76	94.16	99.15	86.14	99.76	99.64	99.94	98.87
		(0.4)	(0.2)	(0.2)	(0.9)	(0.4)	(0.5)	(0.2)	(1.5)	(0.1)	(0.2)	(0.1)	(0.5)

Table 17_continued.

Type of mixture and mixing proportions		ORS 0.33 ERS0.33 MRS 0.33			ORS 0.8 ERS 0.1 MRS 0.1		
Assigned class		ORS	ERS	MRS	ORS	ERS	MRS
Item	Sample size						
4	1200	55.27* (11.1)	78.45* (13.4)	77.32* (8.4)	59.44* (13.2)	69.38* (13.7)	65.47* (10.6)
	3000	59.29 (9.5)	83.16 (9.3)	75.61 (8.7)	73.66* (12.7)	60.48* (11.5)	64.32* (11.5)
	6000	61.6 (10.0)	84.63 (7.7)	74.51 (9.8)	80.27* (14.0)	61.11* (9.5)	55.03* (15.5)
10	1200	83.27 (3.4)	95.73 (1.4)	87.60 (2.9)			
	3000	83.98 (1.9)	96.06 (0.9)	88.65 (1.5)	94.76 (0.7)	87.43 (2.7)	70.49 (3.7)
	6000	84.26 (1.5)	96.14 (0.6)	88.74 (1.2)	95.32 (0.6)	87.56 (1.8)	69.01 (2.6)
20	1200	93.06 (1.5)	98.80 (0.7)	94.14 (1.4)			
	3000	93.27 (0.9)	98.84 (0.4)	94.42 (1.0)	97.18 (0.4)	96.38 (1.1)	86.71 (2.3)
	6000	93.42 (0.8)	98.83 (0.3)	94.43 (0.7)	97.32 (0.3)	96.82 (0.9)	86.75 (2.1)

Type of mixture and mixing proportions		ORS 0.25 ERS 0.25 MRS 0.25 ARS 0.25				ORS 0.7 ERS 0.1 MRS 0.1ARS 0.1			
Assigned class		ORS	ERS	MRS	ARS	ORS	ERS	MRS	ARS
Item	Sample size								
4	1200								
	3000								
	6000								
10	1200								
	3000	84.02 (1.6)	94.14 (0.8)	88.49 (1.4)	95.07 (0.6)				
	6000	84.03 (1.2)	94.04 (0.7)	88.81 (1.0)	95.16 (0.5)	94.76 (0.6)	86.44 (2.3)	70.70 (2.8)	94.20 (1.1)
20	1200	92.66 (1.8)	98.26 (0.9)	94.21 (1.8)	97.37 (1.1)				
	3000	93.24 (1.1)	98.24 (0.5)	94.03 (1.0)	97.53 (0.6)	97.16 (0.4)	95.95 (1.5)	87.94 (2.1)	97.15 (1.0)
	6000	93.42 (0.7)	98.37 (0.3)	94.49 (0.6)	97.52 (0.4)	97.25 (0.3)	96.13 (0.8)	87.47 (1.5)	97.19 (0.6)

Note.* Calculated excluding some of replications for which switched labels were unsolvable.

In the following reports of the factorial ANOVA results, only the effects that were both statistically and practically significant are interpreted for their importance.

Overall classification accuracy. The percentages of correct classification obtained for each class were averaged across latent classes within the given mixture as an index of overall classification accuracy and used as a dependent variable of the factorial ANOVA. Table 18 summarizes the results of the factorial ANOVA on the overall classification accuracy.

Table 18. *Factorial ANOVA Results on Overall Classification Accuracy*

Source	*Type III Sum of Squares*	*Df*	*F*	*p*	η^2
Mixture	1079.91	4	460.53	0.00	0.28
Proportion	104.35	1	178.00	0.00	0.03
Item	1635.07	2	1394.57	0.00	0.42
Sample	0.04	2	0.04	0.97	0.00
mixture * item	422.27	7	102.90	0.00	0.11
proportion * item	48.93	2	41.73	0.00	0.01
item * sample	0.34	4	0.15	0.96	0.00
mixture * proportion	48.69	4	20.77	0.00	0.01
mixture * sample	0.57	8	0.12	0.99	0.00
proportion * sample	1.14	2	0.97	0.39	0.00
Error	17.00	29			
Corrected total	3908.84				

The significant factors on the overall classification accuracy were the main effect of the type of mixture ($F_{(4,29)} = 460.53$, $p < .001$; $\eta^2 = 0.28$) and test length ($F_{(3,29)} = 1394.57$, $p < .001$; $\eta^2 = 0.42$), as well as the interaction effect between type of mixture and test length ($F_{(7,29)} = 102.90$, $p < .001$; $\eta^2 = 0.11$). The effect sizes of

the two main effects were large whereas that of the interaction effect was medium. Table 19 presents the cell means of the classification accuracy at the levels of independent variables of the test length and type of mixture.

Table 19.*Cell Means of the Overall Classification Accuracy*

	Item	Mixture					
		OE	OM	OA	OEM	OEMA	Total
Mean	4	81.49	70.93	93.33	73.13	na	79.72
	10	93.51	87.00	98.42	87.27	89.16	91.07
	20	97.58	94.29	99.56	94.69	95.28	96.28
	Total	91.41	86.10	96.87	88.00	92.98	

For the significant main effects, post-hoc comparisons were conducted. The results of the Tukey HSD (with $\alpha_{FW} = .05$) tests showed that the overall classification accuracy differ significantly among all five different types of mixtures as well as among the three levels of test length. As expected in the earlier sections based on the degrees of heterogeneity in the thresholds plots, the mixtures of ORS and ARS respondents were most accurately classified (96.87 %) while the ORS and MRS mixtures were most difficult to be correctly distinguished (86.10 %). The 3-response-style mixtures showed lower level of overall classification accuracy than the 4-response-style mixtures. It seems to be because of the contribution of the low classification accuracy of the MRS class to the overall classification for the 3-response-style mixtures and also the contribution of the high classification accuracy of the ARS class for the 4-response-style mixtures. Regardless of the type of mixture, the overall classification accuracy was higher than 94% when the test length was $I = 20$.

The interaction effect between the type of mixture and test length was further investigated. In the interaction plot presented in Figure 16, it was observed that the increase in the classification accuracy between the test length of $I = 10$ and $I = 20$ for the ORS-ARS mixture was relatively smaller than that increase for other types of mixture. The pairwise comparisons of the three levels of test length for each type of mixture showed that the increase between $I = 10$ and $I = 20$ for the ORS-ARS mixture was significant at the $p < .05$ whereas that increase for the other four mixtures was significant at the $p < .001$.

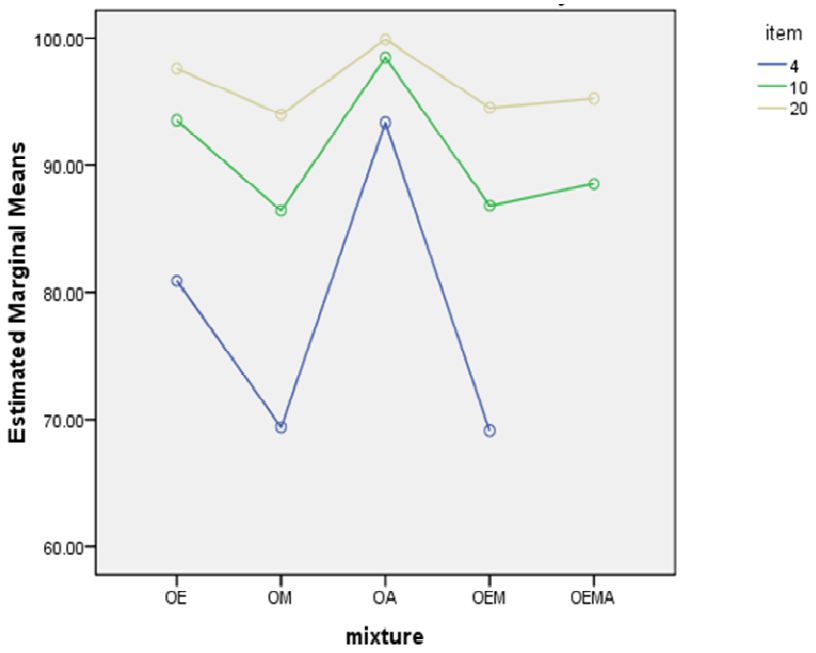

Figure 16. Interaction effect between type of mixture and test length on the overall classification accuracy

Classification accuracy for each response style. In addition to the overall classification accuracy, the classification accuracy for each response-style class was

also evaluated. Table 20 summarizes the four factorial ANOVA results and presents only significant effects that met both the statistical and practical significance criteria.

Table 20. *Effect size (η^2) for the Classification Accuracy Conditional on Statistical Significance ($p < 0.05$)*

Source	ORS	ERS	MRS	ARS
mixture	0.21			
proportion		0.14		
item	0.26	0.35	0.23	0.49
mixture * item	0.14			

Test length was the common factor influencing the classification of ORS, ERS, MRS, and ARS respondents. Regardless of the type of response style, as the number of items increased, the correct classification rate increased with a significant difference: M_4 (86.51) < M_{10} (93.35) < M_{20} (96.93) for ORS: M_4 (78.60) < M_{10} (91.98) < M_{20} (97.65) for ERS: M_4 (64.93) < M_{10} (81.06) < M_{20} (91.31) for MRS: and M_4 (90.05) < M_{10} (96.65) < M_{20} (98.34) for ARS.

The results of the Tukey HSD (with $\alpha_{FW} = .05$) tests on the main effect of the type of mixture showed that 98.38% of ORS respondents were correctly classified in the ORS-ARS mixtures whereas only 86.39% of them were correctly identified in the ORS-ERS-MRS mixtures. In the rest of the mixtures, 92.83 % of ORS respondents on average were correctly classified. These classification accuracy rates were statistically significantly different ($p < .05$). The interaction effect found for the ORS class was in the same pattern as the interaction effect for the overall classification accuracy.

The mixing proportions influenced the classification of ERS respondents. Under the equal proportions conditions, ERS respondents were classified significantly better than under the unequal proportions conditions: $M_{unequal}$ (82.97) $<$ M_{equal} (94.12) ($p < .001$).

A noteworthy result in the classification accuracy analysis was that the sample size was not a significant factor. As may be noticed in Table 17, the differences in the classification accuracy rates at the three sample sizes were negligible in most of the conditions. If this model is used in empirical studies to detect people with different response styles, the number of items of an instrument is the most important factor to be considered. As long as a sufficient number of items (at least ten items) is used, a sample with $N = 1200$ would provide an equivalent level of classification accuracy as a larger sample with $N = 6000$ would provide.

4.3.2. Misclassification

To investigate whether misclassification occurred particularly between certain types of response styles, the 3-response-style mixtures and 4-response-style mixtures were examined with respect to all possible mismatching between true (generated) and assigned class. Since classification rates did not significantly differ at different levels of sample size, Table 21 summarizes the marginal misclassification rates over the three levels of sample size.

Table 21. *Percentages of Misclassified Respondents*

Type of mixture		ORS – ERS – MRS											
True class and mixing proportions		ORS 0.33	ORS 0.33	ERS 0.33	ERS 0.33	MRS 0.33	MRS 0.33	ORS 0.7	ORS 0.7	ERS 0.1	ERS 0.1	MRS 0.1	MRS 0.1
Assigned class		ERS	MRS	ORS	MRS	ORS	ERS	ERS	MRS	ORS	MRS	ORS	ERS
Item	4	17.50	23.78	16.58	1.33	21.53	2.65	9.86	19.00	31.79	4.56	36.59	2.20
	10	6.62	9.53	3.92	0.11	10.38	1.28	2.64	2.33	12.35	0.16	29.68	0.58
	20	3.13	3.60	1.17	0.01	4.96	0.72	2.37	2.29	2.38	0.02	8.90	0.55
	Total	9.08	12.30	7.22	0.48	12.29	1.55	4.96	7.87	15.50	1.58	25.06	1.11
Type of mixture		ORS – ERS – MRS – ARS											
True class and mixing proportions		ORS 0.25	ORS 0.25	ORS 0.25	ERS 0.25	ERS 0.25	ERS 0.25	MRS 0.25	MRS 0.25	MRS 0.25	ARS 0.25	ARS 0.25	ARS 0.25
Assigned class		ERS	MRS	ARS	ORS	MRS	ARS	ORS	ERS	ARS	ORS	ERS	MRS
Item	4												
	10	6.46	9.19	0.32	3.67	0.07	2.12	10.17	1.30	0.04	0.40	4.52	0.01
	20	3.17	3.85	0.04	1.19	0.01	0.56	5.01	0.87	0.01	0.08	2.48	3.17
	Total	4.82	6.52	0.18	2.43	0.04	1.34	7.59	1.30	0.87	0.03	0.08	0.40
True class and mixing proportions		ORS 0.7	ORS 0.7	ORS 0.7	ERS 0.1	ERS 0.1	ERS 0.1	MRS 0.1	MRS 0.1	MRS 0.1	ARS 0.1	ARS 0.1	ARS 0.1
Assigned class		ERS	MRS	ARS	ORS	MRS	ARS	ORS	ERS	ARS	ORS	ERS	MRS
Item	4												
	10	2.75	2.39	0.10	11.47	0.23	1.85	28.71	0.58	0.01	1.65	4.14	0.01
	20	1.72	1.07	0.01	3.47	0.00	0.50	12.10	0.19	0.01	0.31	2.52	0.00
	Total	2.24	1.73	0.06	7.47	0.12	1.18	20.41	0.39	0.01	0.98	3.33	0.01

For the 3-response-style mixtures, the most commonly occurring misspecification was the misclassification of MRS respondents within the ORS class (MO) under unequal mixing proportions, followed by the misclassification of ERS respondents within the ORS class (EO) under unequal mixing proportions (hereafter a misclassification of "*A*" respondents within the "*B*" class is referred to as AB while a misclassification of "*B*" respondents within the "*A*" class is referred to as BA). On the other hand, EM and ME rarely occurred. Especially, when the test length was long and, thus, overall classification accuracy was high, the chance of EM was essentially zero. This rare occurrence of EM was consistent regardless of the mixing proportions.

When the mixing proportions were equal, OE and EO as well as OM and MO did not differ significantly. However, when the mixing proportions were unequal (i.e., 10 % of population was MRS or ERS respondents while the majority was ORS respondents), MO and EO significantly increased (25.06 % and 15.50 %, respectively). It seems that it was easier for the distorted response-style respondents to be misclassified within the normal response-style respondent if the distorted group was a small sized group. However, this trend was not observed for the ARS class.

Under the 4-response-style mixture, the chance of MO and EO was also significantly high (20.41% and 7.47 %, respectively) as well as EM and ME again rarely occurred (0.08 % and 0.85 %, respectively). In addition, there were several other misclassifications that were associated with essentially zero chance of occurrence. They were OA (0.12 %), MA (0.44 %), AO (0.5%), and AM (0.2%).

4.4 Threshold Parameter Recovery

Recovery of item thresholds was evaluated with respect to the RMSE, Pearson correlation, and SE. Initially, these three evaluation measures were assessed for each of the four thresholds. The evaluation measures were then averaged across the four thresholds for use in the ANOVA analysis. The averaged evaluation measures are provided in the following sections. Sections 4.1.1, 4.4.2, and 4.4.3 discuss the results for each of the evaluation criteria based on the factorial ANOVA.

4.4.1. Evaluation of the RMSE.

The averaged RMSE is presented in Table 22 followed by the factorial ANOVA results for each latent class in the subsequent section.

Table 22. *RMSE of Threshold Parameter Estimates*

Type of mixture			ORS ERS		ORS MRS		ORS ARS		ORS ERS MRS			ORS ERS MRS ARS			
Class			ORS	ERS	ORS	MRS	ORS	ARS	ORS	ERS	MRS	ORS	ERS	MRS	ARS
Mixing	Sample	Item													
Proportions	1200	4	.194	.238											
		10	.144	.201	.153	.281	.137	.275	.197	.252	.348				
		20	.140	.195	.144	.230	.139	.280	.179	.238	.291	.209	.283	.339	.411
		4	.117	.162	.207	.536	.095	.222	.305	.231	.438				
Equal	3000	10	.092	.125	.102	.166	.090	.193	.123	.155	.196	.146	.185	.245	.270
		20	.087	.125	.095	.144	.090	.197	.112	.148	.177	.131	.173	.204	.249
		4	.085	.114	.196	.434	.060	.172	.212	.166	.318				
	6000	10	.064	.090	.078	.115	.066	.156	.089	.113	.141	.102	.130	.166	.190
		20	.062	.087	.072	.103	.066	.159	.080	.106	.124	.098	.125	.146	.139
	1200	4													
		10	.104	.518											
		20	.103	.461											
		4	.091	.511											
Unequal	3000	10	.067	.315	.072	.614	.068	.445	.077	.331	.557				
		20	.066	.285	.070	.391	.068	.403	.070	.287	.332	.076	.287	.382	.420
		4	.057	.358	.196	.455	.050	.411							
	6000	10	.046	.225	.054	.357	.051	.301	.053	.229	.370	.060	.253	.343	.351
		20	.046	.201	.052	.252	.050	.283	.051	.206	.258	.057	.231	.260	.289

ORS class. The factorial ANOVA results of the RMSE of threshold parameter estimates for the ORS class (RMSE-threshold-ORS) are presented in Table 23.

Table 23. *Factorial ANOVA Results on the RMSE of Threshold Estimates for ORS Class*

Source	*Type III Sum of Squares*	*Df*	*F*	*P*	η^2
Mixture	0.029	4	102.08	0.00	0.16
proportion	0.005	1	67.53	0.00	0.03
Sample	0.017	2	120.08	0.00	0.09
Item	0.024	2	168.67	0.00	0.13
mixture * item	0.032	7	65.02	0.00	0.18
proportion * item	0.000	2	0.61	0.55	0.00
sample * item	0.001	4	2.89	0.04	0.01
mixture * proportion	0.001	4	4.18	0.01	0.01
mixture * sample	0.002	8	3.28	0.01	0.01
proportion * sample	0.000	2	3.53	0.04	0.00
Error	0.002	27			
Corrected total	0.182	63			

The significant factors on the RMSE-threshold-ORS were the main effect of the type of mixture ($F_{(4,27)} = 102.08$, $p < .001$; $\eta^2 = 0.16$), sample size ($F_{(2,27)} = 120.08$, $p < .001$; $\eta^2 = 0.09$), test length ($F_{(2,27)} = 168.67$, $p < .001$; $\eta^2 = 0.13$), as well as the interaction effect between type of mixture and test length ($F_{(7,27)} = 65.02$, $p < .001$; $\eta^2 = 0.18$). Table 24 presents the cell means of the RMSE at the levels of independent variables of the type of mixture, sample size, and test length.

Table 24.*Cell Means of the RMSE of Threshold Estimates for the ORS Class*

Sample	Item	Mixture				
		OE	OM	OA	OEM	OEMA
1200	4	0.194	na	na	na	na
	10	0.124	0.153	0.137	0.197	na
	20	0.122	0.144	0.139	0.179	0.209
	total	0.137	0.149	0.138	0.188	0.209
3000	4	0.104	0.207	0.082	0.305	na
	10	0.080	0.087	0.079	0.100	0.146
	20	0.077	0.083	0.079	0.091	0.104
	total	0.087	0.109	0.080	0.137	0.118
6000	4	0.071	0.196	0.055	0.212	na
	10	0.055	0.066	0.059	0.071	0.081
	20	0.054	0.062	0.055	0.066	0.078
	total	0.060	0.108	0.056	0.097	0.079

In general, the RMSE-threshold-ORS decreased consistently as the sample size and test length increased in each type of mixture. For the significant main effects, the post-hoc comparisons were conducted. The results of the Tukey HSD (with $\alpha_{FW} = .05$) test showed that RMSE-threshold-ORS differed as following: M_{OA} (0.078) < M_{OE} (0.092) < M_{OEMA} (0.110) = M_{OM} (0.115) < M_{OEM} (0.129), where inequality sign indicates a significant difference and equality sign indicates an insignificant difference

Regarding the main effect of the sample size, the decrease in the RMSE-threshold-ORS as sample size increased was significant between all three levels based on the Tukey HSD test (with $\alpha_{FW} = .05$): M_{1200} (0.154) > M_{3000} (0.103) > M_{6000} (0.080). Regarding the main effect of the test length, the decrease in the RMSE-threshold-ORS was significant as the test length increased from $I = 4$ to $I = 10$ but was not significant as the test length increased from $I = 10$ to $I = 20$: M_4 (0.138) > M_{10} (0.093) = M_{20} (0.093).

The significant interaction effect between the type of mixture and test length is depicted in Figure 17. In the figure, clearly seen is the superior recovery of the ORS threshold parameters in the ORS-ARS mixture even at the $I = 4$ level. The pairwise comparisons of the three levels of test length for each type of mixture showed that the increase in the RMSE from $I = 4$ to $I = 10$ as well as that from $I = 10$ to $I = 20$ was not statistically significant for the ORS-ARS mixture while the decrease in the RMSE from $I = 4$ to $I = 10$ was significant for all other types of mixture.

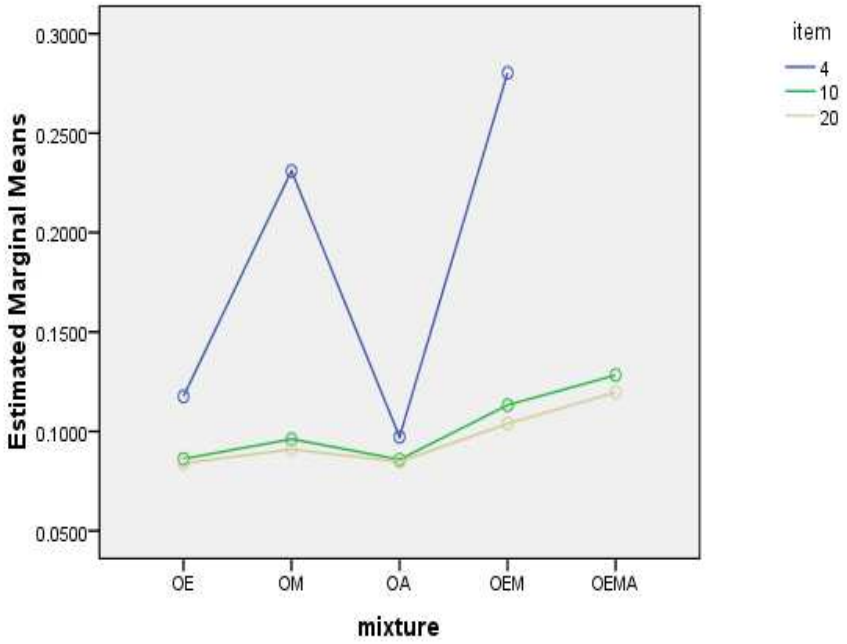

Figure 17. Interaction effect between type of mixture and test length on the RMSE of threshold estimates for the ORS class

ERS class. The factorial ANOVA results of the RMSE of threshold parameter estimates for the ERS class (RMSE-threshold-ERS) are presented in Table 25.

Table 25. *Factorial ANOVA Results on the RMSE of Threshold Estimates for the ERS Class*

Source	*Type III Sum of Squares*	*Df*	*F*	*p*	η^2
Mixture	0.008	2	30.06	0.00	0.02
Proportion	0.133	1	995.61	0.00	0.32
Sample	0.074	2	275.62	0.00	0.18
Item	0.029	2	107.41	0.00	0.07
mixture * item	0.001	3	2.46	0.11	0.00
proportion * item	0.016	2	60.20	0.00	0.04
sample * item	0.002	4	2.93	0.07	0.00
mixture * proportion	0.001	2	1.99	0.18	0.00
mixture * sample	0.001	4	2.06	0.15	0.00
proportion * sample	0.018	2	66.68	0.00	0.04
Error	0.002	12			
Corrected total	0.417	36			

The significant factors on the RMSE-threshold-ERS were the main effect of the mixing proportions ($F_{(1,12)} = 995.61$, $p < .001$; $\eta^2 = 0.32$), sample size ($F_{(2,12)} = 275.62$, $p < .001$; $\eta^2 = 0.18$), and test length ($F_{(2,12)} = 107.41$, $p < .001$; $\eta^2 = 0.07$). Table 26 presents the cell means of the RMSE at the levels of independent variables of the mixing proportions, sample size, and test length.

Table 26. *Cell Means of the RMSE of Threshold Estimates for the ERS Class*

Sample	Item	Mixing Proportions	
		Equal	Unequal
1200	4	0.238	na
	10	0.227	0.518
	20	0.239	0.461
	total	0.235	0.490
3000	4	0.197	0.511
	10	0.155	0.323
	20	0.149	0.286
	total	0.163	0.336
6000	4	0.140	0.358
	10	0.111	0.236
	20	0.106	0.213
	total	0.116	0.243

The Tukey HSD (with $\alpha_{FW} = .05$) test showed the same patterns of significant differences as those that were observed for the RMSE-threshold-ERS. Regarding the main effect of the sample size, the decrease in the RMSE-threshold-ERS as sample size increased was significant between all three levels: M_{1200} (0.298) > M_{3000} (0.237) > M_{6000} (0.176). Regarding the main effect of the test length, the decrease in the RMSE-threshold-ERS was significant as the test length increased from $I = 4$ to $I = 10$ but was not significant as the test length increased from $I = 10$ to $I = 20$: M_4 (0.254) > M_{10} (0.223) = M_{20} (0.215).

The main effect of the mixing proportions showed a smaller RMSE when the mixing proportions were equal: $M_{Unequal}$ (0.313) > M_{Equal} (0.166). The mixing proportion was not a significant factor for the ORS class. It was a significant factor for the ERS class as well as the other two classes. It makes sense because the ORS class

always took on the larger proportion of the generated samples while the ERS, MRS, and ARS took on only 10% of the respondents.

MRS class. The factorial ANOVA results of the RMSE of threshold parameter estimates for the MRS class (RMSE-threshold-MRS) are presented in Table 27.

Table 27.*Factorial ANOVA Results on the RMSE of Threshold Estimates for the MRS Class*

Source	*Type III Sum of Squares*	*Df*	*F*	*p*	η^2
Mixture	0.005	2	3.97	0.05	0.01
Proportion	0.077	1	120.46	0.00	0.13
Sample	0.128	2	99.95	0.00	0.22
Item	0.096	2	74.69	0.00	0.17
mixture * item	0.012	3	6.32	0.01	0.02
proportion * item	0.039	2	30.22	0.00	0.07
sample * item	0.008	3	3.93	0.04	0.01
mixture * proportion	0.005	2	4.09	0.05	0.01
mixture * sample	0.002	4	0.76	0.58	0.00
proportion * sample	0.016	1	25.66	0.00	0.03
Error	0.006	10			
Corrected total	0.573	32			

As was found for the ERS class, the significant factors on the RMSE-threshod-MRS were the main effect of the mixing proportions ($F_{(1,10)} = 120.46$, $p < .001$; $\eta^2 = 0.13$), sample size ($F_{(2,10)} = 99.95$, $p < .001$; $\eta^2 = 0.22$), and test length ($F_{(2,10)} = 74.69$, $p < .001$; $\eta^2 = 0.17$). While the most influencing factor was the mixing proportions for the ERS class, the sample size was the most important factor for the MRS class. Table 28 presents the cell means of the RMSE at the levels of independent variables of the mixing proportions, sample size, and test length.

Table 28. *Cell Means of the RMSE of Threshold Estimates for the MRS class*

Sample	Item	Mixing Proportions	
		Equal	Unequal
1200	4	na	na
	10	0.31	na
	20	0.29	na
	total	0.30	na
3000	4	0.49	na
	10	0.20	0.59
	20	0.18	0.37
	total	0.26	0.46
6000	4	0.38	0.46
	10	0.14	0.36
	20	0.12	0.26
	total	0.19	0.33

The Tukey HSD (with $\alpha_{FW} = .05$) test showed the same patterns of significant differences as those were observed for the previous two classes. Regarding the main effect of the sample size, the significant differences were as following: M_{1200} (0.256) > M_{3000} (0.298) > M_{6000} (0.337). Regarding the main effect of the test length, the significant differences were as following: M_4 (0.436) > M_{10} (0.300) = M_{20} (0.242). In addition, the main effect of the mixing proportions showed the significant difference: $M_{Unequal}$ (0.381) > M_{Equal} (0.193).

ARS class. The factorial ANOVA results of the RMSE of threshold parameter estimates for the ARS class (RMSE-threshold-ARS) are presented in Table 29.

Table 29. *Factorial ANOVA Results on the RMSE of Threshold Estimates for the ARS Class*

Source	*Type III Sum of Squares*	*Df*	*F*	*p*	η^2
Mixture	0.011	1	43.47	0.00	0.06
Proportion	0.094	1	360.78	0.00	0.49
Sample	0.055	2	105.98	0.00	0.29
Item	0.008	2	16.29	0.01	0.04
mixture * item	0.001	1	5.65	0.08	0.01
proportion * item	0.004	2	7.65	0.04	0.02
sample * item	0.000	3	0.02	1.00	0.00
mixture * proportion	0.000	1	0.01	0.92	0.00
mixture * sample	0.006	2	10.76	0.03	0.03
proportion * sample	0.004	1	16.38	0.02	0.02
Error	0.001	4			
Corrected total	0.190	20			

The significant factors on the RMSE-threshold-ARS were the type of mixture ($F_{(1,4)}$ = 43.47, p < .001; η^2 = 0.06), mixing proportions ($F_{(1,4)}$ =360.78, p < .001; η^2 = 0.49), and sample size ($F_{(2,4)}$ = 105.98, p < .001; η^2 = 0.29). Unlike for the other classes, test length was not significant for the ARS class. Table 30 presents the cell means of the RMSE at the levels of independent variables of the mixing proportions, sample size, and test length.

The Tukey HSD (with α_{FW} = .05) test showed the decrease in the RMSE-thr-ARS from N = 1200 to N = 3000 was not significant while the decrease from N = 3000 to N = 6000 was significant: M_{1200} (0.336) = M_{3000} (0.322) > M_{6000} (0.241). Regarding the main effect of the test length, the significant differences were found between I = 4 and I = 10: M_4 (0.358) > M_{10} (0.273) = M_{20} (0.279). In addition, the main effect of the mixing proportions showed the significant difference: $M_{Unequal}$ (0.387) > M_{Equal} (0.163).

Table 30.*Cell Means of the RMSE of Threshold Estimates for the ARS class*

Proportion	Sample	Type of mixture	
		OA	OEMA
Equal	1200	0.278	0.411
	3000	0.204	0.260
	6000	0.162	0.165
	Total	0.207	0.252
Unequal	1200	na	na
	3000	0.492	0.420
	6000	0.317	0.320
	Total	0.405	0.353

4.4.2. Evaluation of the correlation

The second criterion used to evaluate the threshold parameter recovery was the Pearson correlation coefficient between generated and estimated thresholds. Table 31 reports the correlations that were averaged across the four thresholds.

The factorial ANOVA conducted on the correlation measures showed that the significant factors for each response-style class considering both statistical and practical importance turned out to be the same as those that were found to be significant on the RMSE measures. The factorial ANOVA results for the correlation measures are presented in a single table concisely in Table 31 instead of providing four analysis results in separate ANOVA tables.

Table 31.*Correlations Between Generated and Estimated Threshold Parameters*

Type of mixture			ORS ERS		ORS MRS		ORS ARS		ORS ERS MRS			ORS ERS MRS ARS			
Class			ORS	ERS	ORS	MRS	ORS	ARS	ORS	ERS	MRS	ORS	ERS	MRS	ARS
Mixing	Sample	Item													
Proportions	1200	4	.794	.694											
		10	.888	.858	.836	.852	.857	.986	.760	.792	.808				
		20	.885	.862	.839	.880	.853	.986	.786	.802	.831	.736	.748	.790	.966
		4	.910	.836	.832	.780	.930	.994	.691	.755	.758				
Equal	3000	10	.932	.933	.925	.937	.935	.995	.883	.905	.916	.848	.870	.887	.986
		20	.951	.930	.924	.946	.930	.994	.892	.908	.922	.864	.878	.903	.988
		4	.949	.908	.882	.836	.967	.997	.840	.855	.837				
	6000	10	.966	.965	.959	.971	.965	.995	.934	.947	.953	.919	.931	.937	.993
		20	.963	.964	.960	.973	.963	.997	.943	.949	.960	.920	.935	.939	.994
	1200	4													
		10	.911	.546											
		20	.909	.587											
Unequal		4	.945	.469			.964	.954							
	3000	10	.961	.705	.961	.632	.963	.963	.950	.699	.653				
		20	.959	.738	.959	.767	.960	.968	.954	.741	.751	.947	.746	.763	.965
		4	.980	.603	.895	.780	.980	.985							
	6000	10	.982	.818	.980	.840	.982	.984	.976	.827	.790	.973	.790	.803	.978
		20	.979	.843	.979	.862	.980	.985	.976	.848	.853	.975	.802	.820	.978

Table 32. *Effect size (η^2) for Correlation for Thresholds Parameters Conditional on Statistical Significance ($p < 0.05$)*

Factor	ORS	ERS	MRS	ARS
Mixture	0.15			
Proportion		0.29	0.24	0.67
Sample	0.16	0.18	0.27	0.33
Item	0.09	0.14	0.11	
mixture * item	0.08			

ORS class. The factorial ANOVA results of the correlation of threshold parameter estimates for the ORS class (Correlation-threshold-ORS) showed that the main effect of the type of mixture ($F_{(4,27)} = 73.33$, $p < .001$; $\eta^2 = 0.15$), sample size ($F_{(2,27)} = 154.60$, $p < .001$; $\eta^2 = 0.16$), test length ($F_{(2,27)} = 82.97$, $p < .001$; $\eta^2 = 0.09$), as well as the interaction effect between type of mixture and test length ($F_{(7,27)} = 22.98$, $p < .001$; $\eta^2 = 0.08$) were significant. Table 33 presents the cell means of the RMSE at the levels of independent variables of the type of mixture, sample size, and test length.

The main effect of the type of mixture differed from each other as following: M_{OA} (0.945) = M_{OE} (0.933) > M_{OM} (0.918) > M_{OEMA} (0.898) > M_{OEM} (0.882). Regarding the main effect of the sample size, the increase in the Correlation-threshold-ORS as sample size increased was significant between all three levels: M_{1200} (0.838) < M_{3000} (0.919) > M_{6000} (0.954). Regarding the main effect of the test length, the increase in the Correlation-threshold-ORS was significant as the test length increased

from $I = 4$ to $I = 10$ but was not significant as the test length increased from $I = 10$ to $I = 20$: M_4 (0.897) < M_{20} (0.923) = M_{10} (0.954).

Table 33.*Cell Means of the RMSE of Threshold Estimates for the ORS Class*

Sample	Item	Mixture				
		OE	OM	OA	OEM	OEMA
1200	4	0.794	na	na	na	na
	10	0.900	0.836	0.857	0.760	na
	20	0.897	0.839	0.853	0.786	0.736
	total	0.877	0.838	0.855	0.773	0.736
3000	4	0.928	0.832	0.947	0.691	na
	10	0.947	0.943	0.949	0.917	0.848
	20	0.955	0.942	0.945	0.923	0.906
	total	0.943	0.920	0.947	0.874	0.886
6000	4	0.965	0.889	0.974	0.840	na
	10	0.974	0.970	0.974	0.955	0.946
	20	0.971	0.970	0.972	0.960	0.948
	total	0.970	0.943	0.973	0.934	0.947

The significant interaction effect between the type of mixture and test length showed the same pattern as the interaction effect found in the RMSE-threshold-ORS evaluation. The interaction was basically due to the superior recovery for the ORS thresholds for even as the case in which only four items were used.

ERS class. The factorial ANOVA results of the Correlation-threshold-ERS showed that the main effect of the mixing proportions ($F_{(1,12)} = 1210.70$, $p < .001$; $\eta^2 = 0.29$), sample size ($F_{(2,12)} = 378.92$, $p < .001$; $\eta^2 = 0.18$), test length ($F_{(2,12)} = 286.76$, $p < .001$; $\eta^2 = 0.14$) were significant. Table 34 presents the cell means of the correlation at the levels of independent variables of the mixing proportions, sample size, and test length.

Table 34. *Cell Means of the Correlation for the ERS Class*

Sample	Item	Mixing Proportions	
		Equal	Unequal
1200	4	0.694	na
	10	0.825	0.546
	20	0.804	0.587
	total	0.793	0.567
3000	4	0.796	0.469
	10	0.903	0.702
	20	0.905	0.742
	total	0.877	0.683
6000	4	0.882	0.603
	10	0.948	0.812
	20	0.949	0.831
	total	0.932	0.790

The main effect of the mixing proportions showed a higher correlation of threshold parameters when the mixing proportions were equal: $M_{Unequal}$ (0.717) < M_{Equal} (0.874). Regarding the main effect of the sample size, the increase in the Correlation-threshold-ERS as sample size increased was significant between all three levels: M_{1200} (0.736) < M_{3000} (0.794) < M_{6000} (0.866). Regarding the main effect of the test length, the increase in the Correlation-threshold-ERS was significant as the test length increased from $I = 4$ to $I = 10$ but was not significant as the test length increased from $I = 10$ to $I = 20$: M_4 (0.731) < M_{10} (0.828) = M_{20} (0.830).

MRS class. The factorial ANOVA on the Correlation-threshold-MRS showed that the main effect of the mixing proportions ($F_{(1,10)} = 140.08$, $p < .001$; $\eta^2 = 0.24$), sample size ($F_{(2,10)} = 79.03$, $p < .001$; $\eta^2 = 0.27$), test length ($F_{(2,10)} = 31.05$, $p < .001$; $\eta^2 = 0.11$) were significant. Table 35 presents the cell means of the correlation at the levels of independent variables of the mixing proportions, sample size, and test length.

Table 35.*Cell Means of the RMSE of Threshold Estimates for the MRS Class*

Sample	Item	Mixing Proportions	
		Equal	Unequal
1200	4	na	na
	10	0.830	na
	20	0.834	na
	total	0.832	na
3000	4	0.769	na
	10	0.913	0.643
	20	0.924	0.760
	total	0.881	0.713
6000	4	0.837	0.780
	10	0.954	0.811
	20	0.957	0.845
	total	0.926	0.821

The main effect of the mixing proportions showed a higher correlation when the mixing proportions were equal: $M_{Unequal}$ (0.886) < M_{Equal} (0.776). Regarding the main effect of the sample size, the increase in the Corr-thr-MRS between N = 3000 and N = 6000 was significant but not significant between N = 1200 and N = 3000: M_{1200} (0.832) = M_{3000} (0.817) < M_{6000} (0.877). Regarding the main effect of the test length, the increase in the Correlation-threshold-MRS was significant as the test length increased from I = 4 to I = 10 but was not significant as the test length increased from I = 10 to I = 20: M_4 (0.798) < M_{10} (0.845) = M_{20} (0.864).

ARS class. The factorial ANOVA results of the Correlation-threshold-ARS showed that the main effect of the mixing proportions ($F_{(1,5)}$ = 125.59, p < .001; η^2 = 0.67) and sample size ($F_{(2,5)}$ = 35.56, p < .001; η^2 = 0.33) were significant. Table 36

presents the cell means of the correlation at the levels of independent variables of the mixing proportions and sample size.

Table 36. *Cell Means of the RMSE of Threshold Estimates for the ARS Class*

Sample	Proportions	
	Equal	Unequal
1200	0.979	na
3000	0.991	0.963
6000	0.995	0.982
total	0.990	0.973

Regarding the main effect of the sample size, the increase in the Correlation-threshold-MRS between $N = 3000$ and $N = 6000$ was significant but not significant between $N = 1200$ and $N = 3000$: M_{1200} (0.979) = M_{3000} (0.979) < M_{6000} (0.989).

4.4.3. Evaluation of the standard error

The third criterion used to evaluate the threshold parameter recovery was the standard error of estimates (SE), which was the calculated standard deviation of the estimated thresholds provided from all replications. Table 37 reports the SE that was averaged across the four thresholds. The factorial ANOVA results for the SE measures are present in a single table concisely in Table 38 instead of providing four analysis results in separate ANOVA tables.

Table 37. *SE of Threshold Parameter Estimates*

Type of mixture			ORS ERS		ORS MRS		ORS ARS		ORS ERS MRS			ORS ERS MRS ARS			
Class			ORS	ERS	ORS	MRS	ORS	ARS	ORS	ERS	MRS	ORS	ERS	MRS	ARS
Mixing	Sample	Item													
Proportions	1200	4	.118	.133											
		10	.053	.071	.056	.104	.047	.080	.078	.090	.135				
		20	.036	.046	.037	.060	.034	.062	.047	.057	.078	.054	.067	.093	.088
		4	.069	.091	.060	.274	.055	.102	.268	.152	.327				
	3000	10	.032	.044	.035	.063	.030	.053	.047	.051	.077	.057	.065	.097	.089
		20	.027	.029	.025	.036	.021	.040	.029	.035	.047	.033	.039	.054	.055
		4	.052	.064	.042	.199	.038	.073	.179	.114	.246				
	6000	10	.024	.031	.024	.043	.021	.038	.035	.040	.057	.041	.046	.064	.064
		20	.015	.021	.016	.027	.017	.027	.021	.024	.034	.023	.026	.044	.045
	1200	4													
		10	.035	.180											
		20	.024	.113											
		4	.064	.334			.043	.354							
	3000	10	.024	.106	.029	.240	.024	.150	.047	.051	.077				
		20	.016	.073	.016	.100	.016	.099	.029	.035	.047	.019	.067	.106	.095
		4	.033	.195	.045	.211	.029	.221							
	6000	10	.016	.079	.017	.138	.017	.097	.035	.040	.057	.021	.076	.137	.105
		20	.011	.047	.011	.067	.011	.067	.021	.023	.033	.012	.049	.070	.068

Table 38. *Effect size (η^2) for the SE Conditional on Statistical Significance ($p < 0.05$)*

Factor	ORS	ERS	MRS	ARS
mixture	0.22			
proportion		0.14		0.32
item	0.23	0.31	0.34	0.36
Sample size		0.09	0.09	0.08
Mixture * item	0.28			
Proportion * item		0.10		0.12

ORS class. The factorial ANOVA results of the SE-threshold-ORS showed that the main effects of the type of mixture ($F_{(4,27)} = 80.03$, $p < .001$; $\eta^2 = 0.22$) and test length ($F_{(2,27)} = 173.73$, $p < .001$; $\eta^2 = 0.23$), as well as the interaction effect between type of mixture and test length ($F_{(7,27)} = 59.30$, $p < .001$; $\eta^2 = 0.28$) were significant. Table 39 presents the cell means of the SE at the levels of independent variables of the type of mixture and test length.

Table 39. *Cell Means of the SE of Threshold Estimates for the ORS Class*

Item	Mixture				
	OE	OM	OA	OEM	OEMA
4	0.067	0.049	0.041	0.224	na
10	0.031	0.032	0.028	0.053	0.040
20	0.022	0.021	0.020	0.032	0.028
total	0.038	0.032	0.029	0.088	0.033

The main effect of the type of mixture differed from each other as following: M_{OA} (0.029) = M_{OM} (0.032) = M_{OEMA} (0.033) = M_{OE} (0.038) < M_{OEM} (0.038). Regarding the main effect of the test length, the increase in the Correlation-threshold-ORS was significant at the three levels: $M_4 = 0.078 > M_{10} = 0.035 > M_{20} = 0.024$ in

MRS class (M_{1200} = 0.067 < M_{3000} = 0.116 < M_{6000} = 0.232), and in ARS class (M_{1200} = 0.080 < M_{3000} = 0.119 < M_{6000} = 0.276).

The significant interaction effect between the type of mixture and test length was mainly due to the poor stability for the I = 4 short test to estimate ORS thresholds in the mixture of more than two latent class parameters. The interaction plot is present in Figure 18. Pairwise comparison showed that the difference in the SE between any of the two mixtures was not significant for the I = 20 conditions.

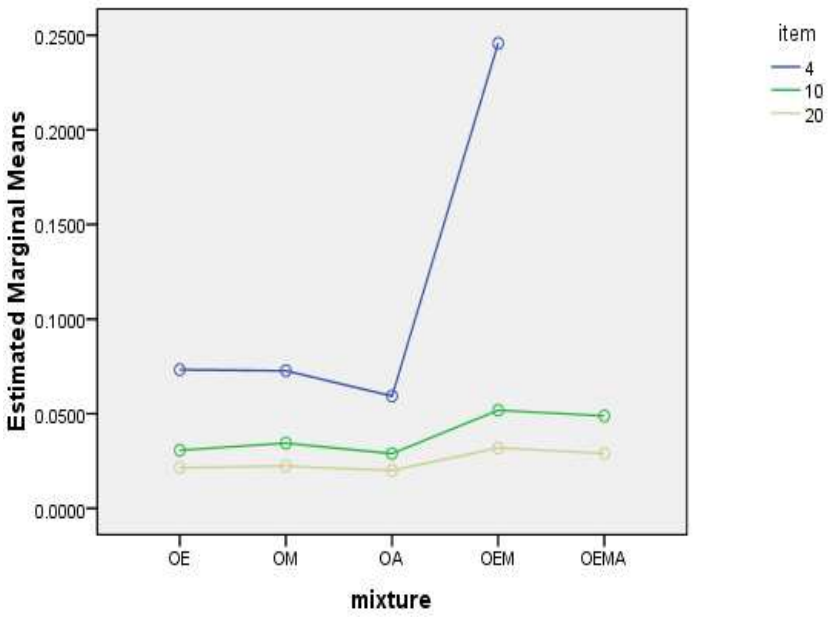

Figure 18. Interaction effect between type of mixture and test length on the SE of threshold estimates for the ORS class

ERS class. The factorial ANOVA results of the SE-threshold-ERS showed that the main effect of the mixing proportions ($F_{(1,12)}$ = 79.18, p < .001; η^2 = 0.14), test length ($F_{(2,12)}$ = 88.07, p < .001; η^2 = 0.31), and sample size ($F_{(1,12)}$ = 25.89, p < .001;

η^2 = 0.09) as well as the interaction effect between mixing proportion and test length ($F_{(2,12)}$ = 29.18, p < .001; η^2 = 0.10). Table 40 presents the cell means of the SE at the levels of independent variables of the mixing proportion, test length, and sample size.

Table 40. *Cell Means of the RMSE of Threshold Estimates for the ERS Class*

Sample	Item	Mixing Proportions	
		Equal	Unequal
1200	4	0.133	na
	10	0.081	0.180
	20	0.057	0.113
	total	0.077	0.147
3000	4	0.122	0.334
	10	0.053	0.079
	20	0.034	0.058
	total	0.063	0.111
6000	4	0.089	0.195
	10	0.039	0.065
	20	0.024	0.040
	total	0.046	0.073

The main effect of the mixing proportions showed a larger standard error when the mixing proportions were unequal: $M_{Unequal}$ (0.098) > M_{Equal} (0.061). Regarding the main effect of the sample size, the decrease in the SE was significant between N = 3000 and N = 6000 and was not significant between N = 3000 and N = 1200: M_{1200} (0.095) = M_{3000} (0.084) > M_{6000} (0.058). Regarding the main effect of the test length, the decrease was significant at all three levels: M_4 (0.155) > M_{10} (0.069) > M_{20} (0.047).

The significant interaction effect between the mixing proportion and test length was also because of the disproportionate increase in the SE for $I = 4$ condition. The interaction plot is presented in Figure 19.

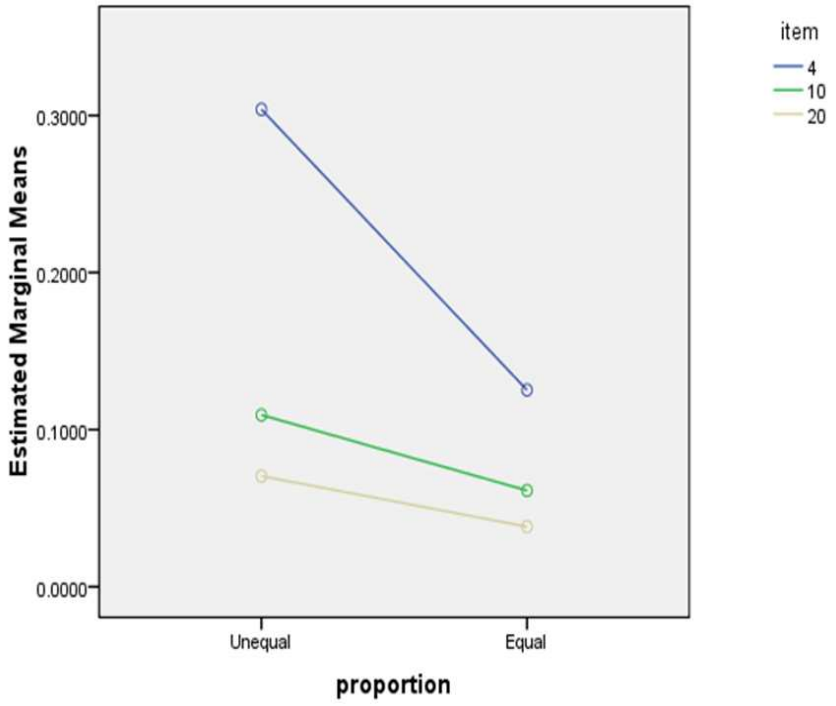

Figure 19. Interaction effect between type of mixing proportion and test length on the SE of threshold estimates for the ERS class

MRS class. The factorial ANOVA results of the SE-threshold-MRS showed that the main effect of the sample size ($F_{(2,10)} = 22.35$, $p < .001$; $\eta^2 = 0.09$) and test length ($F_{(2,10)} = 85.26$, $p < .001$; $\eta^2 = 0.34$). Table 41 presents the cell means of the SE at the levels of independent variables of the mixing proportion, test length, and sample size.

Table 41. *Cell Means of the SE of Threshold Estimates for the MRS Class*

Sample	Item	
	4	na
1200	10	0.120
	20	0.077
	total	0.094
3000	4	0.301
	10	0.111
	20	0.065
	total	0.119
6000	4	0.219
	10	0.083
	20	0.046
	total	0.095

Regarding the main effect of the test length, the decrease was significant at all three levels: M_4 (0.251) > M_{10} (0.099) > M_{20} (0.060). Based on the Tukey HSD (with α_{FW} = .05) test any of the difference in the SE between the three levels of sample size was significant: M_{1200} (0.094) = M_{3000} (0.095) > M_{6000} (0.119).

ARS class. The factorial ANOVA results of the SE-threshold-ARS showed that the same effects on the ERS class were also significant for the ARS class. The significant factors were the main effect of the mixing proportions ($F_{(1,5)}$ = 128.61, p < .001; η^2 = 0.32), test length ($F_{(2,5)}$ = 73.97, p < .001; η^2 = 0.36), and sample size ($F_{(2,5)}$ = 17.18, p < .001; η^2 = 0.08) as well as the interaction effect between mixing proportion and test length ($F_{(2,5)}$ =25.00, p < .001; η^2 = 0.12). Table 42 presents the

cell means of the SE at the levels of independent variables of the mixing proportion, test length, and sample size.

Table 42. *Cell Means of the SE of Threshold Estimates for the ARS Class*

Sample	Item	Mixing Proportions	
		Equal	Unequal
1200	4	Na	0.354
	10	0.080	0.150
	20	0.075	0.097
	total	0.077	0.175
3000	4	0.102	0.221
	10	0.071	0.101
	20	0.048	0.068
	total	0.068	0.112
6000	4	0.073	0.288
	10	0.051	0.117
	20	0.036	0.082
	total	0.049	0.140

4.5 Person Trait Parameter Recovery

The *mdltm* that uses the marginal MLE method provides as many class-specific person trait estimates (θ) as the number of classes specified in the model for each respondent. The assigned θ estimate is the one that is associated with the class for which his or her posterior probability of class membership is the highest. If a respondent is incorrectly classified, he or she is given an improper θ estimate that is estimated within those who may be qualitatively different from himself or herself.

The current study analyzed the accuracy of θ recovery for the following groups of respondents: *i*) the whole group of respondents based on their assigned class membership (i.e., all misclassified respondents were included), and *ii*) a group of correctly classified respondents. In real world data analysis, respondent's true latent

class membership is unknown information, and, hence how inaccurately his or her θ is assessed due to incorrect classification is never known. These separate analyses of θ recovery provided not only the results of the accuracy of θ recovery but also the quantification of the impact of misclassification on θ recovery. Recovery of person trait parameters was evaluated with respect to Bias, RMSE, and Pearson correlation.

4.5.1. Evaluation of the bias

The factorial ANOVA results on bias of person trait estimates showed that any of the main effects of the four manipulated factors and their two-way interaction effects were neither statistically nor practically significant. Table 43 reports the marginal bias for all respondents (whole group) and for the correctly classified respondents (selected group) in each simulation condition. As can be seen in Table 43, the bias was very small and fluctuating around zero across all simulation conditions.

Table 43. *Theta Recovery for All Respondents and Correctly Classified Respondents*

Type of mixture and Mixing proportions			ORS 0.5 ERS 0.5				ORS 0.9 ERS 0.1			
			Whole Group		Selected Group		Whole Group		Selected Group	
			ORS	ERS	ORS	ERS	ORS	ERS	ORS	ERS
Item	Sample									
		Bias	-.001	.004	.005	.001				
	1200	RMSE	.540	.559	.497	.487				
		Corr	.776	.876	.805	.882				
		Bias	-.001	.003	-.001	.002	-.002	.005	-.001	.009
4	3000	RMSE	.507	.560	.492	.487	.503	.658	.495	.530
		Corr	.785	.878	.807	.884	.839	.899	.846	.887
		Bias	.002	.003	.003	.002	.013	.009	-.002	-.006
	6000	RMSE	.503	.559	.490	.486	.516	.573	.496	.521
		Corr	.786	.879	.807	.885	.859	.861	.849	.888
		Bias	-.001	-.001	-.001	.000	.000	-.008	.000	-.009
	1200	RMSE	.368	.450	.343	.368	.358	.607	.347	.382
		Corr	.911	.916	.924	.932	.926	.908	.932	.933
		Bias	.003	.001	.002	.000	.001	.008	.001	.005
10	3000	RMSE	.366	.449	.340	.367	.355	.609	.345	.372
		Corr	.912	.917	.925	.933	.926	.909	.932	.933
		Bias	.000	.000	.001	.000	.000	.000	.000	.003
	6000	RMSE	.365	.450	.341	.365	.356	.607	.346	.369
		Corr	.912	.917	.924	.933	.926	.910	.931	.934
		Bias	-.002	.001	-.002	.000	.001	-.005	.001	-.001
	1200	RMSE	.270	.370	.252	.293	.262	.551	.255	.288
		Corr	.957	.941	.962	.957	.962	.917	.964	.958
		Bias	.001	.001	.001	.000	.001	.000	.001	.002
20	3000	RMSE	.270	.368	.253	.289	.262	.555	.255	.290
		Corr	.956	.941	.962	.958	.962	.917	.964	.958
		Bias	.000	.000	.000	.000	.001	.001	.001	.001
	6000	RMSE	.269	.367	.252	.290	.262	.554	.256	.289
		Corr	.957	.942	.962	.958	.962	.919	.964	.958

Table 43_Continued

Type of mixture and Mixing proportions			ORS 0.5 MRS 0.5				ORS 0.9 MRS 0.1			
			Whole Group		Selected Group		Whole Group		Selected Group	
Assigned class			ORS	MRS	ORS	MRS	ORS	MRS	ORS	MRS
Item	Sample									
	1200	Bias								
		RMSE								
		Corr								
4	3000	Bias	.004	.008	.003	.006				
		RMSE	.576	.733	.510	.757				
		Corr	.857	.087	.871	.080				
	6000	Bias	.000	-.010	.000	-.010	0.001	0.002	0.001	0.000
		RMSE	.575	.731	.508	.755	0.576	0.732	0.566	0.754
		Corr	.858	.032	.871	.032	0.858	0.025	0.859	0.028
10	1200	Bias	.016	-.007	.000	.003				
		RMSE	.511	.733	.358	.472				
		Corr	.861	.716	.938	.840				
	3000	Bias	-.001	.003	.001	.000	-0.001	-0.005	-0.001	0.000
		RMSE	.418	.484	.358	.468	0.370	0.532	0.354	0.501
		Corr	.931	.827	.938	.844	0.933	0.779	0.936	0.768
	6000	Bias	.000	-.001	.000	.002	0.000	0.014	0.000	0.014
		RMSE	.418	.477	.356	.469	0.373	0.487	0.356	0.482
		Corr	.932	.830	.939	.844	0.933	0.746	0.936	0.776
20	1200	Bias	-.001	.002	.001	-.003				
		RMSE	.416	.478	.268	.354				
		Corr	.933	.830	.965	.921				
	3000	Bias	-.001	-.004	.001	-.001	-0.001	0.000	-0.001	-0.002
		RMSE	.319	.368	.268	.354	0.281	0.390	0.268	0.360
		Corr	.959	.913	.965	.922	0.962	0.891	0.965	0.907
	6000	Bias	.000	-.001	.000	.001	0.000	0.002	0.000	0.001
		RMSE	.317	.367	.265	.353	0.281	0.383	0.268	0.358
		Corr	.959	.914	.965	.922	0.962	0.888	0.964	0.905

Table 43_Continued

Type of mixture and Mixing proportions			ORS 0.5 ARS 0.5				ORS 0.9 ARS 0.1			
			Whole Group		Selected Group		Whole Group		Selected Group	
Assigned class			ORS	ARS	ORS	ARS	ORS	ARS	ORS	ARS
Item	Sample									
		Bias								
	1200	RMSE								
		Corr								
		Bias	-0.005	-0.001	-0.003	-0.001	0.002	0.007	0.002	0.000
4	3000	RMSE	0.508	0.585	0.498	0.585	0.503	0.632	0.505	0.654
		Corr	0.876	0.788	0.859	0.771	0.863	0.686	0.865	0.746
		Bias	-0.003	0.008	0.000	-0.001	0.003	-0.001	0.002	0.000
	6000	RMSE	0.505	0.590	0.501	0.586	0.504	0.612	0.506	0.615
		Corr	0.864	0.812	0.860	0.765	0.863	0.706	0.866	0.761
		Bias								
	1200	RMSE								
		Corr								
		Bias	-0.002	0.025	-0.004	0.026	-0.001	0.037	-0.001	0.036
10	3000	RMSE	0.357	0.428	0.353	0.425	0.354	0.428	0.356	0.431
		Corr	0.936	0.905	0.934	0.902	0.935	0.889	0.936	0.896
		Bias	-0.001	0.024	-0.003	0.024	0.000	0.027	0.000	0.027
	6000	RMSE	0.356	0.429	0.352	0.426	0.356	0.423	0.358	0.426
		Corr	0.936	0.905	0.934	0.902	0.935	0.889	0.936	0.892
		Bias	0.001	0.027	-0.002	0.026				
	1200	RMSE	0.264	0.879	0.266	0.328				
		Corr	0.964	0.978	0.964	0.946				
		Bias	0.001	0.028	0.000	0.028	0.001	0.027	0.001	0.027
20	3000	RMSE	0.268	0.330	0.266	0.327	0.267	0.325	0.268	0.332
		Corr	0.964	0.947	0.964	0.947	0.964	0.944	0.964	0.946
		Bias	0.002	0.026	0.001	0.026	0.000	0.025	0.000	0.024
	6000	RMSE	0.267	0.329	0.264	0.326	0.267	0.321	0.268	0.326
		Corr	0.965	0.947	0.965	0.947	0.964	0.945	0.964	0.946

Table 43_Continued

Type of mixture and mixing proportions			ORS 0.33ERS 0.33MRS 0.33						ORS 0.8 ERS 0.1MRS 0.1					
			Whole Group			Selected Group			Whole Group			Selected Group		
Assigned class			ORS	ERS	MRS	ORS	ERS	MRS	ORS	ERS	MRS	ORS	ERS	MRS
Item	Sample													
		Bias												
	1200	RMSE												
		Corr												
		Bias	-.016	.002	-.004	-.002	.001	-.012						
4	3000	RMSE	.701	.591	.681	.520	.483	.664						
		Corr	.755	.871	.610	.823	.883	.624						
		Bias	.010	.001	.040	-.004	-.002	.014						
	6000	RMSE	.633	.589	.664	.505	.487	.657						
		Corr	.777	.873	.584	.830	.883	.615						
		Bias	.000	-.002	-.006	.003	-.001	-.013						
	1200	RMSE	..429	.482	.496	.345	.366	.475						
		Corr	.900	.908	.824	.927	.932	.842						
		Bias	-.001	.001	.001	.000	.000	.001	.000	.009	.007	.000	.003	.012
10	3000	RMSE	.420	.483	.485	.343	.365	.471	.680	.824	.776	.666	.692	.705
		Corr	.905	.909	.830	.928	.933	.845	.733	.772	.520	.738	.763	.571
		Bias	.000	-.003	.001	.001	-.001	.001	.000	.001	.006	.000	.003	.005
	6000	RMSE	.420	.480	.480	.343	.365	.469	.375	.603	.491	.346	.365	.481
		Corr	.906	.910	.830	.928	.933	.845	.922	.909	.757	.932	.935	.785
		Bias	.002	-.001	.001	.002	-.001	.002						
	1200	RMSE	.314	.397	.370	.254	.289	.355						
		Corr	.949	.935	.913	.962	.958	.921						
		Bias	.000	.000	.001	.001	-.001	.000	.000	-.006	.011	.000	.002	.002
20	3000	RMSE	.310	.396	.369	.251	.290	.353	.276	.552	.398	.255	.291	.356
		Corr	.950	.935	.914	.963	.958	.922	.959	.916	.885	.964	.958	.910
		Bias	.000	-.002	.003	.000	-.001	-.002	-.008	-.004	.012	.000	-.001	.000
	6000	RMSE	.312	.394	.370	.252	.290	.352	.311	.334	.436	.256	.287	.287
		Corr	.950	.936	.913	.963	.958	.923	.951	.945	.907	.964	.959	.959

Table 43_Continued

	Type of mixture and mixing proportions		ORS 0.25 ERS 0.25 MRS 0.25 ARS 0.25								ORS 0.7 ERS 0.1 MRS 0.1 ARS 0.1							
			Whole Group				Selected Group				Whole Group				Selected Group			
	Assigned class		ORS	ERS	MRS	ARS	ORS	ERS	MRS	ARS	ORS	ERS	MRS	ARS	ORS	ERS	MRS	ARS
Item	Sample																	
		Bias																
	1200	RMSE																
		Corr																
		Bias																
4	3000	RMSE																
		Corr																
		Bias																
	6000	RMSE																
		Corr																
		Bias																
	1200	RMSE																
		Corr																
		Bias	.001	.000	.002	-.002	-.011	-.014	-.002	.006								
10	3000	RMSE	.334	.370	.465	.418	.408	.541	.481	.440								
		Corr	.908	.901	.801	.865	.915	.903	.815	.875								
		Bias	-.011	-.014	-.002	.006	-.012	-.007	.006	.001	.001	-.001	.000	.001	.001	-.001	.000	.001
	6000	RMSE	.338	.371	.469	.420	.408	.541	.481	.440	.378	.632	.508	.442	.347	.372	.479	.424
		Corr	.915	.903	.815	.875	.915	.903	.815	.875	.921	.905	.760	.868	.931	.933	.791	.879
		Bias	-.001	-.001	.001	-.003	-.001	-.001	.001	.003								
	1200	RMSE	.254	.291	.355	.317	.312	.419	.399	.324								
		Corr	.950	.932	.905	.939	.950	.932	.905	.938								
		Bias	.001	.000	.002	-.002	.000	.001	.002	-.003	-.001	.001	.005	-.005	-.001	.001	.005	-.005
20	3000	RMSE	.252	.288	.358	.313	.313	.419	.399	.324	.278	.553	.405	.328	.255	.288	.379	.318
		Corr	.950	.932	.905	.938	.950	.932	.905	.938	.959	.917	.886	.935	.964	.959	.902	.940
		Bias	.000	.002	-.002	.029	.000	.001	.002	-.003	.001	.002	.000	.001	.001	.002	.000	.001
	6000	RMSE	.251	.289	.352	.313	.313	.418	.396	.320	.278	.566	.381	.323	.255	.288	.357	.315
		Corr	.950	.932	.905	.938	.950	.932	.905	.938	.959	.916	.893	.937	.964	.958	.908	.941

4.5.2. Evaluation of the RMSE

The factorial ANOVA was conducted on the RMSE measures for both whole group and selected group. Since the same factors were found to be significant in these two analyses, the factorial ANOVA results for the whole group were reported in this section. The results showed that the test length was the common significant factor across all four response-style classes and also was the only significant factor for the ORS, MRS, and ARS classes. The type of mixture was another significant factor for the ERS class. These ANOVA results are presented in a single table concisely in Table 44.

Table 44. *Effect size (η^2) for the RMSE of Theta Estimates Conditional on Statistical Significance ($p < 0.05$)*

Factor	ORS	ERS	MRS	ARS
mixture		0.11		
item	0.39	0.10	0.39	0.53

The test length was the significant factor on the RMSE for person trait parameters in the ORS class ($F_{(2,24)} = 99.09$, $p < .001$; $\eta^2 = 0.39$), ERS class ($F_{(2,10)} = 8.66$, $p < .001$; $\eta^2 = 0.10$), MRS class ($F_{(2,9)} = 33.05$, $p < .001$; $\eta^2 = 0.39$), and ARS class ($F_{(2,4)} = 3234.06$, $p < .001$; $\eta^2 = 0.53$). In addition to the main effect of the test length, the type of mixture was significant for the ERS class ($F_{(2,10)} = 10.24$, $p < .001$; $\eta^2 = 0.11$).

Based on the Tukey HSD $_{(FW\alpha=0.5)}$ test, the RMSE difference between $I = 4$ and $I = 10$ as well as between $I = 10$ and $I = 20$ were significant in the ORS class: M_4 (0.544) > M_{10} (0.395) > M_{20} (0.282), ERS class: M_4 (0.584) > M_{10} (0.515) > M_{20} (0.393), MRS class: M_4 (0.708) > M_{10} (0.532) > M_{20} (0.384), and ARS class: M_4 (0.601) > M_{10} (0.425) > M_{20} (0.326). For the ERS class, the main effect of the type of mixture differed from each other as following: M_{OEMA} (0.312) < M_{OEM} (0.510) = M_{OM} (0.510).

4.5.3. Evaluation of the Correlation

As was found in the factorial ANOVA on the RMSE in Section 4.5.2, the test length was the significant factor on the correlation between generated and estimated person trait parameters in the ORS class ($F_{(2,25)} = 52.36$, $p < .001$; $\eta^2 = 0.35$), ERS class ($F_{(2,10)} = 5.28$, $p < .001$; $\eta^2 = 0.20$), MRS class ($F_{(2,9)} = 166.97$, $p < .001$; $\eta^2 = 0.61$), and ARS class ($F_{(2,4)} = 3859.53$, $p < .001$; $\eta^2 = 0.72$).

Table 45. *Effect size (η^2) for the Correlation of Theta Estimates Conditional on Statistical Significance ($p < 0.05$)*

Factor	ORS	ERS	MRS	ARS
Item	0.35	0.20	0.61	0.72

Based on the Tukey HSD $_{(FW\alpha=0.5)}$ test, the correlation difference between $I = 4$ and $I = 10$ as well as between $I = 10$ and $I = 20$ were significant in the ORS class: M_4 (0.832) < M_{10} (0.911) < M_{20} (0.958), ERS class: M_4 (0.877) = M_{10} (0.901) < M_{20}

(0.935), MRS class : M_4 (0.328) < M_{10} (0.772) < M_{20} (0.899), and ARS class : M_4 (0.757) < M_{10} (0.890) < M_{20} (0.946).

4.5.4. Impact of misclassification on person trait estimation

To test the impact of the misclassification on person trait parameter recovery, the discrepancies in the RMSE and correlation measures between the whole and selected group were tested. A paired *t*-test was conducted for each latent class on the marginal discrepancies over all manipulated factors. Table 46 and Table 47 present the descriptive statistics of the RMSEs and the correlations for the whole and selected groups, respectively. The results of the paired *t*-test are presented in Table 48. The effect size was evaluated using Cohen's *d* (*d=mean difference / standard deviation of mean difference),*which indicates a small effect size if $d > 0.2$, a medium effect size if $d > 0.5$, or a large effect size if $d > 0.8$. In table 48, Cohen's *d* is presented when the mean difference is statistically significant at $p < .05$

Table 46. *Cell Means of the RMSE of theta estimates*

Type of mixture	Group	N	Mean	SD
ORS	Whole	60	0.386	0.119
	Selected	60	0.361	0.105
ERS	Whole	34	0.478	0.126
	Selected	34	0.406	0.113
MRS	Whole	31	0.495	0.136
	Selected	31	0.472	0.135
ARS	Whole	20	0.425	0.114
	Selected	20	0.428	0.116

Table 47. *Cell Means of the Correlation of Theta Estimates*

Type of mixture	Group	N	Mean	SD
ORS	Whole	61	0.911	0.060
	Selected	61	0.919	0.054
ERS	Whole	34	0.911	0.035
	Selected	34	0.923	0.039
MRS	Whole	30	0.772	0.219
	Selected	30	0.790	0.222
ARS	Whole	20	0.882	0.083
	Selected	20	0.882	0.074

Table 48. *Paired t-test Results on the Impact of Misclassification on Theta Recovery*

Type of mixture	Evaluation measures	*t*	*df*	*p*	Cohen's *d*
ORS	RMSE	4.204	59	.000	0.54
	Correlation	-4.079	60	.000	0.49
ERS	RMSE	2.838	33	.008	0.38
	Correlation	-3.522	33	.001	0.17
MRS	RMSE	2.107	30	.044	0.52
	Correlation	-3.108	29	.004	0.60
ARS	RMSE	-0.755	19	.460	
	Correlation	0.131	19	.897	

As can be seen in Table 48, the impact of misclassification was statistically significant for all response-style classes except the ARS class. The reason that the theta recovery was not impacted for the ARS class is because the classification accuracy was high. The effect size was generally medium level except for the correlation for the ERS class.

4.6. Model-based Correction of Score Bias

Figure 20 depicts the relation between sum score and estimated θ for each class of the 3-response-style mixture. The data for this figure was obtained from the equal proportions, 10-items with a sample size of $N = 6000$ condition. This figure showed how the θ estimates of the MPCM would provide a tool to correct the sum score bias due to response styles. For example, if a respondent's θ level is above the mean (i.e., $\theta > 0$) and belongs to ERS class his or her estimated θ is lower than when he or she belongs to the ORS class. Since the sum score is likely to be inflated by his or her endorsement of a higher extreme category, his or her θ should be adjusted downward to correct the inflated sum score. Conversely, if a respondent's θ level is below the mean (i.e., $\theta < 0$) and belongs to ERS class, the estimated θ is higher than when he or she belongs to the ORS class. Because he or she would have selected a lower extreme response category more often, his or her estimated θ should be adjusted upward to compensate the deflated sum score.

If a respondent with a θ level that is higher than the mean belongs to MRS class, his or her estimated θ is higher than when he or she belongs to the ORS class. His or her tendency to select middle categories would have deflated sum score, and, therefore, the correction is made to compensate his or her score lost due to the response tendency. Conversely, if a respondent's θ level is below the mean and belongs to MRS class, the estimated θ is lower than when he or she belongs to the ORS class. Because he would have selected the middle-category despite his or her

lower θ level, his or her estimated θ should be adjusted downward to correct the inflated sum score.

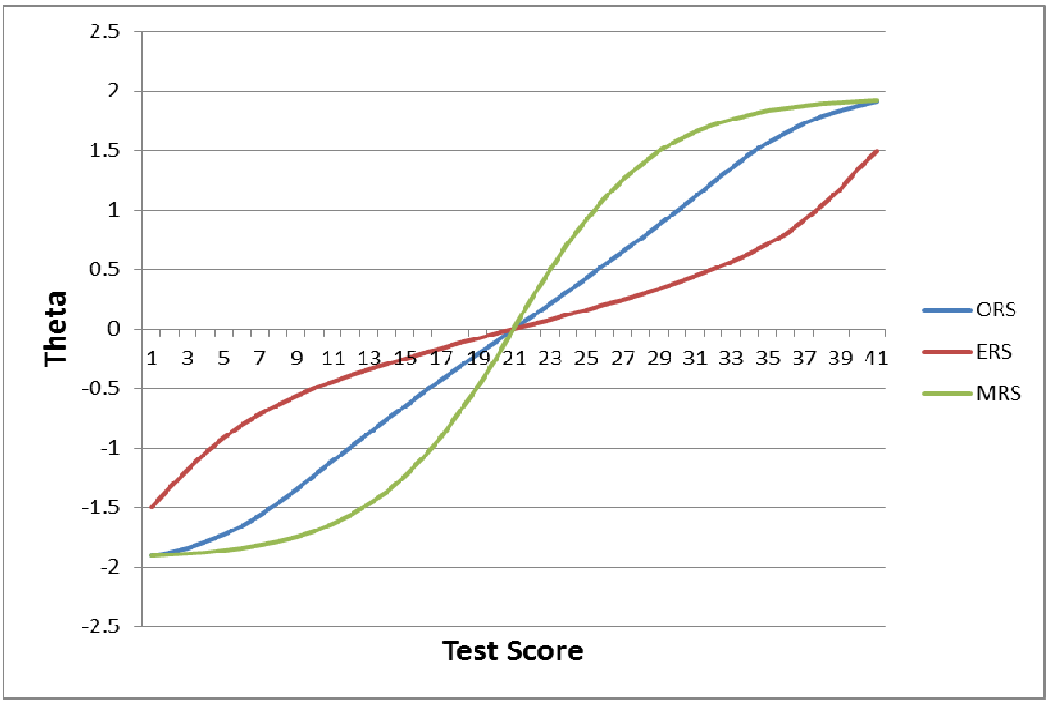

Figure 20. Theta estimates as a function of sum score for ORS, ERS, and MRS Class

Figure 21 represents the relation between sum scores and estimated θ under the 4-response-style mixture. In this figure, a function for the ARS class was added. It appears that the correction for the ARS class is very much alike the correction for the MRS class. This is understandable because the ARS responses were generated assuming a balanced scale that intended to cancel out ARS respondent's directional choice of response categories. Although ARS respondents endorse higher extreme categories only, the use of a balanced scale causes the sum scores to regress toward the mean score.

Figure 21. Theta estimates as a function of sum score for ORS, ERS, MRS, and ARS Class

The plots in which the relation between ORS-ERS, ORS-MRS, and ORS-ARS in the 2-response-style mixtures because they the relations appeared the same as those were depicted in Figures 20 and 21.

Chapter 5: Discussion

The primary goal of the current study was to investigate the performance of the mixture distribution polytomous Rasch model in accurately recovering model parameters under the heterogeneous population conditions in which people differed in their response styles, or individual tendencies in responding to the formal aspects of rating scales. The current study examined the mixture polytomous Rasch model with two, three, and up to four latent classes within each of which a different response style was manifested. One of the latent classes was simulated to represent ordinary response style (ORS), which did not manifest a distorted use of response categories of a rating scale. The rest of the latent classes were characterized by one of the following distorted response styles, i.e., extreme response style (ERS), middle category response style (MRS), and acquiescent response style (ARS).

Response styles have been recognized as a source of systematic measurement bias. Ignoring or failing to adequately account for the impact of the response styles in latent trait measurement leads to various psychometric problems such as invalidating test score differences at both individual and group levels, inflating test reliability, obscuring structural relations among psychological constructs of interest, and confounding the interpretation of the findings in comparative studies.

As a model-based approach to control for these adverse effects, mixture polytomous Rasch models, particularly the mixture partial credit model (MPCM) has been increasingly applied in empirical research where ordered polytomous item responses were analyzed. The MCPM was suggested as a method for classifying

people according to their response styles as the model was proposed by Rost (Rost, 1991). Cumulative results from previous studies have evidenced that respondents who share the ERS constitute a latent class while the other class(es) is often composed of respondents with a non-extreme response style. Once different response styles are detected within different latent classes, the subsequent analysis of a psychological construct of interest can be conducted under the control of response styles. It is promising that the application of the MCPM has potential for a better estimation of person trait as well as a better prediction of relevant criteria.

In addition, the MPCM is a flexible modeling framework in that the nature of latent classes does not need to be known a priori. “What are the types of response styles manifested in this data set?” and “which response style do most people present in this group?” are explored and answered during the course of the MPCM analysis. Although previous empirical studies have detected relatively simple structures of the mixture of response styles, i.e., mostly a combination of ERS and another style characterized as a rather moderate response style (perhaps MRS or ORS), this flexibility of the MPCM extends the potential for identifying more diverse response styles that might exist in a data set.

There is a need for a simulation study to evaluate the performance of the MPCM including accurate recovery of the model parameters, thereby assessing the soundness of the application of the MPCM to account for various types of response style effects that may be presented in real world testing situation. Little information is known thus far, however, regarding the accuracy of model parameter recovery in the

MPCM and testing conditions that can exert an influence on the model parameter recovery.

The current simulation study, therefore, focused on the evaluation of: *i*) the accuracy of recovering class membership, threshold parameters, and person trait parameters in various testing conditions, and *ii*) the model-based correction of score bias due to response styles. Of particular importance, the current study included more complex and realistic, mixture structure where multiple classes of ORS, ERS, MRS, and ARS coexist.

The following sections include a summary of the findings, discussion of the important issues surrounding interpretations of the MPCM results, recommendation for applied researchers, as well as limitations of the current study and implications for future research.

5.1 Summary of Findings

Non-convergence and boundary threshold estimates. Estimation problems were examined as a preliminary analysis of the simulation results. First, the rate of non-convergence, which may very well be indicative of problems in model identifiability and instability of parameter estimates, was obtained. This non-convergence rate was 0 % for 80 out of 90 simulation conditions when the data sets were correctly estimated with the data generation model. The other 10 simulation conditions showed the non-convergence rate ranging between 1 % and 9 %. The ORS-ERS mixture conditions never encountered non-convergence while the highest rate of non-convergence occurred under the ORS-ARS mixture conditions.

When the generated data set was under parameterized, non-convergence problems occurred only for two conditions of the 4-response-style mixtures. Conversely, when the generated data was over parameterized, thirty-five out of ninety simulation conditions showed non-convergence ranging between 1 % and 20 %. A high percentage of these problems occurred with the ORS-MRS mixtures.

Boundary threshold estimates were also monitored and screened. Extreme thresholds exceeding 9.0 or -9.0 in the provided *mdltm* outputs were filtered out. Boundary estimates never occurred when the 2-response-style data sets were under parameterized. When the generated data was correctly parameterized, the occurrence of boundary estimates was closely related to the sample size, more specifically the expected category frequencies. A high percentage of boundary estimate problems ranging between 48 % and 96 % occurred mostly for the response categories in the MRS and ARS class for which the expected response frequency was essentially zero. Nearly all of the simulation conditions presented boundary threshold estimates when the data sets were over parameterized.

As a result of checking non-convergence and boundary threshold estimates problems, ten simulation conditions were removed from the design. These excluded conditions were associated with a sample size of $N = 1200$ and the unequal mixing proportions condition (except one condition with four response styles and equal mixing proportions). These results would seem to indicate that an appearance of implausible threshold values in an empirical data analytic study may be an indication of over parameterization (i.e., estimating a model with too many latent classes) or an

insufficient sample size to estimate parameters for a given model, or a combination of the two conditions.

Label Switching. The current study tackled the label switching problems by jointly applying two different algorithmic approaches, each of which utilizes different source of information. The first algorithm developed by the author used the characteristic features of the order of four thresholds in each response style class. The second approach developed by Tueller et al. (2011) used the results of respondent classification results. By incorporating these two algorithms, the efficiency of the automated process of detecting and correcting switched labels was enhanced.

Thirteen simulation conditions turned out to have a large proportion of replications in which switched labels were unresolved. It was found that there was a great deal of overlap between the cases where switched labels were not corrected and the BIC and CAIC were unable to correctly identify the data generation model. A close investigation of this overlap allowed the researchers to better understand the hidden structures of the subpopulation distributions as well as the capabilities and limitations of the MPCM in modeling those population heterogeneities.

Model selection. Among the three information criterion statistics, AIC, BIC, and CAIC, the BIC and CAIC performed nearly equally well in identifying the data generation model with a slightly higher accuracy for the BIC. Across all of the simulation conditions, the AIC showed high rates of over-identification of the latent classes. Based on the current simulation results, the AIC should not be recommended for use in model selection under the MPCM.

In general, the BIC was found to most accurately identify the correct number of latent classes in the MPCM. Under the simulation conditions in which neither estimation problems nor unresolved label switching problems occurred, the data generation model was identified 100% of the time based on the BIC. The simulation conditions in which the BIC did not perform perfectly were associated with at least one of the following conditions: *i*) the test length $I = 4$, *ii*) the sample size $N = 1200$, and *iii*) the mixing proportions were unequal.

Classification accuracy. Generally, the ORS-ARS mixtures allowed for accurate classification under all simulation conditions while the ORS-MRS mixtures were the least accurate in providing correct classification of respondents followed by the ORS-ERS mixtures. Misclassification of ERS respondents within the MRS class (EM) and misclassification of MRS respondents within the ERS class (ME) rarely occurred. In addition to EM and ME, the chance of OA, MA, AO, and AM was also essentially zero.

The most important factor influencing respondent classification accuracy was test length. The effect size of test length was extraordinary large ($\eta^2 = 0.42$). Under the least complex, 2-response-style mixtures, when test length was $I = 4$, ORS-ERS, ORS-MRS, and ORS-ARS mixtures allowed for an average classification accuracy rate of 81%, 70 %, and 93%, respectively. As the number of items increased to $I = 10$, the average classification accuracy increased to 94%, 87%, and 98%, respectively. While for the test length, $I = 20$, it reached 98%, 94%, and almost 100%, respectively. Under the 3-response-style mixtures, as test length increased from $I = 4$ to $I = 10$ and

then from $I = 10$ to $I = 20$, the corresponding average classification accuracy improved from 73% to 87%, and then to 95%, respectively. Under the most complex, 4-response-style mixtures, classification accuracy was 89% and 95% when $I = 10$ and 20, respectively. Significant interaction effects were mainly due to the outstanding classification accuracy for the ARS class even under the $I = 4$ condition.

***Threshold recovery*.** Generally, as the sample size increased from $N = 1200$ to $N = 3000$, then to $N = 6000$, threshold recovery tended to be more accurate. While the increase in the test length from $I = 4$ to $I = 10$ improved threshold recovery significantly, the increase from $I = 10$ to $I = 20$ did not result in a significant difference. Threshold recovery for the ARS class was quite accurately achieved under even $I = 4$ condition and, consequently, the test length was not found to be an influencing factor for this class. When the distorted response styles, i.e., ERS, MRS, and ARS presented with a small proportion in a sample of respondents, the threshold recovery was significantly less accurate for those small latent class. ORS thresholds were most accurately recovered under the ORS-ARS mixtures and least accurately recovered under the ORS-ERS-MRS mixtures. Therefore, it may not be necessarily true that thresholds of a more complex model are less accurately recovered. Standard error of threshold estimates dramatically increased for the models with 3 response-style classes when the test length of $I = 4$ was considered.

***Person trait recovery*.** The factor that most affected the accuracy of θ recovery was the test length. The accuracy of θ recovery in each response-style class increased as the test length increased.

Overall, the person trait θ was well recovered when the test length was $I = 10$ or $I = 20$. A sample size of $N = 1200$ provided relatively lower correlations between generated and estimated θ parameters. Across the three levels of test length, the mean RMSE ranged from 0.28 to 0.54 for the ORS class; 0.39 to 0.58 for the ERS class; 0.38 to 0.53 for the MRS class; and 0.33 to 0.43 for the ARS class. The mean correlations ranged from 0.83 to 0.96 for the ORS class, 0.88 to 0.94 for the ERS class, and 0.77 to 0.90 for the MRS class, and 0.76 to 0.95 for the ARS class.

When the accuracy of θ recovery was computed for those who were correctly classified, there was always an increase in the level of accuracy compared to when the accuracy was computed for all respondents including misclassified cases. The discrepancies in the accuracy level between all respondent group and correctly classified respondent group were tested. The results of the paired *t*-test showed statistically significant impact of misclassification on the person trait estimation.

Correction of score bias. In an empirical rating scale data, respondent's sum scores may be biased if his or her particular response style operates while responding to the response categories. The most practical benefits of employing the MPCM is that sum scores that might have been biased due to the compounding effects of the response styles can be corrected through the class-specifically estimated θ.

The current study showed that the MPCM provides θ estimates that were corrected for the sum score bias caused by the different response styles. In general, the inflated score bias that occurred for ERS respondents with a higher θ level and for MRS and ARS respondents with a lower θ level were adjusted downward whereas the

deflated score bias occurred for ERS respondents with a lower θ and for MRS and ARS respondents with a higher θ level were adjusted upward.

5.2 Discussion

The current study showed that the model parameters of the MPCM were recovered well and that classification accuracy was reasonably relatively high. Of particular importance, rather complex mixture structure where up to four different response-style subpopulations were mixed appeared to be reasonably well modeled by the MPCM under the simulated testing conditions that were considered in this study. This observed model performance support the potential utility of this model in real world data analysis situation where there is a possibility that there exist hidden subpopulations that differ from each other with respect to response styles.

Previous empirical studies have shown the utility of this mixture modeling approach in various researches in the fields of study including personality, organizational, and clinical psychology. The latent groupings identified in those studies could be attributed to social desirability, faking, structural differences, and different response styles. By examining the thresholds plots for each estimated latent class and analyzing the contents of the items for which latent classes specifically show differences, there seems to be the potential for new findings and insights in psychological constructs that can be revealed beyond the presence of response styles.

Testing conditions and MPCM performance. The preliminary examinations of the estimation issues and label switching solutions, as well as the model selection analysis provided coherent information regarding the structure of the response-style

mixture distributions and testing conditions that allowed the MPCM to adequately deal with the response style problems.

As more profound differences in response styles were manifested across latent classes, the easier for the MCPM to detect the differences. Thus, the structural differences in the thresholds between ORS and ARS class appeared to be more easily identified than those between ORS and ERS while the differences between ORS and MRS were the most difficult to be distinguished. As the structural differences were harder to detect, the higher rates of the occurrence of boundary estimates, unresolved label switching as well as the lower rates of the correct model selection based on the BIC were observed. When the nature of the response-style mixture distribution imposed a challenge on the parameter estimation, a larger sample size and/or a larger number of test items were required for reasonable parameter estimation.

The current simulation study showed that when the test length was $I = 10$ and the sample size of $N = 3000$, the MPCM performed fairly well in recovering model parameters for the most complex 4-response-style mixtures with equal proportions. The MPCM performance shown under this nature of mixture distribution and those testing conditions are the following: *i*) the correct model selection rate based on BIC was 100%, *ii*) classification accuracies were 84%, 94%, 88%, and 95% , *iii*) the mean RMSE of the four thresholds were 0.15, 0.19, 0.25, and 0.25, *iv*) the mean correlation for the four thresholds were 0.85, 0.87, 0.89, and 0.99, v) the mean SE of the four thresholds were 0.06, 0.07, 0.10, and 0.09, *vi*) the biases of θ estimates were -0.01, -0.01. 0.00. 0.00, *vii*) the RMSEs of θ were 0.41, 0.54, 0.48, and 0.44, and *viii*) the

correlations of θ were 0.92, 0.90, 0.82, and 0.88, for the ORS, ERS, MRS, and ARS class, respectively.

Based on the findings in the current study, some recommendations are suggested for applied researchers. Regarding the common issues in measurement, 'how large should the sample size be?' and 'how many items should be asked?', 10 items with a 5-category Likert scale and the number of respondents of 3000 was found to warrant reasonably accurate parameter estimation and respondent classification when up to four different response styles among ORS, ERS, ARS, MRS were presented in a data set under equal proportions. If the data being analyzed includes less diverse types of response styles, the same level of parameter estimation and respondent classification could be achieved with less than 3000 respondents. If the relative sizes of different response-style group are unequal, more than 3000 respondents may be needed to achieve the same level of accuracy.

Comparisons of person trait across latent classes. One of the arguments that had been raised in the mixture IRT domain was whether person trait θ estimates obtained from different classes could be legitimately compared based on their magnitudes. This argument revolves around the notion that the continuous variable measured within each class is qualitatively different in mixture IRT models. As was discussed by Rost et al. (1997), the comparisons could certainly be problematic if the profiles of the item locations (i.e., the means of the thresholds) were substantially different across latent classes. This would indicate that people in different classes present different cognitive structures or psychological constructs. In these cases, since

questionnaires could not claim to be measuring the same trait in different populations, trait estimates obtained through the use of questionnaires could not be used to compare differences among respondents across the latent response-style classes.

When the item difficulties were very much the same across latent classes, however, what distinguished latent classes was the dispersion of item responses, not the difficulty of an item (Rost et al., 1997). When this condition held, the comparison of person trait across different classes could be justified because the class specific θ values only adjusts for the effects of the class-specific dispersion of responses.

In practice, item location profiles should be checked across latent classes before attempting any interpretation of latent class differences. If the item location profiles from each class locate significantly different positions, the difference across latent classes may better be characterized with respect to certain latent traits rather than response styles.

Correction of score bias and predictability. The current study showed that the MPCM provided the corrected θ estimates that clearly differentiated the effects of different response styles. Given that the model provided this alternative, "purified" score for each response style, an important issue to address is whether using the "purified" θ improves predictability of relevant criterion variable. This idea was addressed by Maij-de-Meij et al. (2008). Improved predictability is a question that awaits an answer from empirical research in various fields. The current simulation provided results that help in building a foundation upon which this practical utility of the MPCM can be further investigated among applied researchers.

5.3 Limitations of the current study and implications for future research

The current study included extreme simulation conditions with an intention to explore possible limitations of the MPCM performance. The combination of test length of $I = 4$, sample size of $N = 1200$, and unequal mixing proportions that allows only 10% of the respondents to be members of the smaller classes were highly challenging conditions to achieve good parameter estimation in the context of mixture distribution polytomous IRT modeling. While setting up these extreme conditions helped in revealing some limitations in the application of the MPCM, it caused several cell means to be unavailable, limiting the interpretations of the factorial ANOVA results regarding the effects of the manipulated factors.

The interpretations of the current results that involved the acquiescent responses should be limited to the testing situation where a well-constructed balanced scale was used. From a methodological perspective, the current simulation results were meaningful in that the aberrant response behavior could possibly be controlled through the use of a balanced scale and the MPCM. The results showed that the ARS respondents were almost perfectly differentiated from other types of respondents and received a corrected θ similar to what MRS respondents would receive. However, whether the corrected θ contains the same psychological meaning for this group of respondents is evidently a question that calls for a degree of informed judgment among experts in the content area where the psychological test results would be scrutinized.

The generated item locations within each class had small variability in the current study. In the MPCM, between-class variability not only in the order of thresholds and threshold distances but also in the item locations among test items may contribute to the recovery accuracy of the parameters (e.g., Rost, 1991). This small between-group variability in item locations might have contributed positively or negatively to the parameter recovery results of this study. In this study, polytomous item responses obtained with a 5-category Likert scale items were used. It has been previously investigated in the literature that the parameter recovery of the partial credit model differed depending on the number of categories on the rating scale that was used. The simulation results could possibly be different if different numbers of response categories were used. The effects of the variability in item locations within latent classes and the effects of different numbers of response categories warrant further studies.

Future studies can also explore the other mixture distribution IRT model than the Rasch family models. Researchers have pointed out that the equal discrimination assumption of the Rasch models can be easily violated in real data analytic situations. The extension of other polytomous IRT models to mixture distributions would have the potential for allowing researchers to have a more complete view of hidden structural differences including personality or cognitive constructs, faking and social desirability tendencies, non-invariant items, as well as response styles. Empirical studies needs to be conducted to evaluate whether trait estimates of the mixture IRT

models corrected for the confounding effects of different response styles can improve predictability of criteria variables in various social behavioral research.

Appendix A

Table A.1. *Category probabilities for individual items for ERS class*

Item	Category1	Category2	Category3	Category 4	Category 5
1	0.3734	0.1065	0.0402	0.1065	0.3734
2	0.3829	0.0880	0.0582	0.0880	0.3829
3	0.4120	0.0614	0.0532	0.0614	0.4120
4	0.4173	0.0675	0.0306	0.0675	0.4173
5	0.3956	0.0800	0.0488	0.0800	0.3956
6	0.3958	0.0914	0.0257	0.0914	0.3958
7	0.4363	0.0514	0.0247	0.0514	0.4363
8	0.4037	0.0777	0.0370	0.0777	0.4037
9	0.3727	0.1020	0.0506	0.1020	0.3727
10	0.4069	0.0785	0.0293	0.0785	0.4069
Mean	0.3997	0.0804	0.0410	0.0804	0.3997

Table A.1. *Category probabilities for individual items for MRS class*

Item	Category1	Category2	Category3	Category 4	Category 5
1	0.0254	0.0889	0.7713	0.0889	0.0254
2	0.0343	0.1118	0.7079	0.1118	0.0343
3	0.0492	0.0614	0.7788	0.0614	0.0492
4	0.0852	0.0976	0.6345	0.0976	0.0852
5	0.0519	0.0661	0.7640	0.0661	0.0519
6	0.0546	0.0752	0.7405	0.0752	0.0546
7	0.0368	0.1064	0.7136	0.1064	0.0368
8	0.0502	0.1052	0.6892	0.1052	0.0502
9	0.0602	0.1441	0.5913	0.1441	0.0602
10	0.0591	0.1088	0.6642	0.1088	0.0591
Mean	0.0507	0.0966	0.7055	0.0966	0.0507

Table A.1. *Category probabilities for individual items for ARS class*

Item	Category1	Category2	Category3	Category 4	Category 5
1	0.7136	0.1424	0.1270	0.0124	0.0046
2	0.0046	0.0124	0.1270	0.1424	0.7136
3	0.7669	0.1065	0.0758	0.0421	0.0087
4	0.0087	0.0421	0.0758	0.1065	0.7669
5	0.8269	0.1045	0.0313	0.0205	0.0170
6	0.0170	0.0205	0.0313	0.1045	0.8269
7	0.7102	0.1535	0.0906	0.0420	0.0036
8	0.0036	0.0420	0.0906	0.1535	0.7102
9	0.7338	0.2144	0.0356	0.0096	0.0066
10	0.0066	0.0096	0.0356	0.2144	0.7338
Mean	0.7503	0.1443	0.0721	0.0253	0.0081

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716-723.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 42,* 69-81.

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika, 47,* 105-113.

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modeling of responses to NEO-FFI items. *Personality and Individual Differences, 40,* 1235-1245.

Austin, E. J., Deary, I. J., Gibson, G. J., McGregor, M. J., Dent, J. B. (1998). Individual response spread in self-report scales: personality correlations and consequences. *Personality and Individual Differences, 24,* 421-438.

Bachman, J. G., & O'Mally, P. M. (1984) 'Yes-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly, 48,* 491-509.

Baker, F. B., & Kim, S. H. (2004). *Item Response Theory*: Parameter Estimation Techniques (2nd ed.). New York, Dekker.

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

Berg, I. A., & Collier, J. S. (1953). Personality and group differences in extreme response sets. *Educational Psychological Measurement, 13,* 164-169.

Billet, J. B., & McClendon, M.J. (2009). Modeling acquiescence in measurement models for two balanced sets of items. *Structural equation modeling: A multidisiplinary Journal, 7,* 608-628.

Bock, R. D. (1972). Estimaing item parameters and latent ability when responses are scored in two or more nominal categories, *Psychometrika, 37,* 29-51.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39,* 331-348.

Bolt, D. M., Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement, 33,* 335-352.

Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement, 7,* 814-833.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and tis analytical extensions. *Psychometrika, 52,* 345-370.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153-168.

Brown, S. W., Garry, M., Silver, B., & Loftus, E. (1997). *Conceptions and misconceptions of what and how we remember: Survey results.* Paper presented

at the annual conference of the American Psychological Society, Washington, DC.

Buckley, J. (2009). *Cross-national response styles in international educational assessment: Evidence from PISA 2006.* NCES Conference on the Program for International Student Assessment: What we can learn from PISA. (Downloaded from http://edsurvey.rti.org/PISA/ on January 12, 2012).

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 Sciences, 6,* 170-175.

Cheung, G. W., Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology, 31,* 187-212.

Cho, Y., Jiao, H., & Macready, G. (2012a*). Assessing the effects of different item parameter profiles in mixture Rasch models*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, Canada.

Cho, Y., Jiao, H., & Macready, G. (2012b). *Simultaneous effects of different item discrimination profiles and item difficulty profiles in mixture 2PL models.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Chronbach, L. J. (1946). Response set and test validity. *Educational and Psychological Measurement, 6,* 475-494.

Clarke III, I. (2000). Extreme response style in cross-cultural research: An empirical investigation. *Journal of Social Behavior and Personality, 15,* 137-152.

Cohen, J. (1988). Statistical power analysis for the behavior sciences (2nd ed.) Routledge.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory and NEO Five Factor Inventory.* Professional Manual, Odessa, Florida: Psychological Assessment Resources Inc.

Couch, A., & Kenison, K. (1960). Yeasayers and Naysayers: Agreeing response set as a personality variable. *Journal of Abnormal Social Psychology*, *60,* 151-174.

De Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: Guilford Press.

De Jong, M. G, Steenkamp, J.-B. E. M., 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.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, ser. B, 39,* 1-38.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49,* 71-75.

Eid, M., Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment, 16,* 20-30.

Egberink, I. J. L., Meijer, R. R., Veldkamp, B.P. (2010). Conscientiousness in the workplace: Applying mixture IRT to investigate scalability and predictive validity. *Journal of Research in Personality, 44*, 232-244.

Embretson, S. E., & Reise, S. P. (2000). *Item reponse theory for psychologist.* Mahwah, NJ: Lawrence Erlbaum Associates.

Fahrenberg, J., Hampel, R., & Selg, H. (1989). *Das Freiburger Persönalichkeitsinventar FPI* [Freiburg Personaliry Inventory FPI] (5th ed.). Göttingern, Germany: Hogrefe.

Gollwitzer, M., Eid, M., Jürgensen, R. (2005). Response styles in the assessment of anger expression. *Psychological Assessment, 17,* 56-69.

Greenleaf, E. A. (1992). Measuring extreme response style. *Public Opinon Quarterly,* 56, 328-351.

Hofstede, G. (1980). *Culture's Consequences. International Differences in Work-Related Values.* London: SAGE Publications.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style, *Journal of Cross-Cultural Psychology,* 20, 296-309.

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology, 36*, 264–277.

Kulas, J. T., & Stachowski, A. A. (2008). Moddle category endorsement in odd-numbered Likert response scales: Associated item characteristics, cognitive

demands, and preferred meanings. *Journal of Research in Personality, 43*, 489-493.

Harzing, A.-W. (2006). Response styles in cross-naitonal survey research: A 26-contry study. *International Journal of Cross Cultural Psychology, 20*, 296-309.

Lau, A. (2009*). Using a mixture IRT model to improve parameter estimates when some examinees are amotivated* (Unpublished doctoral dissertaion). James Madison Univeristy, Harrisonburg, VA.

Lazarsfeld, P. F., & Henry, N. W. (1968). Latent structure analysis. Boston, MA: Houghton Mifflin.

Lee, C., & Green, R. T. (1991). Cross-cultural examination of the fishbein behavioral intentions model. *Journal of International Business Studies, 25,* 289-305.

Lee, J. W., Jones, P. S., Meneyama, Y., & Zhang, X.E. (2002). Cultual difference in responses to a Likert scale. *Research in Nursing & Health, 25*, 295-306.

Lewis, N. A., & Taylor, J. A. (1955). Anxiety and extreme response preference. *Educational and Psychological Measurement, 15,* 111-116.

Li, F., Cohen, A. S., Kim, S-H, & Cho, S-J. (2009). Model selection methods for mixture dichotomour IRT models. *Applied Psychological Measurement, 33* (5), 353-373.

Likert, R. (1932). A technique for the measurement of attitude. *Archives of Psychology, 140,* 1-55.

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

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2005). Latent-trait latent-class analysis of self-disclosure in the work environment. *Multivariate Behavioral research, 40*, 435-459.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnair data: characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement, 32*, 611-631.

Marin, G., Gamba, R. J., & Marin, B. V. (1992). Extreme response style and acquiescence among Hispanic: The role of acculturation and education. *Journal of Cross-cultural Psychology*, *23,* 498-509.

Masters, G. N. (1984). A Rasch model for partial credit scoring. *Psychometrika,* 47, 149-174.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Meiser, T., & Machunski, M. (2008). The personal structure of personal need for structure: A mixture-distribution Rasch analysis. *European Journal of Psychological Assessment, 24*, 27-34.

Messick. S. (1991). Psychology and methodology of response styles. In Snow & Wiley (Eds.), *Improving Inquiry in Social Science: A Volume in Honor of Lee J. Chronbach*. Hillsdale, New Jersey: Lawrence Erlbaum Association. 161-

200.

Mirowsky, J., & Ross, C. E. (1991). Elimination defense and agreement bias from measures of the sense of control: A 2×2 index. *Social Psychology Quarterly, 54,* 127-145.

Mislevy, R. J., & Verhelst, N. D. (1990). Modeling item responses when differentsubjects employ different solution strategies. *Psychometrika, 55,* 195-215.

Molenaar, I. W. (1995). Some background for item response theory and the Rasch model. In G.H. Fischer & I.W. Molenaar (Eds). *Rasch models: Foundations, recent development, and application (pp.3-14).* New York: Springer.

Moors, G. (2003). Diagnosing response style behavior by means of a latent-class factor approach. Socio-demographic correlates of ethnic discrimination reexamined. *Quality & Quantity, 37,* 277-302.

Moors, G. (2004). Facts and artefacts in the comparison of attitudes among ethnic minorities. A multigroup latent class structure model with adjustment for response style behavior. *European Sociological Review, 20,* 303-320.

Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality & Quantity, 42,* 779-794.

Norman, R. P. (1969). Extreme response tendency as function of emotional adjustment and stimulus ambiguity. *Journal of Counseling and Clinical Psychology, 33,* 406-410.

Nunally, J. C. (1978). *Psychometric Theory.* 2nd Ed. New York: McCraw-Hill.

Paulhus, D. L. (1991). Measurement and control of response bias. In Robinson, Shaver, & Wright. (Eds). *Measures of Personality and Social Psychological Attitudes* (17-59). San Diego, CA: Academic Press.

Preinerstorfer, D., & Formann, A. K. (2012). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology, 65,* 251-262.

Rasch, G. (1960/80*). Probablistic models for some intelligence and attainment test.* (Copenhagen, Danish Institute for Educational Research), expandid edition (1980) with forward and afterward by B.D. Wright. Chicago: The university of Chicago Press.

Reise, S. P., & Gomel, J. N. (1995). Modeling qualitative variation within latent trait dimensions: Application of mixed-measurement to personality assessment. *Multivariate Behavioral research, 30,* 341-358.

Ross, C. E., & Mirowsky, J. (1984). Socially desirable responses and acquiescence in a cross cultural survey of mental health. *Journal of Health and Social Behavior, 25,* 189-197.

Rost, J. (1988). Measuring attitudes with a threshold model drawing on a traditional scaling concept. *Applied Psychological Measurement, 12,* 397-409.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14,* 271-282.

Rost, J. (1991). A logistic mixture distribution model for polytomous item responses. *British Journal of Mathematical and Statistical Psychology, 44,* 75-92.

Rost, J. (1997). Logistic Mixture Models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 449-463). New York, Springer.

Rost, J., Carstensen, C., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds). *Application of latent trait and latent class models in the social sciences* (pp. 324-332). Münster, Germany: Waxmann.

Schmitt, M. J., & Ryan, A. M. (1993). The big five in personnel selection: Factor structure in applicant and non-applicant populations. *Journal of Applied Psychology, 78,* 966-974.

Schwarz, G. (1978). Estimating the dimension of a model. *Analysis of Statistics, 6,* 461-464.

Smith, E. V., Ying, Y., & Brown, S. W. (2012). Using the mixed Rasch model to analyze data from the beliefs and attitudes about memory survey. *Journal of Applied Measurement, 13,* 23-40.

Subedi, D. R. (2010). *Investigating unobserved heterogeneity using item response theory mixture models* (Unpublished doctoral dissertation). Michigan State University, Lansing, MI.

Lau, A. (2009*). Using a mixture IRT model to improve parameter estimates when some examinees are amotivated* (Unpublished doctoral dissertaion). James Madison Univeristy, Harrisonburg, VA.

Spielberger, C. D. (1988). *STAXI. State-Trait Anger Expression Inventory.*Tempa, FL: Psychological Assessment Resources.

Temple, D. E., & Geisinger, K. F. (1990). Response latency to computer-administered inventory items as an indicatory of emotional arousal. *Journal of Personality Assessment, 54,* 289-297.

Van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response styles in rating scales: Evidence of method bias in data from 6 EU countries. *Journal of Cross-Cultural Psychology 35,* 346-360.

von Davier, M. (2000). WINMIRA 2001 [Computer software]. St. Paul, MN: Assessment Systems.

von Davier, M. (2005a). *mdltm: Software for the general diagnostic model and for estimating mixture of multidimensional discrete latent traits models* [Computer software]. Princeton, NJ: Educational Testing Service.

von Davier, M. (2005b). A general diagnostic model applied to language testing data (ETS Research Report No. PR-05-16). Princeton, NJ: Educational Testing Service.

von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In G.H. Fischer & I.W. Molenaar (Eds). *Rasch models: Foundations, recent development, and application (pp.371-379).* New York: Springer.

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,* 52-88.

Wu, P-C , & Huang, T-W. (2010). Person heterogeneity of the BDI-II-C and its effects on dimensionality and construct validity: Using mixture item response models. *Measurement and Evaluation in Counseling and Development, 43,* 155-167.

Yamamoto, K. Y. (1987). *A model that combines IRT and latent class models.* Unpublished doctoral dissertation, University of Illinois Urbana-Champaign.

Yang, Y., Harkness, J. A., Chin, T-Y., & Villar, A. (2010). Response styles and culture. In J. A., Harkness et al. (Eds). *Survey Methods in Multinational, Multiregional, and Multicultural Contexts.* John Wiley & Sons, Inc.

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods, 7,* 168-190.

Zickar, M. J., & Robie, C. (1999). Modeling faking food on personality items: An item-level analysis. *Journal of Applied Psychology, 84,* 551-563.