Psychological Test and Assessment Modeling, Volume 61, 2019 (1), 91-123

A specialized confirmatory mixture IRT modeling approach for multidimensional tests

Minjeong Jeon[1]

Abstract

Finite-mixture models are typically utilized in educational and psychological research to explore potential latent classes that may be present in the data under investigation. However, mixture models can also be applied to test out or confirm researchers' theories or hypotheses about latent classes. In this paper, we discuss a specialized confirmatory mixture IRT modeling approach for multidimensional tests with a set of pre-arranged constraints on item parameters that are devised to differentiate latent classes. Two types of multidimensional classification scenarios are discussed: (1) a single membership case where subjects strictly have one latent class membership for all test dimensions, and (2) a mixed membership case where subjects are allowed to have different latent class memberships across test dimensions. We illustrate maximum likelihood estimation of the two types of confirmatory mixture models with an empirical dataset.

Keywords: Mixture IRT modeling, confirmatory approach, multiple dimensions, single membership, mixed membership, saltus modeling

[1]*Correspondence concerning this article should be addressed to:* Minjeong Jeon, PhD, Department of Education, University of California, Los Angeles, 3141 Moore Hall, 457 Portola Avenue, Los Angeles CA, 90024, USA; email: mjjeon@ucla.edu

Finite mixture IRT models

Mixture item response theory (IRT) models have become a popular tool to investigate various issues in educational and psychological assessment (Bolt et al., 2001; Cohen & Bolt, 2005; Finch & French, 2012). Mixture IRT models postulate that subjects are drawn from two or more unknown (or latent) populations that present systematic differences in their item response behavior. Hence, mixture IRT models are usually utilized to identify sub-populations of subjects whose differences are characterized or captured based on differences in their item parameters.

Mixture IRT models are usually employed in an exploratory fashion because the number and nature of latent classes are unknown a priori; therefore, users of mixture IRT models aim to *explore* the possible presence of latent classes in their data. Although relatively less common compared with an exploratory approach, there have been confirmatory uses of mixture IRT modeling; in this case, the number and character of latent classes are pre-specified by researchers based on their theory or hypothesis about data. Therefore, researchers aim to *confirm* the presence and characteristics of the hypothesized latent classes by applying the mixture model. For example, educational researchers have applied a confirmatory mixture IRT model to investigate two different types of item solving strategies that examinees may apply during speeded or non-speed tests (e.g., guessing-based and ability-based strategies as latent classes) (e.g., Mislevy & Verhelst, 1990; Schnipke & Scrams, 1997; Yamamoto & Everson, 1997; Boughton & Yamamoto, 2007). Molenaar et al. (2016) hypothesized two modes of intelligence as latent classes based on differences in response times (i.e., slow and fast modes of intelligence) and investigated how examinees apply different types of intelligence during tests. Tijmstra et al. (in press) assumed and analyzed two kinds of response styles that respondents may apply when responding to Likert-type rating scale items with confirmatory mixture modeling. Jin et al. (2018) also applied a similar approach to rating-scale data to differentiate an inattentive response behavior from normal response behavior.

A specialized confirmatory mixture IRT model

Another use of confirmatory mixture modeling is found in psychometrics literature (Wilson, 1989; Mislevy & Wilson, 1996; Draney, 2007; Draney & Wilson, 2008). This, the so-called *Saltus modeling* is unique in the sense that a special set of test items are utilized to differentiate hypothesized latent classes. For instance, Wilson (1989) proposed imposing a set of constraints on the item parameters of a confirmatory mixture Rasch model to examine the developmental stages of children based on their performance on particular item sets of a cognitive test.

One may believe that such a confirmatory use of latent classes and test items is somewhat restrictive. However, in confirmatory factor analysis, which is a common practice in applied research, we typically assume that the number of factors and a factor-item relationship (or a factor structure) are known prior to data analysis. The goal of confirmatory factor analysis is to validate a factor structure that researchers hypothesize

and further examine relationship between factors. This goal is clearly different from exploratory factor analysis that aims to identify an unknown factor structure. Similarly, we argue that it would be reasonable to adopt for a confirmatory approach for latent classes and test items in mixture IRT modeling when researchers wish to corroborate a hypothesis on the number and nature of latent classes and further examine relationships between latent classes. For instance, suppose we have a behavior checklist that contains a set of items that are designed to identify patients with severe depressive symptoms. In this case, we are interested in differentiating patients with severe symptoms from those with mild symptoms (i.e., two latent classes). In addition, it would be reasonable and suitable to utilize those particular check-list items that are designed to distinguish extreme depression symptoms from mild symptoms. Hence, a confirmatory mixture IRT model can be adopted in this situation for differentiating severely depressive patients who need special care and treatments from regular patients.

Purpose

The purpose of this study is to introduce the specialized confirmatory mixture IRT modeling and describe its extension and application for multidimensional tests. Educational and psychological tests are often composed of multiple sub-tests that measure multiple constructs that are related to each other. For example, the mathematics anxiety rating scale (Richardson & Suinn, 1972) is composed of multiple sub-tests that measure situation-specific anxiety factors: (a) anxiety about performing mathematical calculations, (b) anxiety about solving a math problem in public, and (c) anxiety about taking a math test (Lukowski et al., in press). In addition, in the Trends in International Mathematics and Science Study (TIMSS), a well-known, large-scale international educational assessment, mathematics tests are based on multiple dimensions based on three cognitive domains (knowing, applying, and reasoning) as well as three cognitive domains (numbers, geometric shapes and measures, and data display) of mathematics skillsets. Hence, for the purpose of expanding the scope of the discussed confirmatory mixture IRT model's applications, it would be beneficial to consider a multidimensional extension of the model.

To analyze multidimensional assessment data for classification, one may think of a situation where subjects have the same latent class membership for different test dimensions. In this case, it is possible to predict a subject's class membership for one dimension based on her class membership for another dimension. There may be another situation, however, where subjects have different class memberships in different dimensions of the test. For instance, suppose we have a reading test that consists of two sub-tests (that measure vocabulary and comprehension, for instance) and we are interested in classifying examinees into two latent classes that indicate mastery and non-mastery of the skillset that each sub-test intends to measure. Even though the two sub-traits are likely to be positively correlated, it is still possible that some examinees who master one skillset (e.g., vocabulary) do not master the other skillset (e.g., comprehension); in this case, those examinees have different classification memberships for

the two sub-tests (i.e., the mastery class for vocabulary and the non-mastery class for comprehension). Therefore, it would be useful to think about a more general classification scenario where examinees are allowed to have different class memberships across multiple test dimensions. For convenience, we label the first classification type as single membership and the second type as mixed membership classification.

Note that one may consider a type of classification where a single class membership is assigned to subjects in a multidimensional trait space. In this case, a latent class may be characterized with a lower score in one dimension but a higher score in the other dimension (this is likely to be the case if the two dimensions are negatively correlated). Although such a classification method is reasonable, we discuss a different type of classification where class membership is assigned to subjects for each test dimension at a time. Note that this latter type of classification is typical in diagnostic classification modeling (DCM; Rupp et al., 2010) where examinees are classified into one of two classes (e.g., mastery or non-mastery as discussed above) for each of the multiple attributes that are measured with a test. In fact, mixed membership classification (or assigning class membership per dimension at a time) that we discuss in this paper can be seen as a special case of the single classification of subjects into a multidimensional trait space with an increased number of latent classes. This point will be re-visited and discussed later in the discussion section.

In this study, we discuss both single and mixed membership scenarios for a multidimensional extension of the specialized confirmatory mixture IRT model. Although both single and membership classifications have been utilized in the mixture IRT modeling literature (e.g., De Jong & Steenkamp, 2010; Choi & Wilson, 2015; H.-Y. Huang, 2016; Molenaar et al., 2016), the two classification types have rarely been discussed jointly and/or compared in the context of confirmatory mixture analysis.

Model

We first lay out the formulation of the specialized confirmatory mixture IRT model for a unidimensional test. Subsequently, we describe a multidimensional extension of the model in the case of single membership and mixed membership classification, respectively. For all models, we focus on the one-parameter logistic (1PL) parameterization for the sake of simplicity. Extensions to a two-parameter formulation are feasible as illustrated below for a unidimensional case.

Specialized confirmatory mixture IRT model

Denote y_{ij} a binary response to item i for person j and $C_j = g$ is a categorical latent variable that indicates person j's class membership g $(= 1, ..., G)$. A standard exploratory, mixture Rasch model can then be written as follows:

$$logit(\Pr(y_{ij} = 1 | \theta_{jg}, C_j = g)) = \theta_{jg} - \beta_{ig}, \tag{1}$$

where θ_{jg} is a continuous latent variable that indicates person j's latent trait in class g with class-specific mean and variance, $\theta_{jg} \sim N(\mu_g, \sigma_g^2)$ where $\mu_g = 0$ is set for scale determinacy and model identification. The item parameter β_{ig} represents the class-specific difficulty for item i for latent class g. Note that the class-specific item parameters are freely estimated in all latent classes, implying that no structure is imposed in the item parameters. In addition, the number of latent classes (G) is unknown a priori for ordinary exploratory mixture analysis; hence, it needs to be empirically determined based on data analysis.

Suppose a researcher has a strong theory or hypothesis about the number and nature of latent classes for collected data and additionally has identified a particular set of items (or 'item groups') that are designed to differentiate subjects across the latent classes. Let us further illustrate how 'item groups' can be utilized in such a scenario. Suppose a researcher wants to classify children into one of two developmental stages based on the children's scores on a reasoning test. The test is developed based on a design factor that characterizes the complexity or cognitive demand of the test items. According to the test design, individual items belong to one of two item groups: Concrete or Abstract/Counterfactual items (in the order of less to more cognitively demanding items). The researcher hypothesizes that children only in the higher developmental stage are able to solve the more complex, Abstract/Counterfactual items correctly, whereas children in the lower developmental stages can solve only the less complex, Concrete items. In this scenario, it is sensible to use the item groups (Concrete vs. Abstract/Counterfactual items) to differentiate children in the higher developmental stage from those in the lower stage. Note that the specialized confirmatory mixture analysis is more optimal in this case than standard exploratory mixture analysis because the number and nature of latent classes are already known and the latent classes (two developmental stages) would be sufficiently characterized and differentiated based on children's performance on the two item groups (Concrete or Abstract/Counterfactual items).

Suppose we have G latent classes whose characteristics are already known and we want to use H item groups to differentiate the latent classes. We assume $G = H$ for the sake of simplicity. Then model (1) can be revised as follows:

$$logit(\Pr(y_{ij} = 1 | \theta_{jg}, C_j = g)) = \theta_{jg} - \beta_i + \tau_{gh} I_{ih}. \quad (2)$$

Note that in this formulation the item difficulty parameter β_i is set equivalent across all latent classes (hence, β_i does not include subscript g). Importantly, Equation (2) includes an additional structural parameter τ_{gh} that indicates the effect of item group I_{ih} ($h = 1, ..., G$) on the probability of correctly solving the items in the item group for subjects in class g. The structural parameter τ_{gh} represents how difficult or easy the items in item group h are for subjects in latent class g compared with subjects in the reference latent class. In other words, τ_{gh} can also be interpreted as the amount of advantage that subjects in latent class g have in solving the items in item group h. For the reference item group, it is assumed that there is no performance difference between

latent classes.

To further illustrate this specification, Figure 1 is provided to display the item response function (that represents the probability of a correct response as a function of latent trait θ) for two latent classes and for two item groups (the first latent class and the first item group are set to the reference groups and all items in an item group are assumed to have the same difficulty level for the sake of simplicity). This figure shows that for item group 1 which is the reference item group (left panel), the item difficulty level is equal between the two latent classes. For item group 2 (right panel), however, the items are more difficult for latent class 2 than for latent class 1 (which is the reference person group). The difference between the two class-specific curves is captured with the τ_{22} parameter, which also represents the difference in the difficulty level of item group 2 between the two latent classes (or the amount of disadvantage class 2 has in solving item group 2 compared with class 1 subjects).

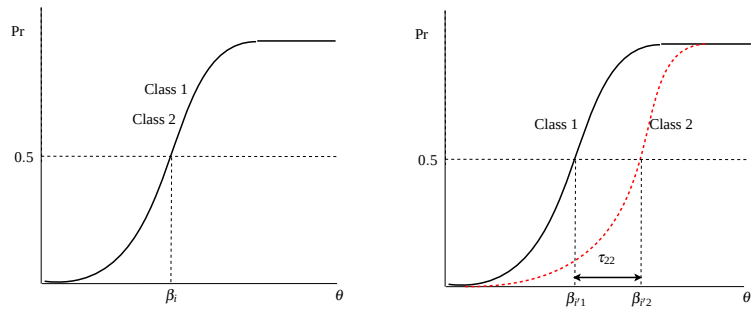

Figure 1:
Item response curves of two latent classes for item group 1 (left) and for item group 2 (right). β_i represents the item difficulty for item group 1 and $\beta_{i'g}$ ($g = 1, 2$) represents the item difficulty for item group 2 for latent class g. τ_{22} represents the difference in the item difficulty for item group 2 between latent class 1 and latent class 2.

Note that all τ_{gh} parameters in Equation (2) can be specified as a $G \times H$ matrix. For instance, when $G = 2$ and $G = 3$, the respective structural parameter matrices can be specified as follows:

$$\begin{bmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \tau_{11} & \tau_{12} & \tau_{12} \\ \tau_{21} & \tau_{22} & \tau_{23} \\ \tau_{31} & \tau_{32} & \tau_{33} \end{bmatrix}. \tag{3}$$

For identification of the model parameters, a set of constraints needs to be imposed on these matrices (Wilson, 1989; Mislevy & Wilson, 1996), such that for the reference latent class g', we set $\tau_{g'h} = 0$ for all h (or $\tau_{g'.} = 0$) and for the reference item group h', we set $\tau_{gh'} = 0$ for all g (or $\tau_{.h'} = 0$). These constraints indicate that we assume

for the reference latent class, any item group does not show a higher or lower difficulty level and for the reference item group, there are no performance differences between latent classes. For instance, suppose the first latent class and the first item group are selected as reference groups; we then set $\tau_{1\cdot} = 0$ and $\tau_{\cdot 1} = 0$. Specifically, for $G = 2$, we impose $\tau_{11} = \tau_{12} = \tau_{21} = 0$. The only to-be-estimated parameter in this case is τ_{22} which represents the amount of advantage (or disadvantage) that people in class 2 have in solving item group 2 compared with people in latent class 1. For $G = 3$, we impose $\tau_{11} = \tau_{12} = \tau_{13} = 0$ and $\tau_{21} = \tau_{32} = 0$. In this case, the structural parameters to be estimated are $\tau_{22}, \tau_{23}, \tau_{32}$, and τ_{33}. Here τ_{22} indicates the amount of (dis)advantage that subjects in class 2 have in solving item group 2 compared with people in latent class 1, τ_{32} indicates the amount of (dis)advantage that subjects in class 3 have in solving item group 2 compared with people in latent class 1, and τ_{33} indicates the amount of (dis)advantage that subjects in class 3 have in solving item group 3 compared with people in latent class 1.

Remarks on structural parameters The structural parameter τ_{gh} represents the difference in item difficulty for a particular item group (group h) between the reference latent class and latent class g. To see this more clearly, let us provide an additional illustration (Jeon, 2018). In the simplest case with $G = H = 2$, the respective models for latent classes 1 and 2 can be written as follows:

$$
\begin{aligned}
logit(\Pr(y_{ij} = 1|\theta_{j1}, C_j = 1)) &= \theta_{j1} \underbrace{-\beta_i + \tau_{11} I_{i1}}_{=-\beta^*_{i1}}, \\
logit(\Pr(y_{ij} = 1|\theta_{j2}, C_j = 2)) &= \theta_{j2} \underbrace{-\beta_i + \tau_{11} I_{i1} + \tau_{22} I_{i2}}_{=-\beta^*_{i2}}.
\end{aligned}
$$

When the first item group and the first latent class are used as reference groups, $\tau_{11} = 0$ is imposed for identification; we then have $\beta^*_{i1} = \beta_i$ when $g = 1$ and $\beta^*_{i2} = \beta_i - \tau_{22} I_{i2}$ with $g = 2$. Hence, $\tau_{22} = \beta^*_{i1} - \beta^*_{i2}$. This shows that the structural parameter τ_{22} is equivalent to the difference between the item difficulty of latent class 1 (β^*_{i1}) and the item difficulty of latent class 1 (β^*_{i2}) for item i that belongs to item group 2 ($h = 2$). Therefore, if τ_{22} is positive and significant, it means that the group 2 items are relatively easier for class 2 subjects than class 1 subjects. In other words, class 2 subjects have advantages in solving the items in item group 2 compared with class 1 subjects.

Now let us consider a more complex scenario with three latent classes ($G = H = 3$). In

this case, we can specify three models for latent classes 1, 2 and 3 as follows:

$$
\begin{aligned}
logit(\Pr(y_{ij}=1|\theta_{j1},C_j=1)) &= \theta_{j1}\underbrace{-\beta_i+\tau_{11}I_{i1}+\tau_{12}I_{i2}+\tau_{13}I_{i3}}_{=-\beta^*_{i1}},\\
logit(\Pr(y_{ij}=1|\theta_{j2},C_j=2)) &= \theta_{j2}\underbrace{-\beta_i+\tau_{21}I_{i1}+\tau_{22}I_{i2}+\tau_{23}I_{i3}}_{=-\beta^*_{i2}},\\
logit(\Pr(y_{ij}=1|\theta_{j3},C_j=3)) &= \theta_{j3}\underbrace{-\beta_i+\tau_{31}I_{i1}+\tau_{32}I_{i2}+\tau_{33}I_{i3}}_{=-\beta^*_{i3}}.
\end{aligned}
$$

When the first item group and the first latent class are used as reference groups, $\tau_{11}=\tau_{12}=\tau_{13}=0$ and $\tau_{21}=\tau_{31}=0$ are imposed for model identification; we then have $\beta^*_{i1}=\beta_i$ when $g=1$, $\beta^*_{i2}=\beta_i-\tau_{22}-\tau_{23}$ when $g=2$, and $\beta^*_{i3}=\beta_i-\tau_{32}-\tau_{33}$ when $g=3$. Therefore, for latent class 2 we can define two structural parameters as $\tau_{22}=\beta^*_{i[2]1}-\beta^*_{i[2]2}$, for item group 2 and $\tau_{23}=\beta^*_{i[3]1}-\beta^*_{i[3]2}$, for item group 3.[1] These parameters represent the amount of advantage that subjects in latent class 2 have in solving item group 2 and item group 3, respectively, compared with subjects in latent class 1. Similarly, we can define two structural parameters for latent class 3 as $\tau_{32}=\beta^*_{i[2]1}-\beta^*_{i[2]3}$ for item group 2, and $\tau_{33}=\beta^*_{i[3]1}-\beta^*_{i[3]3}$ and for item group 3, which present the amount of advantage that subjects in latent class 3 have in solving item group 2 and item group 3, respectively, compared with subjects in latent class 1.[2]

Two-parameter formulation For a two-parameter extension, an additional item discrimination parameter can be added to the specialized confirmatory mixture IRT model (Jeon, 2018). For instance, model (2) can be re-written as:

$$
logit(\Pr(y_{ij}=1|\theta_{jg},C_j=g))=\alpha_{ig}\theta_{jg}-\beta_i+\tau_{gh}I_{ih}. \tag{4}
$$

where α_{ig} is the discrimination parameter for item i for latent class g. To further simplify the model, one can set the item discrimination parameters to be equal across latent classes by replacing α_{ig} with α_i. Note that with an inclusion of the item discrimination parameters, we need to set the variances of the θ_{jg} distributions to 1 per latent class g.

Suppose one can hypothesize a structured difference in the items' discriminating power for a particular item group between latent classes. In this case, an additional structural

[1] To illustrate the three latent class cases, we additionally introduced subscript $[k]$ for $\beta^*_{i[k]h}$ to represent the i-th item's group membership ($k=1,...,H$).

[2] To compare the performance between latent class 2 and latent class 3, an additional data analysis needs to be done with re-defined structural parameters (with latent class 2 or latent class 3 as the reference group). This is similar to how dummy coding is used when a categorical covariate is utilized in regression analysis.

parameter can be introduced to the item discrimination parameters. Equation (4) can then be re-written as

$$logit(\Pr(y_{ij}=1|\theta_{jg},C_j=g))=(\alpha_i+\tau_{gh}^{(\alpha)})\theta_{jh}-\beta_i+\tau_{hk}^{(\beta)}I_{ih}, \quad (5)$$

where the new structural parameter $\tau_{gh}^{(\alpha)}$ represents a systematic difference in the discrimination power of item group h between subjects in latent class g and the reference latent class. For model identifiability, we need to impose a set of constraints to the new structural parameters as $\tau_{1.}^{(\alpha)}=\tau_{.1}^{(\alpha)}=0$ in addition to the usual constraints $\tau_{1.}^{(\beta)}=\tau_{.1}^{(\beta)}=0$ (assuming that the first latent class and the first item group are the reference groups).

Multidimensional extension

For a multidimensional extension of the specialized mixture IRT model, we consider two types of classification scenarios where: (1) subjects are classified into a single latent class across all test dimensions; and (2) subjects are classified into multiple latent classes across dimensions. We first formulate a single membership model that suits the first case scenario and then formulate a mixed membership model for the second case scenario.

Single membership model

Let us first consider the situation where subjects are classified into one of G latent classes across K dimensions of a test. A number of multidimensional mixture IRT models have been proposed based on such single membership classification (e.g., De Jong & Steenkamp, 2010; Finch & Finch, 2013; Choi & Wilson, 2015; H.-Y. Huang, 2016). For an illustration of single membership classification, see Table 1 for a two-dimensional test with two latent classes. Note that only two possible places (in the diagonal) are available in this case for subjects to be classified into.

Table 1:
Single membership classification for a two-dimensional test with two latent classes

Dim1 \ Dim2	class1	class2
class1	[1]	-
class2	-	[2]

Suppose we have H item groups to differentiate latent classes in each dimension of the

test.[3] Model (2) can be extended for single membership classification as follows:

$$logit(\Pr(y_{ij}=1|\boldsymbol{\theta}_{jg},C_j=g))=\sum_{k=1}^{K} r_{ik}\theta_{jkg}-\beta_i+\sum_{k=1}^{K} r_{ik}\tau_{ghk}I_{ihk}, \tag{6}$$

where $C_j = g$ indicates subject j's class membership ($g = 1,...,G$) for the test. r_{ik} indicates the (i,k) element (that takes value 0 or 1) of the $I \times K$ score matrix R and denotes whether the ith item is an indicator for the kth dimension. For instance, suppose we have two items in each of two test dimensions; then the score matrix R can be written as:

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

This matrix indicates that the first two items belong to the first dimension and the last two items belong to the second dimension.

The latent traits for subject j in class g are a vector, $\boldsymbol{\theta}_{jg} = (\theta_{j1g},...,\theta_{jKg})'$ and assumed to follow a multivariate normal distribution with mean vector μ_g and covariance matrix Σ_g. For the identification of this multidimensional model, an additional set of constraints needs to be imposed; that is, the mean of a reference latent class is fixed at 0 in all dimensions to set the reference points of the latent traits. For instance, when the first latent class is the reference class for a two-dimensional test, we set $\mu_{g=1} = (0,0)'$, while the means of the other latent classes are freely estimated.

The structural parameters τ_{ghk} are now defined for dimension k. To explain the dimension-specific structural parameters, let us consider a two-dimensional test where subjects are classified into one of two latent classes ($G = 2$) based on two item groups ($H = 2$). Equation (6) can then be expressed as follows:

$$logit(\Pr(y_{ij}=1|\boldsymbol{\theta}_{jg},C_j=g))=r_{i1}\theta_{j1g}+r_{i2}\theta_{j2g}-\beta_i+r_{i1}\tau_{gh1}I_{ih1}+r_{i2}\tau_{gh2}I_{ih2}. \tag{7}$$

Note that we have two sets of structural parameters τ_{gh1} and τ_{gh2} for dimension 1 and dimension 2, respectively, each of which is in the two 2×2 matrix form:

$$\begin{bmatrix} \tau_{111} & \tau_{121} \\ \tau_{211} & \tau_{221} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \tau_{112} & \tau_{212} \\ \tau_{212} & \tau_{222} \end{bmatrix}. \tag{8}$$

Assuming that the first latent class and the first item group are reference groups, we need to impose identification constraints as follows: $\tau_{1.k} = \tau_{.1k} = 0$ for each dimension k. Therefore, we have two structural parameters to estimate: τ_{221} and τ_{222}. They represent the advantages (or disadvantages) that class 2 subjects have in solving item group 2 in

[3] As in the unidimensional model case, we assume $H = G$. In addition, we assume that there are the same number of item groups in each test dimension.

dimension 1 and dimension 2, respectively. For example, suppose we obtain significant parameter estimates and $\tau_{221} > \tau_{222} > 0$; this means that class 2 subjects have general advantages in solving the second item group compared with class 1 subjects, while the amount of advantages is greater in dimension 1 than in dimension 2. Figure 2 illustrates this scenario by positioning the two item groups on the θ logit scale of each dimension and for each of the two latent classes. This figure shows that (1) for item group 1, there is no difference in performance between the two latent classes in both dimensions, and (2) item group 2 is easier for class 2 in both dimensions, while the amount of advantage (τ) that class 2 has is larger in dimension 1 than in dimension 2.

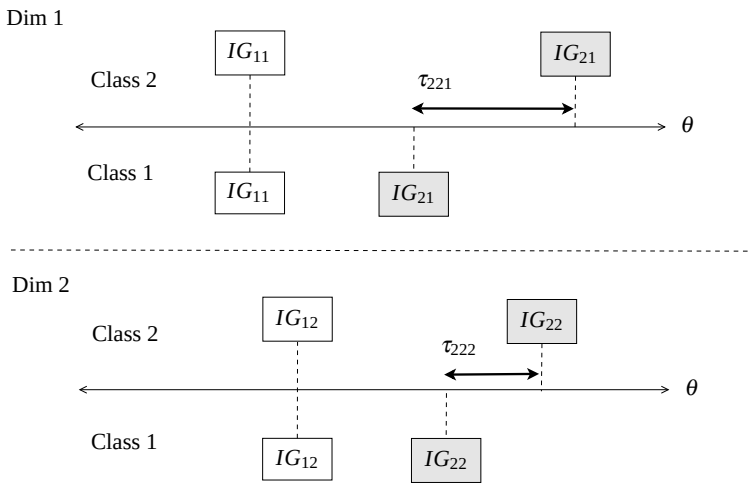

Figure 2:
Positions of two item groups on the θ logit scale (the solid line with an arrowhead on both ends) per dimension. The upper panel is for dimension 1 (Dim1) and the lower panel is for dimension 2 (Dim2). The solid line differentiates two latent classes (the space below the line is for latent class 1, while the space above the line is for latent class 2). Item groups located on the left side of the scale are easier than item groups located on the right side of the scale. IG_{hk} represents item group h in dimension k. τ_{22k} represents the amount of advantage that class 2 subjects have in solving item group 2 items compared with the reference latent class subjects in dimension k

Mixed membership model

We now consider the mixed membership scenario where subjects are allowed to have different class memberships across multiple dimensions. Since we classify subjects into G latent classes for each of K dimensions, we need to consider a total of $T = G^K$ possible class membership combinations in the mixed membership situation. For example, with two latent classes ($G = 2$) for two dimensions ($K = 2$), four classification combinations are possible: [$t = 1$] class 1 in dimension 1 and class 1 in dimension 2; [$t = 2$] class 2 in dimension 1 and class 1 in dimension 2; [$t = 3$] class 1 in dimension

1 and class 2 in dimension 2; [$t = 4$] class 2 in dimension 1 and class 2 in dimension 2. See also Table 2. Note that all places of the 2 by 2 table are all available as classification possibilities. This contrasts with the single membership case where only the diagonal places are available for classification (Table 1).

Table 2:
Multiple membership classification for a two-dimensional test with two latent classes per dimension

Dim1 \ Dim2	class1	class2
class1	[1]	[3]
class2	[2]	[4]

To allow for mixed classifications across dimensions, Equation (6) needs to be revised with $T = G^K$ possible classification combinations as follows:

$$logit(\Pr(y_{ij} = 1|\boldsymbol{\theta}_{jt}, C_j = t)) = \sum_{k=1}^{K} r_{ik}\theta_{jkt} - \beta_i + \sum_{k=1}^{K} r_{ik}\tau_{thk}I_{ihk}. \quad (9)$$

Note that the only difference of the mixed membership model (9) from the single membership model (6) is that latent class g is replaced with class membership combination t ($= 1,...,T$). As in the single membership model, the mean of a reference latent class should be fixed at 0 for any dimension to set the reference points and identify the model. For instance, μ_{tk} = 0 when t includes a reference latent class in dimension k (this will be further illustrated with an example later). As in the single membership model, the structural parameters τ_{thk} are defined per dimension but in each of t latent class combination.

To explain the mixed membership model in a simpler scenario, let us assume that we want to classify subjects into one of two latent classes ($G = 2$) based on two item groups ($H = 2$) for a two-dimensional test ($K = 2$). Equation (9) can be expressed as follows:

$$logit(\Pr(y_{ij} = 1|\boldsymbol{\theta}_{jt}, C_j = t)) = r_{i1}\theta_{j1t} + r_{i2}\theta_{j2t} - \beta_i + r_{i1}\tau_{th1}I_{ih1} + r_{i2}\tau_{th2}I_{ih2}. \quad (10)$$

In this case, the subjects' latent traits are assumed to follow a bivariate normal distribution in each of four latent class combinations: $\boldsymbol{\theta}_{jt} \sim N(\mu_t, \Sigma_t)$. As in the single membership model, the mean of a reference latent class should be fixed at 0 for any dimension to set the reference points and identify the model. That is, μ_{tk} = 0 when t includes a reference latent class in dimension k. For example, suppose the first latent class and the first item group are reference groups. We then impose the following identification constraints on the means of the latent trait distributions: $\mu_1 = (0,0)'$, when $t = 1$ (class 1 both dimensions 1 and 2), $\mu_2 = (\mu_{12}, 0)'$, when $t = 2$ (class 2 in dimension 1 and class 1 in dimension 2), and $\mu_3 = (0, \mu_{23})'$, when $t = 3$ (class 1 in dimension 1

and class 2 in dimension 2). When $t = 4$, subjects belong to class 2 in both dimensions; thus, $\mu_4 = (\mu_{14}, \mu_{24})'$ are freely estimated without constraints.

We have two sets of structural parameters τ_{th1} and τ_{th2} for dimensions 1 and 2, each of which is a 4×2 matrix as follows:

$$\begin{bmatrix} \tau_{111} & \tau_{121} \\ \tau_{211} & \tau_{221} \\ \tau_{311} & 0 \\ \tau_{411} & \tau_{421} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \tau_{112} & \tau_{122} \\ \tau_{212} & 0 \\ \tau_{312} & \tau_{322} \\ \tau_{412} & \tau_{422} \end{bmatrix}. \tag{11}$$

When the first class and the first item group are reference groups in each dimension k, the usual constraints for the structural parameters are imposed in each dimension as follows: $\tau_{1.k} = \tau_{.1k} = 0$. Note that an additional set of constraints are imposed for the mixed membership model: $\tau_{321} = 0$ in $k = 1$ and $\tau_{222} = 0$ in $k = 2$ (these are already applied in the matrices above). This is because by definition when $t = 2$, subjects belong to class 2 in dimension 1 and class 1 in dimension 2; when $t = 3$, subjects belong to class 1 in dimension 1 and class 2 in dimension 2 (see Table 2). In other words, the $t = 2$ subjects do not differ in dimension 2 and the $t = 3$ subjects do not differ in dimension 1. Therefore, it make sense to set $\tau_{222} = 0$ for dimension 2 when $t = 2$ and $\tau_{321} = 0$ for dimension 1 when $t = 3$.

Accordingly, we have a total of four structural parameters to estimate with the mixed membership model: τ_{221}, τ_{322}, τ_{421} and τ_{422}. Here τ_{221} represents the advantage (or disadvantage) that subjects have in solving item group 2 in dimension 1, and τ_{322} represents the advantage (or disadvantage) that subjects have in solving item group 2 in dimension 2. τ_{421} and τ_{422} represent the amount of advantage (or disadvantage) that subjects have in solving item group 2 in both dimensions (dimension 1 and dimension 2, respectively). Note that the single membership model estimates only two structural parameters.

Figure 3 illustrates a fictitious joint mixture distribution for a single membership model and for a mixed membership model with two latent classes for a two-dimensional test. The figure shows that the two latent traits (θ_1 and θ_2) marginally follow a mixture of normal distributions (on the x-axis and the y- axis, respectively). The single membership model shows two subject clusters, while the mixed membership model shows four subject clusters. Note that these clusters are created to visually illustrate and contrast a single membership model with a mixed membership model in terms of possible latent classes that each model can investigate with. It is worth mentioning that the reference points of the scale should be fixed for discussing quantitative differences between latent classes.

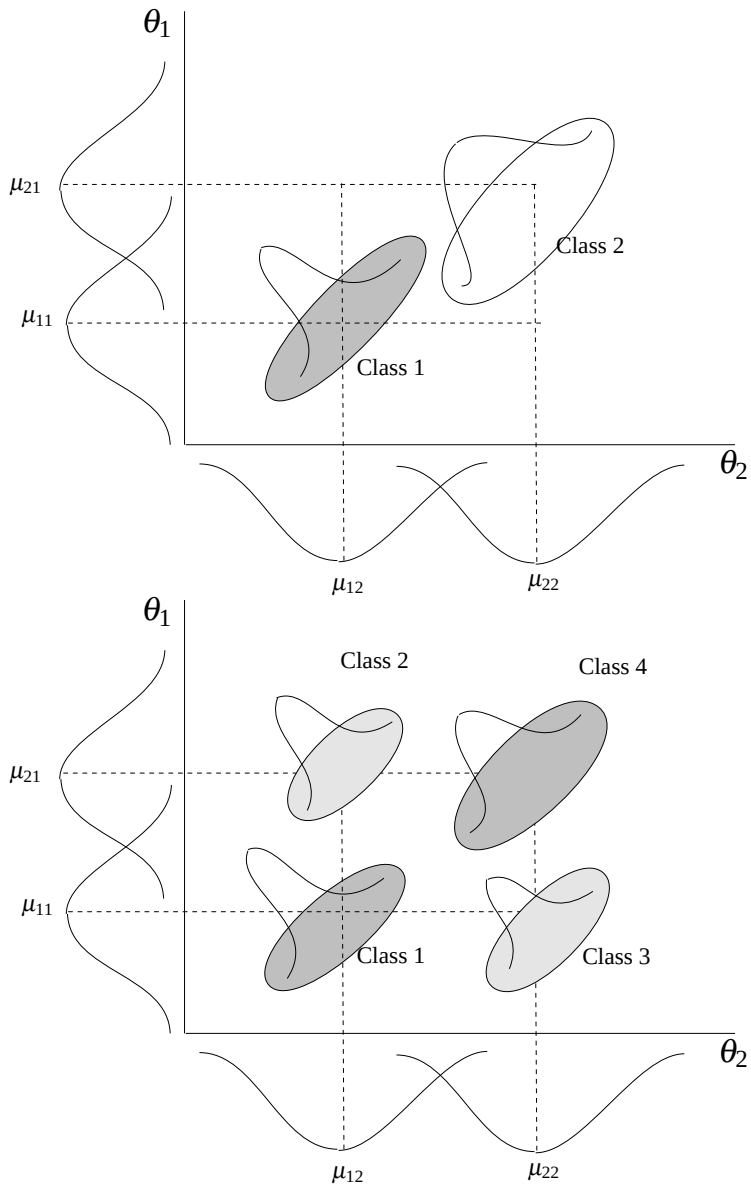

Figure 3:
A hypothetical joint mixture distribution of the two latent traits (θ_1 and θ_2) for Class 1 (C1) and Class 2 (C2) for a single membership model (top) and a mixed membership model (bottom)

Illustrations

Data

To illustrate the two types of confirmatory mixture models for a multidimensional test, we utilize the verbal aggression dataset (Vansteelandt, 2000; De Boeck & Wilson, 2004). This dataset has frequently been utilized in the literature to introduce new IRT models or procedures (e.g., De Boeck & Wilson, 2004; Braeken et al., 2007; Magis et al., 2010; Choi & Wilson, 2015; Jeon & Rijmen, 2016). Hence, we selected this well-known dataset for our illustration in the hope that interested researchers can readily apply the proposed models.

In short, the verbal aggression dataset includes 24 items for 316 respondents (243 female and 73 male). The test items are designed to measure the source of verbal aggression and its inhibition. Each item gives a situation related to one of two blaming types ('Other-to-blame' and 'Self-to-blame'), one of two behavioral modes ('Want' and 'Do'), one of three types of verbally aggressive behavior ('Curse', 'Scold', and 'Shout') and one of four frustrating situations ('Bus', 'Train', 'Store', and 'Operator'). For example, "A bus fails to stop for me. I would want to shout" involves the 'Want' behavior mode, the 'Shout' behavior, the 'Self-to-blame' type, and the 'Bus' situation. Each item asks respondents whether they would agree to give an aggressive verbal response in a given situation. Three response options are provided: 'No', 'Perhaps', and 'Yes'. For current analysis, we dichotomized the item responses by combining 'Perhaps' and 'Yes' categories.

We apply the following rationale to construct a multidimensional confirmatory mixture IRT model with a specialized set of constraints for the described verbal aggression data. First, the to-be-measured trait, verbal aggression, is a multidimensional, rather than a unidimensional construct. This is because the amount of angry behavior that a person displays in a given situation is conceived as a function of the features of the situation as well as the person (Vansteelandt & Van Mechelen, 2004). We consider two dimensions of verbal aggression based on the 'Other-to-blame' and 'Self-to-blame' situation types. These two situations differ in terms of the presence or absence of another person in the situation and are likely to induce a different amount of frustrations to subjects (Vansteelandt & Van Mechelen, 2004).

Second, the amount of subjects' anger elicited by frustrating situations and their thresholds to those situations are likely to differ across subjects (Vansteelandt & Van Mechelen, 2004). This means that the subjects may come from different sub-populations (or latent classes) that are characterized by a different order of the situations and the behaviors in terms of the amount of angry behavior evoked.

Third, there are likely substantial individual differences in the way people express their anger experience (Vansteelandt & Van Mechelen, 2004). In particular, 'Do' vs 'Want' behavior types may differ in terms of the ease with which a behavior is displayed. Specifically, 'Do' behaviors are likely to show a higher response threshold than 'Want'

behaviors because the former implies a higher risk of causing actual damage to the given situation. Hence, we utilize the 'Do' items to differentiate the types and levels of verbal aggression that subjects possess. In other words, we assume two sub-populations (or latent classes) of subjects are differentiable based on their response thresholds to the 'Do' vs. 'Want' behavior items.

Based on this rationale, we specify both single- and mixed-membership models for illustrative purposes. With the single membership model, we hypothesize that subjects belong to a single latent class across two dimensions of verbal aggression. That is, irrespective of situation types (that involve self or others), subjects are assumed to display the same thresholds to the 'Do' behavior items. With the multiple membership model, we relax the single membership assumption and allow subjects to belong to different latent classes across two dimensions of verbal aggression. In other words, interactions between behavior types and situations are allowed with the mixed-membership model. Note that it is possible and perhaps more realistic to presume that subjects differ in their thresholds to the 'Do and 'Want' items depending on the situation types.

Estimation

For the estimation of the proposed models, we utilized Mplus version 7.4 (Muthén & Muthén, 2008) with full information maximum likelihood estimation. We provided annotated Mplus code for both single- and mixed-membership models in Appendix A. To ensure that the parameter estimates were not obtained at local maxima of the log-likelihood function, we utilized multiple random starting values and monitored the convergence. Both the single- and mixed-membership models were successfully converged and estimated for this dataset. To verify parameter recovery of the two models, we additionally conducted a small simulation study. We found that the model parameters were generally well recovered for both types of models under the considered conditions, assuring the reliability of the parameter estimates. Details of the simulation study procedure and results are provided in the supplementary material.

For illustrative purposes, we describe the results obtained from both single- and mixed-membership models although in practice, a researcher may want to choose a single, better-suited model based on theoretical and/or practical considerations as well as relative fit statistics such as Akaike information criterion (AIC) and Bayesian information criterion (BIC). In theory, it would also be informative to evaluate absolute model fit statistics for the two models; however, Mplus currently does not provide absolute fit statistics for complex models including mixture IRT models. In addition, developing absolute fit measures for complex IRT models is an ongoing area of research. Therefore, we leave investigating the absolute fit of the models that we discuss in this article for future studies.

Results: Single membership analysis

The single membership model showed a log-likelihood of -4059.240 (p=35, where p is the number of parameters) with a AIC of 8188.480 and a sample-size adjusted BIC of 8208.920. The model classified approximately 21.8 % subjects into class 1 and 78.2 % subjects into class 2. The performance difference between the two latent classes was quantified with the dimension-specific structural parameters, with $\tau_{221} = -0.927$ (SE= 0.331) and $\tau_{222} = -0.787$ (SE= 0.219). This means that the class 2 subjects had a lower probability of endorsing the 'Do' items compared with class 1 in both dimensions (i.e., for the items that involve 'Other-to-blame' situations (dimension 1) and 'Self-to-blame' situation (dimension 2)).

There were also some differences in the latent trait distributions between the two latent classes. For class 2, the estimated means were 0.960 (SE=0.427) and -0.215 (SE=0.296) for dimensions 1 and 2, respectively (for class 1, the means were fixed at 0 in both dimensions since class 1 was set as the reference group). That is, class 2 subjects showed a higher overall mean in their latent trait distribution of dimension 1 compared to class 1 subjects. The standard deviations were 1.243 (dimension 1) and 0.442 (dimension 2) for class 1, while they were 1.289 (dimension 1) and 2.032 (dimension 2) for class 2. The correlation between the two dimensions was 0.980 and 0.787 for class 1 and class 2, respectively. These results suggest that (1) between-class differences existed in the latent trait variation. In particular, class 1 showed a relatively smaller variation in dimension 2 (which involves 'Self-to-blame' situations) compared with class 1; and (2) the between-dimension correlation was quite large in both classes, although the correlation was slightly larger in class 1 than in class 2. The latter may be interpreted as that verbal aggression levels were generally consistent across situation types, while the consistency was a little stronger for class 1 subjects than class 2. Note that this kind of heterogeneity in subjects' latent trait distributions would not be found if a unidimensional mixture model or a regular multidimensional model was applied to analyze the data.[4]

Figure 4 displays the estimated item difficulty parameter (β_i) values from the single-membership model. Clearly, there were systematic differences in the β_i parameter estimates between class 1 and class 2. Specifically, the 'Do' items were more difficult for class 2 subjects than class 1 subjects in dimension 1 (items 7 to 12) and dimension 2 (items 19 to 24). Those differences amount to the dimension-specific structural parameter estimates, τ_{221} and τ_{222} for dimensions 1 and 2, respectively.

[4]Both uni- and multi-dimensional exploratory 1PL mixture models (with 2 classes) were not converged unless additional constraints were imposed. A regular (non-mixture) 1PL multidimensional model showed AIC and BIC values (log-likelihood=-4080.149 (p=27), AIC= 8214.298 BIC= 8230.066) which were worse than those of the single-membership model.

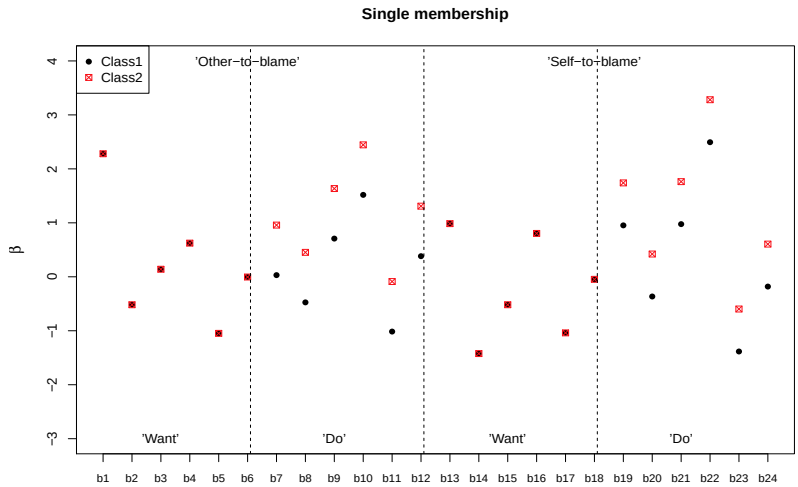

Figure 4:
Estimated item difficulty parameter (β_i) values from the single-membership model for the verbal aggression data. b1 to b24 represent the β_i parameters for $i = 1, \ldots, 24$.

Results: Mixed Membership Analysis

The mixed membership model showed a log-likelihood of -4049.377 (p=46) with a AIC of 8190.755 and a sample-size adjusted BIC of 8217.619. The mixed membership model showed similar fit to the single membership model in the current example.

From the mixed membership analysis results, we first checked the estimated structural parameter values. They were $\tau_{221} = 0.769$ (SE=0.336) for class 2 (dimension 1), $\tau_{322} = -0.528$ (SE=0.491) for class 3 (dimension 2), $\tau_{421} = -1.570$ (SE= 0.661) for class 4 (dimension 1), and $\tau_{422} = -1.291$ (SE=0.484) for class 4 (dimension 2). The estimates were all significant at the 5 % level except τ_{322} for class 3. This result suggests that (1) class 2 had a higher probability of endorsing the 'Do' items that involve the 'Other-to-blame ' situations (dimension 1) compared with class 1; (2) there was insufficient evidence that class 3 had a lower probability of endorsing for the 'Do' items involved with the 'Self-to-blame ' situations (dimension 2) compared with class 1; and (3) class 4 showed a lower probability of endorsing the 'Do' items compared with class 1, irrespective of situation types ('Other-to-blame' vs. 'Self-to-blame ' situations). This is a quite intriguing result in that the 'Do' items were found to be easier only in 'Other-to-blame' situations for some group of subjects (class 2). This implies that there may be interactions between situations and behavior types in terms of the amount of anger evoked by subjects.

We found that the participants were classified into four latent groups, approximately 43.0 % in class 1, 29.8 % in class 2, 11.1 % in class 3, and 16.1 % in class 4. It is interesting to see that there were a non-negligible number of subjects (about 40.9 %) who showed a mixed membership profile across two dimensions (class 1 in dimension 1 but class 2 in dimension 2 or vice versa). Note that such mixed membership people could not be identified with the single membership analysis that forces subjects to have only one class membership across all dimensions.

In addition, the four latent classes presented some differences in their latent trait distributions. Table 3 lists the estimated means, standard deviations, and between-dimension correlation for the two latent traits in each latent class.

Table 3:
The estimated means, standard deviations (SD), correlation between the two latent traits (Dim1 and Dim2) from the mixed membership model.

	Mean (SE)		SD		Cor
	Dim1	Dim2	Dim1	Dim2	
Class 1	0	0	1.378	1.196	1.00
Class 2	-1.287(0.332)	0	1.550	3.108	0.99
Class 3	0	-1.096(0.631)	1.372	0.425	0.53
Class 4	0.938(0.473)	0.687(0.522)	0.968	1.096	1.00

From Table 3, we found that (1) the latent trait means varied across the four latent class combinations. For instance, class 2 subjects showed a smaller mean than class 1 in dimension 1; class 3 subjects showed a smaller mean than class 1 in dimension 2; and class 4 subjects showed a larger mean than class 1 in dimension 1, (2) the latent trait standard deviations also varied across the latent classes. Interestingly, class 2 showed a larger variation but class 3 showed a smaller variation than class 1 in dimension 2; and (3) the correlation between the two dimensions was close to 1.0 in all classes except class 3 that showed a correlation of 0.53. Note that in the single membership analysis, we found that (1) the mean difference in dimension 2 between latent classes 1 and 2 was not significantly different from 0, (2) the subject heterogeneity of the latent trait distribution was larger in dimension 2 than in dimension 1 for class 2, and (3) class 2 showed a relatively smaller between-dimension correlation than class 1. These differences in the results might stem from that the mixed membership model relaxes a somewhat stringent assumption taken by the single membership model.

Figure 5 displays the estimated item difficulty parameter (β_i) values from the mixed membership model. Observe that the β_i parameter estimates were structurally different for the 'Do' items across three latent classes. For the 'Want' items, there were no differences in the difficulty parameter estimates between latent classes. For class 2 subjects, the 'Do' items in dimension 2 (items 19 to 24) were easier than class 1 subjects. For class 4 subjects, the 'Do' items were harder in both dimensions (items 7 to 12,

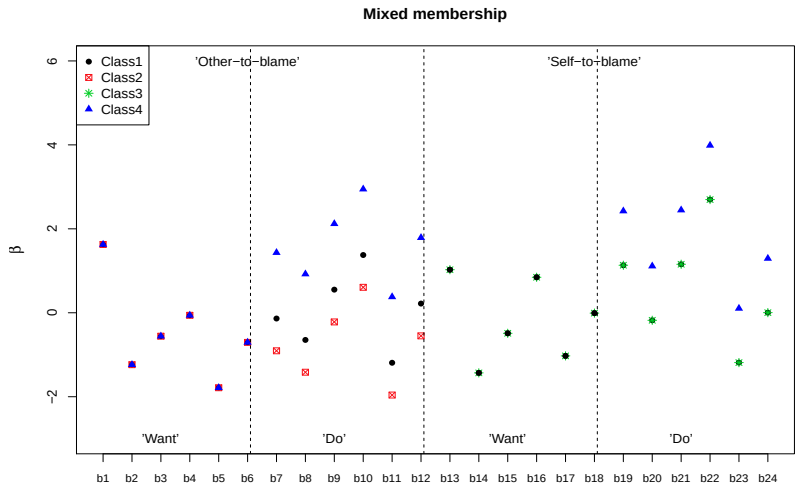

Figure 5:
Estimated item difficulty parameter (β_i) values from the mixed membership model for the verbal aggression data. b1 to b24 represent the β_i parameters for $i = 1,\ldots,24$.

items 19 to 24) than class 1 subjects. This analysis result confirms that systematic class differences across dimensions could be successfully identified with the mixed membership model.

Discussion

In this paper, we discussed specialized confirmatory mixture IRT modeling for multidimensional assessments. The model is confirmatory in two senses: (1) the number/nature of latent classes is known prior to data analysis, and (2) prior knowledge on items (and their characteristics) is used to hypothesize the item-class relationship and to differentiate latent classes. Researchers who are more accustomed to an ordinary, exploratory use of mixture IRT modeling may feel that such a confirmatory approach seems somewhat unnatural. However, we would like to stress that a confirmatory approach could be introduced in mixture modeling as in confirmatory factor analysis, because it can serve for the purpose of verifying a researcher's hypothesis on the nature of the postulated latent classes and latent class differentiation.

For multidimensional extensions, we discussed two types of classification scenarios: (1) a single membership case where subjects have only one latent class membership across all dimensions, and (2) a mixed membership case where subjects can have different

memberships across multiple dimensions. We discussed and illustrated with an empirical example how these two classification methods could be adopted for specialized confirmatory mixture IRT modeling.

We would like to make a few points regarding mixed membership analysis in contrast to single membership analysis. First, mixed membership models become exponentially complicated as the number of dimensions and/or latent classes increase. For instance, with two latent classes for a two-dimensional test, we have four latent class combinations to consider, while when we have four latent class for a three-dimensional test, we need to consider 16 latent class combinations. With such a large number of latent classes, an extremely large sample size may be needed for reliable estimation of mixed membership models.

Second, from mixed membership analysis, a group of subjects with different class memberships across dimensions may be captured because of the difference in the moments (mean and standard deviation) of their latent trait distributions rather than difference in their performance on the chosen item groups. This may be the case especially when the underlying latent trait distribution is not normal in the data (e.g., Sen et al., 2015). Hence, researchers should be aware of this issue when interpreting results from mixed membership analysis.

Third, mixed membership analysis can be viewed as a special case of single membership analysis when subjects are simultaneously classified into a multidimensional latent space. With a larger number of latent classes (than for usual single membership analysis), all latent class combinations (that mixed membership analysis considers) may be captured with this special single membership analysis, although the estimation of this model is likely to be more challenging than the mixed membership model. Hence, if these two types of analysis procedures produce the same classification results, it may be sensible to choose the mixed membership model because of its relative simplicity.

Note that from a modeling perspective, our model can be seen as a multidimensional extension of the Saltus model (Wilson, 1989; Mislevy & Wilson, 1996). The original Saltus model was proposed for unidimensional assessments and has been extended for polytomous item responses (Draney, 2007) and with item covariates (Draney & Wilson, 2008). Jeon (2018) presented several extensions of the Saltus model with item discrimination parameters, person predictors, and ordinal item responses. Recently, Jeon et al. (in press) presented an application of a multidimensional extension of the Saltus model with item covariates, for analyzing developmental stages of children for a deductive reasoning test. However, those authors focused on single membership classification. A unique contribution of the current work is that we differentiate, formulate, and compare single and mixed classification scenarios that are sensible for the analysis of multidimensional assessments and discuss how these two different classification scenarios can be applied in the context of confirmatory mixture modeling.

The specialized confirmatory mixture modeling approach that we discussed in this paper has merits both in technical and substantive aspects (e.g., for computation and result

interpretation) because of its parsimony. In addition, the proposed multidimensional extensions have a variety of applications; for instance, they can be applied to scenarios where a researcher would like to test a hypothesis on structural differences between latent classes with multidimensional assessments. Suppose we have a mathematics test that consists of multiple sub-tests (e.g., algebra, geometry, and number theory) where each sub-test includes a key item set that are designed to identify students who would need extra support to improve their mathematics skills in each content area. In this case, the proposed modeling can successfully be applied to differentiate those students who are in need of special assistance (from students without needing extra support) as well as to evaluate whether the designed test items function effectively as expected in terms of differentiating a potentially disadvantaged group of students.

On a final note, we would like to mention that it may be beneficial to incorporate additional item discrimination parameters into the discussed models, in order to improve classification precision. This is because applying a one-parameter model to two-parameter data can possibly lead to identifying spurious latent classes (Alexeev et al., 2011). In addition, including person predictors into the model for classification can also be useful to prevent possible misclassification of latent classes (G.-H. Huang & Bandeen-Roche, 2004). We will leave further investigations on the impacts of model misidentification and possible remedies for future studies.

Appendix A

Here we provide example Mplus code for fitting single membership and mixed membership models for the verbal aggression data.

<Single membership model>

```
!! Header of input file
TITLE: Single membership model for verbal aggression data

!! Data file specification
DATA: FILE = verbal.dat;

!! Define variables and specify number of latent classes
VARIABLE:
NAMES = u1-u24;
CATEGORICAL = u1-u24; ! binary item responses
MISSING = ALL(99); ! missing data are coded as 99
CLASSES = c (2) ; ! define number of latent classes

!! Estimation settings
ANALYSIS: TYPE = MIXTURE; ! estimate finite mixture model
ALGORITHM = INTEGRATION; ! 15 default quadrature points
STARTS = 500 10 ; ! use multiple random start (can be increased if needed)

!! Model specification
MODEL:
```

```
! Overall model
%OVERALL%
f1 BY u1-u12@1; ! item loading parameters in dim 1
f2 By u13-u24@1; ! item loading parameters in dim 2

! Model for class 1
%c#1%
f1 BY u1-u12@1;
f2 By u13-u24@1;
[f1@0]; [f@02]; ! factor means fixed at 0 (reference group)
f1; f2;
f1 with f2;
[u1$1-u12$1](a1-a12 ) ; !difficulty parameters in dim 1
[u13$1-u24$1](a13-a24); !difficulty parameters in dim 2


! Model for class 2
%c#2%
f1 BY u1-u12@1;
f2 By u13-u24@1;
[f1]; [f2];
f1; f2; ! factor means freely estimated
f1 with f2;

! Use different difficulty parameter labels for  'Do' items
[u1$1-u12$1] (a1-a6 b7-b12 ); !i7-i12 in dim 1
[u13$1-u24$1](a13-a18 b19-b24);  !i19-i24 in dim 2

! Set model constraints
MODEL CONSTRAINT:

NEW(tau1 tau2); ! define structural parameters in dim 1 and din 2

!! Define structural parameter for dim 1 as difference
!! in difficulty parameters for 'Do' items
!! between class 1 and class 2 (i7-i12)
  tau1 = a7-b7;
  tau1 = a8-b8;
  tau1 = a9-b9;
  tau1 = a10-b10;
  tau1 = a11-b11;
  tau1 = a12-b12;

!! Define structural parameter for dim 2 as difference
!! in difficulty parameters for 'Do' items
!! between class 1 and class 2 (i19-i24)
  tau2 = a19-b19;
  tau2 = a20-b20;
  tau2 = a21-b21;
  tau2 = a22-b22;
  tau2 = a23-b23;
  tau2 = a24-b24;


!! Save posterior probabilities for latent class membership
Savedata:
file is prob1_single.txt ;
save is cprob;
```

<Mixed membership model>

```
!! Header of input file
TITLE: Mixed membership model for verbal aggression data

!! Data file specification
DATA: FILE = verbal.dat;

!! Define variables and specify number of latent classes
VARIABLE:
NAMES = u1-u24;
CATEGORICAL = u1-u24; ! binary item responses
MISSING = ALL(99); ! missing data are coded as 99
CLASSES = c1 (2) c2(2); ! define two latent classes for dim 1 and dim 2

!! Estimation settings
ANALYSIS: TYPE = MIXTURE; ! estimate finite mixture model
ALGORITHM = INTEGRATION; ! 15 default quadrature points
STARTS = 500 10; ! use multiple random start (can be increased if needed)

!! Model specification
MODEL:
! Overall model
%OVERALL%
f1 BY u1-u12@1; ! item loading parameters in dim 1
f2 By u13-u24@1; ! item loading parameters in dim 2

! Model for class 1
%c1#1.c2#1% ! class 1 in dim 1 and class 1 in dim 2
f1 BY u1-u12@1;
f2 By u13-u24@1;
[f1@0]; [f2@0]; ! factor means fixed at 0 (reference group)
f1 (v11); f2 (v12);
f1 with f2 (cov1);

[u1$1-u12$1](a1-a12 ) ; ! difficulty parameters in dim 1
[u13$1-u24$1](a13-a24); ! difficulty parameters in dim 2

! Model for class 2
%c1#2.c2#1%  ! class 2 in dim 1 and class 1 in dim 2
f1 BY u1-u12@1;
f2 By u13-u24@1;
[f1]; [f2@0];
f1 (v21); f2 (v22);
f1 with f2 (cov2);

! Use different difficulty parameter labels for  'Do' items in dim 1
[u1$1-u12$1] (a1-a6 d7-d12 ); ! i7-i12 in dim 1
[u13$1-u24$1](a13-a24);

! Model for class 3
%c1#1.c2#2%  ! class 1 in dim 1 and class 2 in dim 2
f1 BY u1-u12@1;
f2 By u13-u24@1;
[f1@0]; [f2];
f1 (v31); f2 (v32);
```

```
f1 with f2 (cov3);

! Use different difficulty parameter labels for  'Do' items in dim 2
[u1$1-u12$1] (a1-a12 );
[u13$1-u24$1](a13-a18 d19-d24); ! i19-i24 in dim 2

! Model for class 4
%c1#2.c2#2%  ! class 2 in dim 1 and class 2 in dim 2
f1 BY u1-u12@1;
f2 By u13-u24@1;
[f1]; [f2]; ! factor means freely estimated
f1 (v41); f2 (v42);
f1 with f2 (cov4);

! Use different difficulty parameter labels
! for  'Do' items in dim 1 and dim 2
[u1$1-u12$1] (a1-a6 b7-b12 ); !i7-i12 in dim 1
[u13$1-u24$1](a13-a18 b19-b24);  !i19-i24 in dim 2


! Set model constraints
MODEL CONSTRAINT:

! define structural parameters for classes 2, 3, and 4
NEW(tau1 tau2 tau3 tau4);

!! For class 2, define structural parameter for dim 1 as difference
!! in difficulty parameters for 'Do' items
!! between class 1 and class 2 (i7-i12)
  tau1 = a7-d7;
  tau1 = a8-d8;
  tau1 = a9-d9;
  tau1 = a10-d10;
  tau1 = a11-d11;
  tau1 = a12-d12;

!! For class 3, define structural parameter for dim 1 as difference
!! in difficulty parameters for 'Do' items
!! between class 1 and class2 (i19-i24)
  tau2 = a19-d19;
  tau2 = a20-d20;
  tau2 = a21-d21;
  tau2 = a22-d22;
  tau2 = a23-d23;
  tau2 = a24-d24;

!! For class 4, define structural parameter for dim 1 as difference
!! in difficulty parameters for 'Do' items
!! between class 1 and class 2 (i7-i12)
  tau3 = a7-b7;
  tau3 = a8-b8;
  tau3 = a9-b9;
  tau3 = a10-b10;
  tau3 = a11-b11;
  tau3 = a12-b12;

!! For class 4, define structural parameter for dim 2 as difference
!! in difficulty parameters for 'Do' items
!! between class 1 and class 2 (i19-i24)
```

```
  tau4 = a19-b19;
  tau4 = a20-b20;
  tau4 = a21-b21;
  tau4 = a22-b22;
  tau4 = a23-b23;
  tau4 = a24-b24;


!! Save posterior probabilities for latent class membership
Savedata:
file is prob2_mixed.txt ;
save is cprob;
```

Acknowledgement

The authors greatly appreciate the anonymous reviewers whose comments have greatly improved this manuscript.

References

Alexeev, N., Templin, J., & Cohen, A. (2011). Spurious latent classes in the mixture Rasch model. *Journal of Educational Measurement, 3*, 313-332.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics, 26*, 381-409.

Boughton, K. A., & Yamamoto, K. (2007). A HYBRID model for test speededness. In von Davier & C. Carstensen (Eds.), *Multivariate and mixture distribution rasch models* (p. 147-156). New York: Springer-Verlag.

Braeken, J., Tuerlinckx, F., & De Boeck, P. (2007). Copula functions for residual dependency. *Psychometrika, 72*, 393-411.

Choi, I.-H., & Wilson, M. (2015). Multidimensional classification of examinees using the mixture random weights linear logistic test model. *Educational and Psychological Measurement, 75*, 78-101.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*, 133-148.

De Boeck, P., & Wilson, M. (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach.* New York: Springer-Verlag.

De Jong, M. G., & Steenkamp, J.-B. E. M. (2010). Finite mixture multilevel multidimensional ordinal IRT models for large scale cross-cultural research. *Psychometrika, 75*, 3-32.

Draney, K. (2007). The Saltus model applied to proportional reasoning data. *Journal of Applied Measurement, 8*, 438-455.

Draney, K., & Wilson, M. (2008). A LLTM approach to the examination of teachers' ratings of classroom assessment tasks. *Psychology Science, 50*, 417-432.

Finch, W. H., & Finch, M. E. H. (2013). Investigation of specific learning disability and testing accommodations based differential item functioning using a multilevel multidimensional mixture item response theory model. *Educational and Psychological Measurement, 73*, 973-993.

Finch, W. H., & French, B. F. (2012). Parameter estimation with mixture item response theory models: A Monte Carlo comparison of maximum likelihood and Bayesian methods. *Journal of Modern Applied Statistical Methods, 11*, 167-178.

Huang, G.-H., & Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika, 69*, 5-32.

Huang, H.-Y. (2016). Mixture IRT model with a higher-order structure for latent traits. *Educational and Psychological Measurement, 77*, 275-304.

Jeon, M. (2018). A constrained confirmatory mixture irt model: Extensions and estimation of the saltus model using mplus. *The Quantitative Methods for Psychology, 14*, 120-136.

Jeon, M., Draney, K., Wilson, M., & Sun, Y. (in press). Investigation of adolescents' developmental stages in deductive reasoning: An application of a specialized mixture irt approach. *Behavior Research Methods*.

Jeon, M., & Rijmen, F. (2016). A modular approach for item response theory modeling with the R package FLIRT. *Behavior Research Methods, 48*, 742-755.

Jin, K.-Y., Chen, H.-F., & Wang, W.-C. (2018). Mixture item response models for inattentive responding behavior. *Organizational Research Methods, 21*, 197-225.

Lukowski, S. L., DiTrapani, J., Jeon, M., Wang, Z., Schenker, V. J., Doran, M. M., … Petrill, S. A. (in press). Multidimensionality in the measurement of math-specific anxiety and its relationship with mathematical skills. *Learning and Individual Differences*.

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*, 847-862.

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

Mislevy, R. J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika, 61*, 41-71.

Molenaar, D., Oberski, D., Vermunt, J., & De Boeck, P. (2016). Hidden markov IRT models for responses and response times. *Multivariate Behavioral Research, 51*, 606-626.

Muthén, L., & Muthén, B. (2008). *Mplus User's Guide*. Angeles, CA: Muthen & Muthen.

Richardson, F. C., & Suinn, R. M. (1972). The mathematics anxiety rating scale: Psychometric data. *Journal of Counseling Psychology, 19*, 551 - 554.

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guildford Press.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*, 213-232.

Sen, S., Cohen, A. S., & Kim, S. H. (2015). The impact of non-normality on extraction of spurious latent classes in mixture IRT models. *Applied Psychological Measurement, 40*, 98-113.

Tijmstra, J., Bolsinova, M. A., & Jeon, M. (in press). Generalized mixture IRT models with different item-response structures: A case study using Likert-scale data. *Behavior Research Methods*.

Vansteelandt, K. (2000). *Formal models for contextualized personality psychology.* (Unpublished doctoral dissertation, K.U. Leuven, Belgium)

Vansteelandt, K., & Van Mechelen, I. (2004). The personality triad in balance: Multidimensional individual differences in situation-behavior profiles. *Journal of Research in Personality, 38*, 367-393.

Wilson, M. (1989). Saltus: a psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105*, 276-289.

Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.), *Applications of*

latent trait and latent class models in the social sciences (p. 89-98). New York, NY: Waxmann Verlag GmbH.

Supplementary material

Here we discuss the simulation study that we conducted to evaluate parameter recovery of the two types of models that we discussed in the manuscript. To this end, we considered a testing situation analogous to the empirical data setting utilized in Section 3. Specifically, 24 test items of a two-dimensional test were considered with two item groups (6 items per item group in each dimension) with two latent classes. We then considered two sample size conditions $N = 500$ and $N = 1000$ (with equal mixing proportions across latent classes for each model). Showing parameters recovery in relatively small sample size situations is meaningful because one can expect generally improved recovery accuracy in larger sample sizes.

The data generating parameter values were set similar to the parameter estimates obtained from each model fitted to the verbal aggression data. For each model, 100 datasets were generated and estimated with Mplus, with the same maximum likelihood estimation setting as in the empirical study. Potential label switching between runs were checked for the two fitted models.

Figures 6 to 9 display the bias and root mean square error (RMSE) of the estimated model parameters for the models in the two sample size conditions.

For the single membership model with 32 parameters, the bias was not significantly different from 0 at the 5% level except for the two parameters, σ^2_{21} (t = -2.78, p<0.01) and β_{10} (t = -2.37, p = 0.02) when $N = 500$. When $N = 1000$, the bias was insignificant for all model parameters. The RMSE ranged from 0.02 to 0.23 for all model parameters when $N = 500$ and ranged from 0.01 to 0.16 when $N = 1000$.

For the mixed membership model with 40 parameters, the bias was not significantly different from zero at the 5% significance level, except σ^2_{41} ($t = -2.23$, $p = 0.03$) and β_{10} ($t = -2.69$, $p = 0.01$) when $N = 500$. When $N = 1000$, the bias was insignificant except τ_{322} ($t = -2.18$, $p = 0.03$) and β_{11} ($t = -2.15$, $p = 0.03$). For all model parameters, The RMSE ranged from 0.07 to 0.31 when $N = 500$ and ranged from 0.03 to 0.19 when $N = 1000$.

These results assure that the model parameters could generally be well recovered for both types of models under the considered conditions. The bias and RMSE tended to be slightly larger for the mixed membership model than the single membership model. This makes sense given that the mixed membership model is a more complex model and contains more parameters to estimate than the single membership model. For both models, the bias and RMSE tended to decrease when the sample size is $N = 1000$ than $N = 500$. This result suggests that the estimation accuracy can indeed be improved with larger sample sizes.

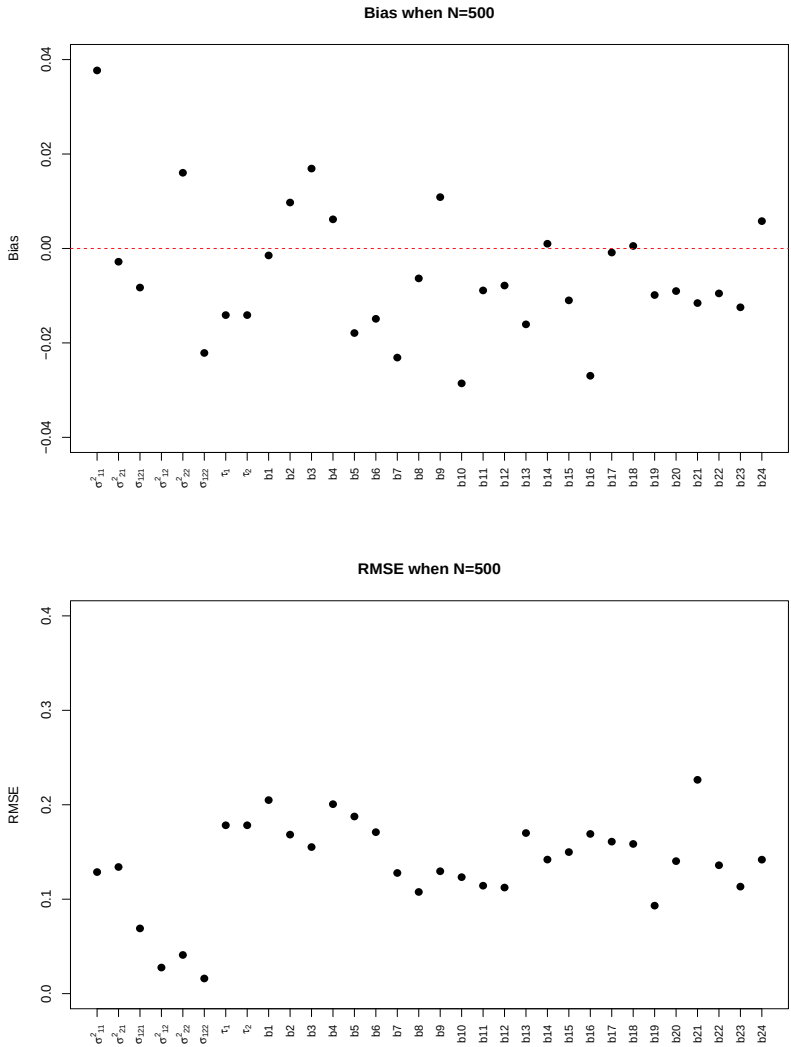

Figure 6:
Bias and RMSE of the estimated model parameters for the single membership model when $N = 500$. τ_g, σ^2_{1g}, σ^2_{2g}, σ_{12g}, and β_i indicate the structural parameters, the factor variances for dimensions 1 and 2, covariance for class g ($g = 1, 2$) and the item difficulty parameters b_i ($i = 1, \ldots, 24$), respectively.

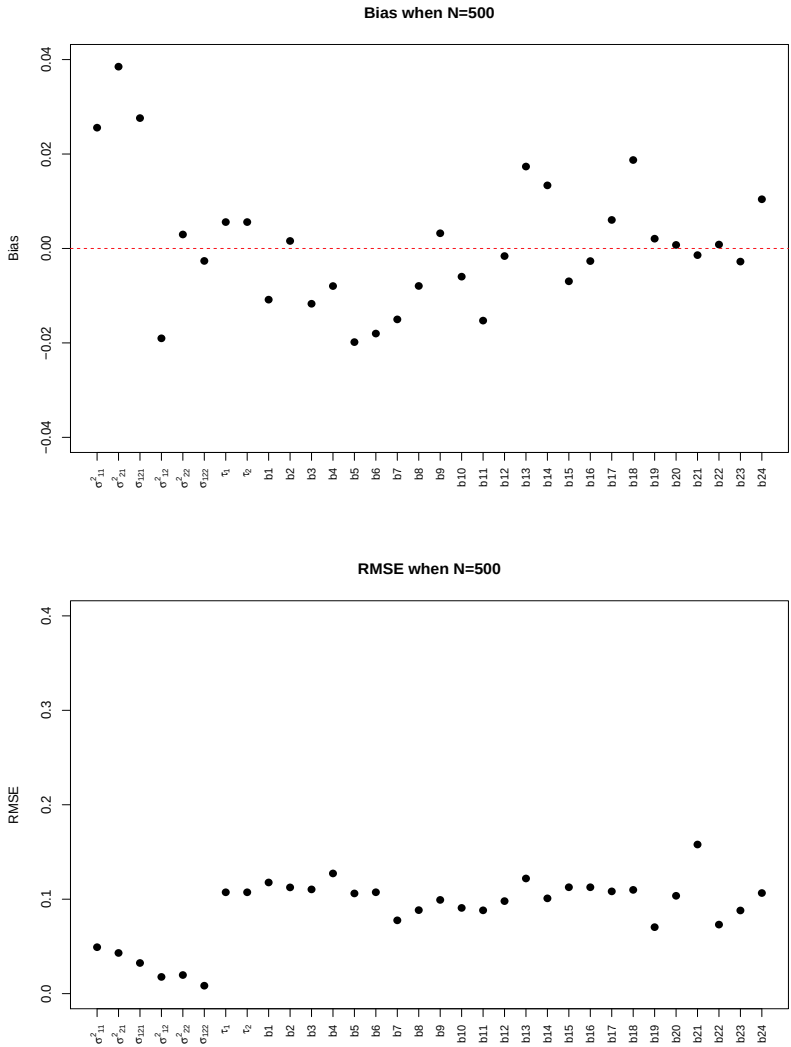

Figure 7:
Bias and RMSE of the estimated model parameters for the single membership model when $N = 1000$. τ_g, σ^2_{1g}, σ^2_{2g}, σ_{12g}, and β_i indicate the structural parameters, the factor variances for dimensions 1 and 2, covariance for class g ($g = 1, 2$) and the item difficulty parameters b_i ($i = 1, \ldots, 24$), respectively.

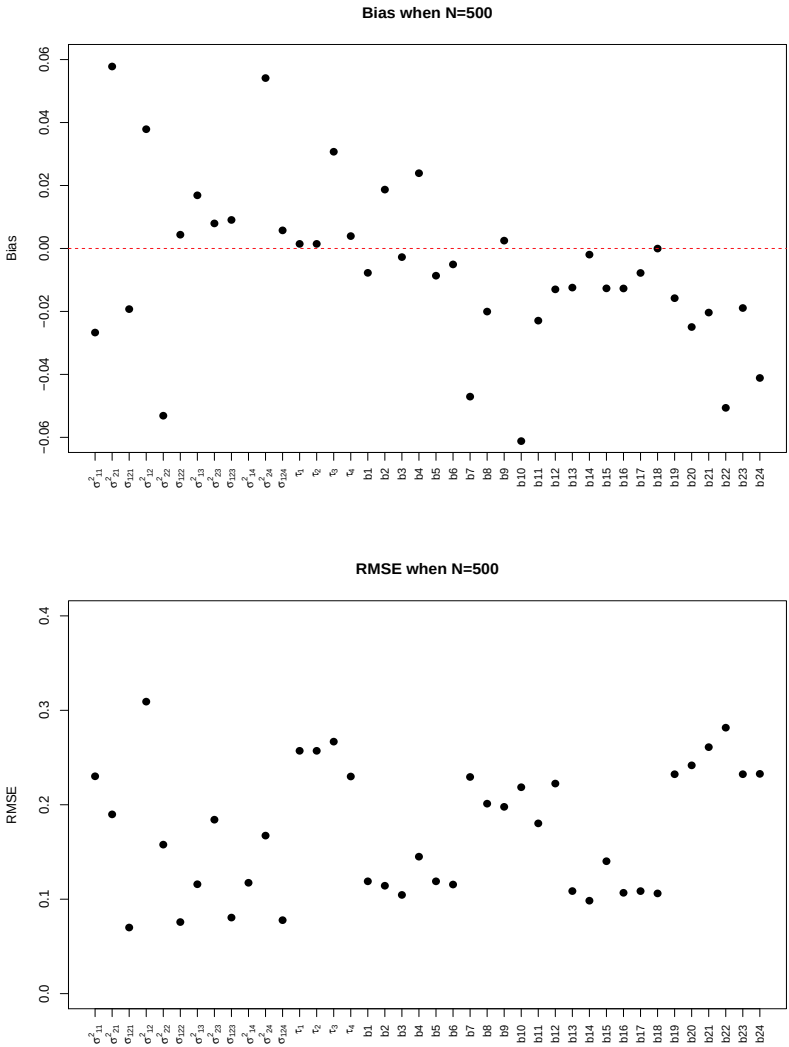

Figure 8:
Bias and RMSE of the estimated model parameters for the mixed membership model when $N = 500$. τ_g, σ^2_{1g}, σ^2_{2g}, σ_{12g}, and β_i indicate the structural parameters, the factor variances for dimensions 1 and 2, covariance for class g ($g = 1, \ldots, 4$) and the item difficulty parameters b_i ($i = 1, \ldots, 24$), respectively.

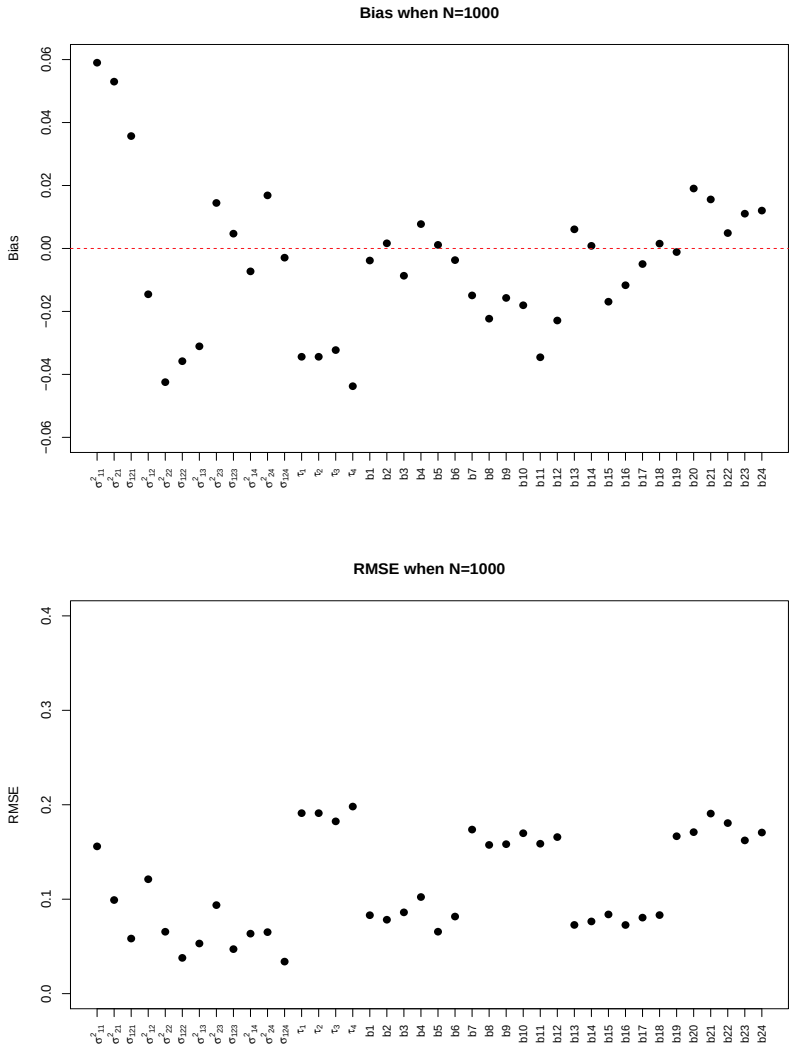

Figure 9:
Bias and RMSE of the estimated model parameters for the mixed membership model when $N = 1000$. τ_g, σ^2_{1g}, σ^2_{2g}, σ_{12g}, and β_i indicate the structural parameters, the factor variances for dimensions 1 and 2, covariance for class g ($g = 1, \ldots, 4$) and the item difficulty parameters b_i ($i = 1, \ldots, 24$), respectively.