Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Spring 2016, Vol. 35, No. 1, pp. 48–61

An NCME Instructional Module on Latent DIF Analysis Using Mixture Item Response Models

Sun-Joo Cho, *Vanderbilt University,* Youngsuk Suh, *Rutgers, The State University of New Jersey,* and Woo-yeol Lee, *Vanderbilt University*

The purpose of this ITEMS module is to provide an introduction to differential item functioning (DIF) analysis using mixture item response models. The mixture item response models for DIF analysis involve comparing item profiles across latent groups, instead of manifest groups. First, an overview of DIF analysis based on latent groups, called latent DIF analysis, is provided and its applications in the literature are surveyed. Then, the methodological issues pertaining to latent DIF analysis are described, including mixture item response models, parameter estimation, and latent DIF detection methods. Finally, recommended steps for latent DIF analysis are illustrated using empirical data.

Keywords: differential item functioning, item response model, mixture model

The assessment of the presence of differential item functioning (DIF) is a key component for test development and validation. An item is said to exhibit DIF when the item functions differently in a focal group in comparison with a reference group after controlling for differences in levels of performance on a latent trait (e.g., ability) of interest (Holland & Wainer, 1993; Scheuneman, 1979).[1] A popular DIF analysis method involves detecting DIF items using multigroup item response models where the groups are manifest groups (Bock & Zimowski, 1997). Procedures for detecting DIF items based on the manifest groups are by now well established in psychometric research (Holland & Wainer, 1993; see Millsap, 2011, for a review).

When DIF is detected based on manifest groups (called *manifest DIF*), the group membership of interest is set a priori for DIF analysis as a known group such as gender and ethnicity groups. Instead of setting the group membership of interest a priori, an unknown homogeneous subgroup, called a latent group,[2] can be found from the data using mixture modeling (McLachlan & Peel, 2000). It is assumed that items function the same way within a latent group. DIF detected based on the latent groups is called *latent DIF*, as opposed to manifest DIF.

The mixture item response model[3] is similar to a multigroup item response model, except that the group of interest is not specified a priori but is determined by the results from the model parameter estimation. As in multigroup item response models, item parameters and latent variable(s) can be different across latent groups in mixture item response models.

Sun-Joo Cho, Department of Psychology and Human Development, Vanderbilt University, Nashville, TN 37240; sj.cho@vanderbilt.edu. Youngsuk Suh, Department of Educational Psychology, Rutgers, The State University of New Jersey, New Brunswick, NJ 08901; youngsuk.suh@gse.rutgers.edu. Woo-yeol Lee, Department of Psychology and Human Development, Vanderbilt University, Nashville, TN 37240; woo-yeol.lee@vanderbilt.edu.

In this Instructional Topics in Educational Measurement Series (ITEMS) module, we present an overview of a latent DIF analysis using mixture item response models. We also provide recommended steps for applying mixture item response models with an empirical illustration for latent DIF analysis. We view the uses of mixture item response models as latent DIF applications when item parameter estimates across latent groups (called item profiles) are compared to detect DIF items and to explain latent groups. In this article, a latent DIF analysis is considered for tests that have multidimensionality unintentionally. DIF for unintentionally multidimensional tests should be clearly distinguished from DIF for intentionally multidimensional tests (Oshima, Raju, & Flowers, 1997). When a test is intentionally multidimensional, two or more dimensions are relevant to the purpose of the test. When a test is unintentionally multidimensional, only one dimension is considered relevant, although some items may be inadvertently sensitive to secondary, that is, nuisance dimensions. Reading ability, for example, can be considered a nuisance dimension in a test designed to measure algebraic symbol manipulation. For items suspected of DIF, there are shifts with DIF magnitudes on item parameters in a focal group. A nuisance dimension other than the primary dimension can be included in an item response model to explain the individual differences for the focal group and DIF items, indicating that multidimensionality exists as a result of DIF at the test level (e.g., De Boeck, Cho, & Wilson, 2011).

Manifest DIF versus Latent DIF

Latent DIF analysis is explained by contrasting it with manifest DIF analysis regarding the analysis purpose and DIF detection procedures.

Analysis Purpose

The main purpose of manifest DIF analysis is to detect DIF items to answer the question, "for whom does an item

function differently?" In manifest DIF analysis, it is assumed that the manifest groups are homogeneous subgroups. However, it has been shown that the assumption of homogeneity is not always met in manifest DIF analysis. Manifest group membership (e.g., gender) associated with DIF often had a weak relationship with the (homogeneous) latent groups detected using mixture item response models (Cohen & Bolt, 2005; de Ayala, Kim, Stapleton, & Dayton, 2002). In latent DIF analysis, a latent group is assumed to be homogeneous because persons classified into the same latent group from mixture item response models show similar item response behaviors. The results of manifest DIF and latent DIF analyses would be similar when there is a large overlap between manifest group membership and latent group membership (identified from data).

Latent groups can be characterized by investigating how they respond to items or item types and to what extent the latent groups can be linked to person characteristics such as demographic information. For this reason, a latent DIF analysis with person characteristics is useful for explaining DIF, in addition to detecting DIF items between latent groups.

De Boeck et al. (2011) listed four a priori reasons to consider latent DIF analysis instead of manifest DIF analysis: the *no idea* reason (one has no idea what group membership is of interest, or the group membership information is missing), an *unobservable* reason (the group membership of interest is not observable), a *reliability* reason (manifest group membership may not be perfectly reliable), and a *validity* reason (manifest group membership may not provide a perfectly valid indication of the true group membership). Example papers of these reasons are discussed in the "Latent DIF Applications: Literature Reviews" section.

DIF Detection Procedures

In manifest DIF analysis, DIF items are detected using statistical analyses for their statistical significance and effect size (see Holland & Wainer, 1993; Millsap, 2011, for details). There are two more steps to go through in latent DIF analysis than in manifest DIF analysis. The first step is a model selection phase where one needs to find the best-fit measurement model (i.e., a mixture item response model) among several candidate models. The candidate models differ in terms of number of latent groups (i.e., number of categorical latent variables), holding the within-class model constant. The second step is to obtain a latent group membership for each person, given the selected measurement model. The following step is to detect DIF items for which we can characterize or interpret latent groups. Researchers have used manifest DIF detection procedures for detecting latent DIF after they identified latent group membership (e.g., Cho, Cohen, & Bottge, 2013; Cho, Cohen, Kim, & Bottge, 2010; Finch & Finch, 2013; Leite & Cooper, 2010; Maij-de Meij, Kelderman, & van der Flier, 2008, 2010).

Latent DIF Applications: Literature Review

In order to report how researchers used mixture item response models in practice, papers published in six journals were reviewed: *Applied Psychological Measurement* (APM), *Educational and Psychological Measurement* (EPM), *Journal of Educational and Behavioral Statistics* (JEBS), *Journal of Educational Measurement* (JEM), *Multivariate Behavioral Research* (MBR), and *Psychometrika* (PMET). The papers were searched with the keyword "mixture item response" on each journal's website. In Table 1, we selected papers in which applications of mixture item response models were used, and we excluded papers that investigated methodological issues of the models.[4]

As surveyed in Table 1, various kinds of mixture item response models have been applied to diverse research questions in educational and psychological domains. According to De Boeck et al. (2011), the applications can be categorized as follows:

(a) *no idea* reason: The mixture item response models allow the estimation of multigroup models with partially missing group membership information. See von Davier and Yamamoto (2004) for specific examples (pp. 396–398).
(b) *unobservable* reason: The mixture item response models can be used to identify and characterize unobservable groups (i.e., latent groups) that have similar item response behaviors regarding item-solving strategies or content knowledge (Bolt, Cohen, & Wollack, 2001; Mislevy & Verhelst, 1990; Rost, 1990), guessing (Meyer, 2010), test speededness (Bolt, Cohen, & Wollack, 2002), types of personality (Choi & Wilson, 2014; Leite & Cooper, 2010; Maij-de Meij et al., 2008), types of psychological symptom (Hong & Min, 2007), and change or growth patterns (Cho et al., 2010; Cho et al., 2013; Rijmen, De Boeck, & van der Maas, 2005; von Davier, Xu, & Carstensen, 2011).
(c) *validity* reason: The mixture item response models can be used to cluster heterogeneous samples (e.g., too many different manifest school types that have different students' and schools' characteristics), which may not indicate the "true" group membership characterized as similar item response behaviors. Examples include identifying homogeneous latent school groups based on students' item response patterns (Bennink, Croon, Keuning, & Vermunt, 2014; Cho & Cohen, 2010; Finch & Finch, 2013; von Davier et al., 2011).

In most applications surveyed above, the number of latent groups and latent DIF items were detected with an exploratory approach based on model fit indices. There are few applications with a confirmatory approach in which the number of latent groups and latent DIF items are hypothesized. For example, Bolt et al. (2002) hypothesized two item types (i.e., items at the beginning of the test and items at the end of the test) and two latent groups (i.e., nonspeeded group and speeded group) in using a mixture Rasch model.

Mixture item response models were also applied to purify measures (specifically, to obtain item parameter estimates from a primary dimension) in the presence of the secondary dimension due to test speededness (De Boeck et al., 2011; Suh, Cho, & Wollack, 2012) or item drift (Wollack, Cohen, & Wells, 2003). A mixture Rasch model was used in a computerized adaptive testing setting (Jiao, Macready, Liu, & Cho, 2012).

All papers surveyed in Table 1, except one, include applications of detecting person latent groups. The one exception is the application in Frederickx, Tuerlinckx, De Boeck, & Magis (2010). In this paper, a mixture model is assumed for the item difficulties such that the items may belong to different item latent groups: DIF or non-DIF latent group.

Table 1. Applications of Mixture Item Response Models

Journal	Paper	Model	Application area	Estimation/Software	DIF Detection
APM	Rost (1990)	Mixture Rasch	Physics knowledge	Conditional MLE	.
APM	von Davier & Yamamoto (2004)	Mixture generalized partial credit	Partially missing data	Marginal MLE	.
APM	Maij-de Meij et al. (2008)	Mixture nominal response; Mixture partial credit	Personality characterization and prediction	/EM	BIC, LRT
APM	Cho et al. (2010)	LTA-mixture Rasch	Growth pattern	Mplus	LRT
APM	Meyer (2010)	Mixture Rasch	Guessing behavior	WinBUGS	.
APM	De Boeck et al. (2011)	Mixture multidimensional two-parameter	Test speededness; arithmetic operations	Latent GOLD	.
APM	Jiao et al. (2012)	Mixture Rasch	Computerized adaptive test	mdltm	.
EPM	Hong & Min (2007)	Mixture rating scale	Depression	WINMIRA	.
EPM	Finch & Finch (2013)	Mixture multilevel multidimensional	Learning disability and testing accommodations	Mplus	GMH
EPM	Choi & Wilson (2014)	Mixture random weight LLTM	Verbal aggression	WinBUGS	.
JEBS	Bolt et al. (2001)	Mixture nominal response	English usage	WinBUGS	.
JEBS	Cho & Cohen (2010)	Mixture multilevel Rasch	Math aptitude	WinBUGS	HPD test
JEBS	Bennink et al. (2014)	Mixture multilevel two-parameter	Large-scale international assessment	Latent GOLD	Wald test
JEM	Bolt et al. (2002)	Mixture Rasch	Test speededness	WinBUGS	.
JEM	Wollack et al. (2003)	Mixture Rasch	Test speededness	WinBUGS	.
JEM	Cohen & Bolt (2005)	Mixture three-parameter	Math placement	WinBUGS	.
JEM	Frederickx et al. (2010)	Mixture Rasch	DIF item detection	WinBUGS	.
JEM	Suh et al. (2012)	Mixture (multidimensional) two-parameter; HYBRID	Test speededness	Latent GOLD	.
MBR	Leite & Cooper (2010)	Mixture multidimensional graded response	Social desirability	Mplus	IC
MBR	Maij-De Meij et al. (2010)	Mixture Rasch	Vocabulary	/EM	Lord test
PMET	Mislevy & Verhelst (1990)	Mixture LLTM	Solution strategy	Marginal MLE	.
PMET	Rijmen et al. (2005)	Mixture longitudinal LLTM	Conservation acquisition	Marginal MLE	.
PMET	von Davior et al. (2010)	Mixture longitudinal two-parameter	School type	mdltm	.
PMET	Cho et al. (2013)	Multilevel LTA-mixture two-parameter	Math classroom assessment	Mplus	BIC, LRT

Notes. "." in the last column indicates that DIF detection method was not applicable or specified clearly. BIC: Bayesian information criterion, LRT: likelihood ratio test method (Thissen, Steinberg, & Wainer, 1988), GMH: generalized Mantel-Haenszel, HPD: highest posterior density interval test, IC: information criteria.

Methods for Latent DIF

Mixture Item Response Models

To explain mixture item response models, the extension of a mixture Rasch model (Rost, 1990; Rost & von Davier, 1995) to a two-parameter mixture item response model is described below. For the two-parameter mixture item response model, the within-class response probability of an item endorsement can be written as

$$\text{logit}[P(y_{ji} = 1|\theta_{jg})] = \alpha_{ig}(\theta_{jg} - \beta_{ig}), \quad (1)$$

where j is an index for a person ($j = 1, \ldots, J$), i is an index for an item ($i = 1, \ldots, I$), g is a latent group or categorical latent variable ($g = 1, \ldots, G$), y_{ji} is a binary item response, θ_{jg} is a group-specific continuous latent variable (e.g., ability), α_{ig} is a group-specific item discrimination parameter, and β_{ig} is a group-specific item location parameter (e.g., difficulty parameter). When DIF involves discrimination and possibly also item location, it is labeled *nonuniform* DIF, and when it involves only the item location, it is labeled *uniform* DIF (Mellenbergh, 1982).

When latent groups are assumed to be mutually exclusive and exhaustive, the latent group structure is written as

$$P(y_{ji} = 1|\theta_j) = \sum_{g=1}^{G} \pi_g P(y_{ji} = 1|\theta_{jg}), \quad (2)$$

where π_g is the group size parameter or "mixing proportion" with constraints $0 < \pi_g < 1$ and $\sum_{g=1}^{G} \pi_g = 1$.

Related Models

Mixture item response models relax assumptions of a latent class analysis (Rost, 1990). In a latent class analysis, items are assumed to exhibit local item independence by having latent categorical variables (i.e., latent groups) on which there is no variability in the latent variable (e.g., ability) within a latent group. In mixture item response models, a continuous latent variable (i.e., θ_{jg} in Equation 1) is introduced to account for item associations within a latent group. In this way, mixture item response models permit within-group variation in the latent variable.

Multidimensional item response models and mixture item response models are two conceptually different item response theory (IRT) approaches for modeling multidimensionality in data. Multidimensional item response models are used to address the heterogeneity in item content that arises in subgroups of items. Mixture item response models address heterogeneity within subgroups of persons. Although the two modeling approaches have different assumptions about the causes of multidimensionality in data, they are closely related and often end up drawing the same conclusion on the data. For example, Rijmen and De Boeck (2005) found that the between-item multidimensional model is formally equivalent to a mixture Rasch model (Rost, 1990).

Model Extensions

As shown in Table 1 and in other publications (not included in Table 1), simple mixture response models, such as a mixture Rasch model and a two-parameter mixture item response model, have been extended for more complex mixture item response models to take data type or complexity into account: three-parameter (Cohen & Bolt, 2005), polytomous response (Bolt et al., 2001; Hong & Min, 2007; Leite & Cooper, 2010; Maij-de Meij et al., 2008; Rost & von Davier, 1995), linear logistic test model (LLTM; Choi & Wilson, 2014; Mislevy & Verhelst, 1990), multilevel (Bennink et al., 2014; Cho & Cohen, 2010; Vermunt, 2008), multidimensional (De Boeck et al., 2011; Leite & Cooper, 2010), longitudinal (Cho et al., 2010; Kadengye, Ceulemans, & Van den Noortgate, 2014; Rijmen et al., 2005), multilevel longitudinal (Cho et al., 2013; von Davier et al., 2011), and multidimensional multilevel (Finch & Finch, 2013).

In addition to these extensions that have different item parameters across latent groups, other kinds of mixture item response models were suggested. Examples include the HYBRID model (Yamamoto & Everson, 1997) in which the first component of the model is an item response model and the second component of the model is a latent class model and saltus model (Wilson, 1989) where there are different latent groups or developmental stages and a subset of items increases or decreases in difficulty from one latent group to the other.

Parameter Estimation and Software

Estimation Methods

Estimation methods of mixture item response models can be categorized into maximum likelihood estimation (MLE) and Bayesian analysis using Markov chain Monte Carlo (MCMC). Different kinds of MLE have been developed for mixture item response models, including conditional MLE (Rost, 1990; von Davier, 2001), joint MLE (Willse, 2009), and marginal MLE (Mislevy & Wilson, 1996; von Davier & Yamamoto, 2004). Implementation of Bayesian analysis for mixture item response models was described in Boughton and Yamamoto (2007) and in several studies using WinBUGS (e.g., Cho, Cohen, & Kim, 2013; Dai, 2013). Finch and French (2012) compared the performance of marginal MLE and Bayesian analysis implemented in Mplus. See Finch and French (2012) for details of the results.

In a conditional MLE, the parameters to be estimated include mixing proportion and group-specific item parameters. In a joint MLE, the parameters to be estimated include mixing proportion, group-specific item parameters, and group-specific continuous latent variables. In a marginal MLE method, the parameters to be estimated are mixing proportion, group-specific item parameters, and population parameters of group-specific continuous latent variables (e.g., mean and variance in a normally distributed continuous latent variable). In MLE methods, for a continuous latent variable (θ_{jg}), scoring methods developed for regular item response models can be applied for mixture item response models (see Baker & Kim, 2004, for details.). For a categorical latent variable (g), persons are assigned to one of categorical latent variables (i.e., latent groups) that have the highest posterior probability of membership. For example, in a marginal MLE method, given the item parameter MLE estimates (e.g., $\widehat{b}_{ig}$ in a mixture Rasch model), predicted continuous latent variable scores ($\tilde{\theta}_{jg}$), and item responses ($\mathbf{y} = [y_{j1}, \ldots, y_{ji}, \ldots, y_{jI}]'$) for each person, the posterior probability of belonging to

each latent group, P_{jg}, is calculated using the following equation:

$$P_{jg} = \frac{\pi_g \cdot \prod_{i=1}^{I}(P_{jig} = 1|\tilde{\theta}_{jg})^{y_{ji}}\,[1 - (P_{jig} = 1|\tilde{\theta}_{jg})^{1-y_{ji}}]}{\sum_{g=1}^{G} \pi_g \cdot \prod_{i=1}^{I}(P_{jig} = 1|\tilde{\theta}_{jg})^{y_{ji}}\,[1 - (P_{jig} = 1|\tilde{\theta}_{jg})^{1-y_{ji}}]}. \quad (3)$$

For each person, $\sum_{g=1}^{G} P_{jg} = 1$. The following example shows the posterior probabilities of memberships and assigned latent group membership with two latent groups for five persons:

ID	P_{j1}	P_{j2}	g
1	1	0	1
2	.477	.523	2
3	.992	.008	1
4	.895	.105	1
5	.062	.938	2

If P_{jg} is close to 1 (e.g., Person ID = 1, 3, 5), then classification of that person into a latent group would be made with small uncertainty. Large uncertainty exists for Person ID 2. A measure of uncertainty in classifying a person j into the latent group with the largest probability is given by $e_j = 1 - max P_{jg}$ (e.g., Duda, Hart, & Stork, 2001). The range of e_j is from 0 to .5. The smaller the value of e_j, the better the person is classified.

In Bayesian analysis using MCMC, all parameters (e.g., for a mixture Rasch model, π_g, g_j; $g_j = 1, \ldots, G$, where G is given), b_{ig}, and θ_{jg}, are sampled simultaneously (see Kim & Bolt 2007 for an introduction to MCMC in item response models). The posterior probability of a person in a latent group can be calculated as the relative frequency of being sampled into a latent group using post-burn-in iterations, and persons were assigned to one of the latent groups that have the highest relative frequency of membership.

Issues in Estimation for Mixture Item Response Models

The primary modeling framework of mixture item response models is a mixture model. There are well-known estimation problems in mixture modeling, including label switching and local maxima (Congdon, 2003; Frühwirth-Schnatter, 2006; McLachlan & Peel, 2000; Vermunt & Magidson, 2005). Below, several selected issues that affect the estimation of the mixture item response models are summarized.

Because the likelihood function is invariant regarding different permutations of model parameters in the mixture modeling, there is a problem called label switching. Two types of label switching occur with mixture modeling (Cho, Cohen, & Kim, 2013). The first type occurs across iterations within a single chain in the Bayesian solution. It can be a serious problem in Bayesian estimation because the labels of the latent groups can change within the MCMC chain on different iterations. That is, the meaning of the latent classes can simply switch at each iteration. If multiple modes exist for any of the distributions of parameters, then label switching may present such that the interpretation of the posterior moments is distorted. See Dai (2013) for a summary of label switching treatment in the mixture modeling literature.

The second type of label switching occurs in different runs in Bayesian analysis and MLE. The second type of label switching in MLE and Bayesian analysis is not problematic if it is detected and treated appropriately. The result of this form of label switching may cause confusion since the latent groups have a different order in different runs. Examples of different runs include replications in simulation studies, separate runs for the same empirical data sets, and separate runs to check convergence or to monitor local solution. This type of label switching can be detected when item profiles and other parameter estimates are compared. By setting one of the runs as a reference run, one can relabel the remaining runs to be consistent with the reference run. See Self-Test Item 3 as an example.

Another problem is that the estimation of likelihoods of mixture models, in general, is apt to yield multiple local maxima (Frühwirth-Schnatter, 2006; McLachlan & Peel, 2000). A typical empirical method used for checking whether a local solution is obtained is to run the model with multiple different starting values (McLachlan & Peel, 2000). Observing the same log likelihood from a multiple set of starting values increases confidence that the solution is not local.

Software

As surveyed in Table 1, the following software has been used for mixture item response modeling: *l*tm (Vermunt, 1997), Latent GOLD (Vermunt & Magidson, 2005), mdltm (von Davier, 2005), Mplus (Muthén & Muthén, 1998–2013), WinBUGS (Spiegelhalter, Thomas, & Best, 2000), and WINMIRA (von Davier, 2001). In addition, mixRasch (Willse, 2009) is available as an R package.

In WINMIRA and mixRasch, mixture Rasch models can be implemented. *l*tm and Latent GOLD were developed mainly for mixture modeling including mixture item response models. (See http://statisticalinnovations.com/products/LGIRT.pdf for mixture item response modeling in Latent GOLD.) The mdltm was developed for a general class of models called a general diagnostic model including mixture item response models (von Davier, 2005, 2008; von Davier & Yamamoto, 2004). (See http://208.76.84.140/ svfklumu/pdfs/Software_Description_mdltm_Mvd_XXu_final.pdf for a brief mdltm description.) Mplus was developed for generalized latent variable modeling including mixture item response modeling. (See Example 7.27 for mixture item response modeling in the Mplus manual, which can be downloaded from http://www.statmodel.com/download/usersguide/.) WinBUGS is flexible software with which many different kinds of mixture item response models can be specified. (See the papers that used WinBUGS in Table 1 on its uses for the mixture item response modeling.)

*l*tm, Latent GOLD, and mdltm implement marginal MLE. WINMIRA and mixRasch implement conditional MLE and joint MLE, respectively, for mixture Rasch modeling. In Mplus, marginal MLE (ML or MLR estimation option in Mplus) and Bayesian analysis using MCMC (Bayes estimation option in Mplus) can be implemented for mixture item response modeling. WinBUGS has been used for Bayesian analysis.

Model Selection

As shown in Table 1, different mixture item response models can be selected for latent DIF analysis for the number of item parameters (i.e., Rasch or one-parameter, two-parameter, three-parameter), number of latent groups (i.e., categorical latent variables), and number of dimensions (i.e., continuous

latent variables). Selecting a particular mixture item response model requires decisions about several aspects.

Refer to de Ayala (2009) for IRT model fit to decide the number of item parameters in item response models using IRT model fit indices, as an example. Regarding the number of latent groups and dimensions, several candidate models, each with a different number of latent groups and dimensions, are considered, and a particular model can be selected according to a theoretical rationale that uses one or more statistical criteria. In mixture item response model applications, information criteria have been widely used to select the model. For example, Lubke and Neale (2008) used Akiaike information criterion (AIC; Akiaike, 1974), Bayesian information criterion (BIC; Schwarz, 1978), sample size adjusted BIC (saBIC; Sclove, 1987), and consistent Akaike information (CAIC; Bozdogan, 1987) in selecting a model for the number of latent groups and dimensions. Given one continuous latent variable within a latent group, a few studies investigated the performance of model information criteria in selecting the number of latent groups (e.g., Li, Cohen, Kim, & Cho, 2009; Preinerstorfer & Formann, 2012). See Li et al. (2009) and Preinerstorfer and Formann (2012) for detailed results.

Latent DIF Analysis

Scale Comparability

In mixture item response modeling, latent groups are commonly characterized by comparing item profiles across the groups. For the comparison, item parameter estimates across the latent groups should be on the same scale (Paek & Cho, 2014; von Davier & Yamamoto, 2004). See Paek and Cho (2014) for examples of the consequences of ignoring scale linkage in mixture item response models. IRT linking methods developed for manifest DIF analysis (Thissen, Steinberg, & Wainer, 1993) can be applied to latent DIF analysis to achieve the scale linkage across the latent groups. For example, an iterative procedure for item purification (Lord, 1980) can be used to detect group-invariance items based on mixture item response models that have the determined latent group memberships. After finding the group-invariance items (i.e., anchor items), equality constraints across the latent groups can be imposed on these items in estimating mixture item response models. The use of the group-invariant items for establishing a common metric is similar to the use of anchor items in a nonequivalent group with anchor test (NEAT) design, although in mixture item response models group-specific item parameters as well as latent group membership are estimated simultaneously.

It should be noted that using the group-invariant items for the scale comparability may decrease the unique item profiles across the latent groups. Consequently, the selected number of groups based on model-fit analysis can be "distorted" due to the group-invariant items. Cho and Paek (2014)[5] suggested two criteria for selecting the group-invariant items in mixture item response model applications with the assumption that group-invariant items do not help the detection of group membership: (a) model comparison criterion (model fit with group-invariant items should be as good as model fit without group-invariant items) and (b) group membership criterion (group membership should not change between two models with and without group-invariant items).

DIF Detection Methods

As shown in Table 1, latent DIF can be detected using DIF detection procedures developed for manifest DIF. See Holland and Wainer (1993) and Millsap (2011) for these procedures.

DIF Explanation Methods

Latent DIF can be explained by interpreting item profiles across latent groups and assessing to what extent the latent groups can be linked to person characteristics such as demographic information.

When item profiles are investigated in an exploratory approach, referring to item-level information is recommended. We illustrate how item types can be considered in interpreting latent groups in our empirical illustration (Step 3) and in Self-Test Item 2.

In addition to interpreting item profiles using item-level information to explain DIF, person information can be used to explain DIF by mapping latent group membership to person characteristics. After latent group membership is obtained, one can test the association between the latent group membership and categorical person characteristics using chi-square statistic and test group mean differences in continuous person characteristics using t or F tests (i.e., two-step procedure). Alternatively, the mixing proportion (π_g in Equation 2) can be explained or predicted by modeling person characteristics as predictors in a (multinomial) logitic regression model (see for more details, Cho, Cohen, & Kim, 2013; Dai, 2013; Dayton & Macready, 1988; Smit, Kelderman, & van der Flier, 2000; Vermunt & Magidson, 2005).

Recommended Analysis Steps

We recommend the following five steps for conducting latent DIF analysis: Step 0: Explain why latent DIF is considered (instead of manifest DIF) for research questions and hypotheses based on a substantive theory, if any; Step 1: Select the "best" measurement model for the numbers of item parameters, latent groups, and dimensions; Step 2: Detect DIF items; Step 3: Explain DIF using item information and person characteristics; and Step 4: Discuss the treatment of (detected and explained) DIF item(s).

Application

Latent DIF using mixture item response models is applied to an empirical data set, following our recommended steps for latent DIF analysis. A verbal aggression data set (Vansteelandt, 2000) was selected because it is available to freely download from https://bearcenter.berkeley.edu/page/materials-explanatory-item-response-models, has well-structured item designs, and has been widely used for item response modeling (e.g., De Boeck & Wilson, 2004), including mixture item response modeling (Choi & Wilson, 2014; Fieuws, Spiessens, & Draney, 2004; Rijmen & De Boeck, 2005).

Data Description

Among the data sets provided at https://bearcenter.berkeley.edu/page/materials-explanatory-item-response-models, the file "data verbal aggression matrix dichot.txt", was used in the current application and was renamed "vadata.txt" for the analyses. For a detailed description of the data, refer to pp. 7–10 in De Boeck and Wilson (2004). Briefly, the data set has binary responses (i.e., yes = 1, no = 0) from 24 items and 316 persons. The item design is a 2 (modes [want vs. do]) × 2 (situation types [other-to-blame vs. self-to-blame]) × 3

(behavior types [curse vs. scold vs. shout]) design with two specific situations in each cell, leading to 24 items in total. As person characteristics, gender information (243 women [coded as 0] and 73 men [coded as 1]) and the trait anger scores derived from a personality inventory (mean = 20, $SD = 4.85$) are included in the data set.

Analysis

For all analyses except DIF detection, Mplus was chosen over other software because of its flexibility in modeling; for example, the DIF magnitude significance test (Wald test) in Step 2 and multidimensional mixture modeling in Step 4. Marginal MLE implemented in Mplus (MLR estimator in Mplus) was selected for analyses. To monitor local maxima in mixture item response modeling, "STARTS = 100 10"; was set in Mplus, which indicates that 100 random sets of starting values for the initial stage and 10 final stage optimizations were used. The difR function (Magis, Béland, & Raîche, 2013) in R was chosen to detect DIF items due to its flexibility of having different kinds of IRT DIF detection methods.

Results

Results are reported for each step as follows:

Step 0: *Explain why latent DIF is considered (instead of manifest DIF) for research questions and hypotheses based on a substantive theory, if any*. Previously, Meulders and Xie (2004) detected DIF items for gender. They found that there was DIF for cursing or scolding in each of the situation using a Rasch DIF model. In this application, latent DIF analysis was chosen with "validity" reason. That is, we are interested in the degree of overlap between gender group and latent groups. The latent groups can be defined as the types of persons who may be distinct in terms of how they responded to items that reflect different modes, situation types, and behavior types.

Step 1: *Select the best measurement model regarding the numbers of item parameters, latent groups, and dimensions*. Table 2 presents the model fit results and the values of the information criteria. For the comparison between the (one-group) Rasch and two-parameter item response models, the Rasch model fits better than the two-parameter item response model based on information criteria, but does not fit better with likelihood ratio test (LRT) (chi-square value = 40.96, $df = 23, p = .012$). However, correlations between estimates from the two models are .998 and .996 for the item location parameters and person scores, respectively. This indicates that the relative ordering of the item location and predicted person scores does not change when item discrimination parameters are introduced. There was a warning message for the mixture Rasch model with a three-group solution in Mplus. Referring to the warning message, there was a data sparseness problem (i.e., empty cells in the joint distribution of the item responses). As a result, the mixture Rasch model with a two-group solution was selected because it fits better than the one-group Rasch model and there was evidence that introducing item discrimination is not necessary. For the selected mixture Rasch model, there was evidence consistent with no local solution indicated "THE BEST LOGLIKELIHOOD VALUE HAS BEEN REPLICATED" in Mplus. The Mplus syntax to fit a mixture Rasch model with a two-group solution is attached in Appendix A.

Step 2: *Detect DIF items*. Given the assigned group membership for each person from the mixture Rasch model with a two-group solution, we use the difR function to detect DIF items based on Lord's test, Raju's test, and LRT. The R code for DIF analysis is as follows:

```
library(difR)
data <- read.table(''C:\\vadata_dif.txt'',
header=T,fill=T)
difLord(data, group=''gender'',focal.
name=1,model=''1PL'',c=NULL,engine
=''ltm'', purify=TRUE,
nrIter=10)
difRaju(data, group=''group'',focal.
name=1,model=''1PL'',purify=TRUE,
nrIter=10)
difLRT(data, group=''group'',focal.name=1,
purify=TRUE,nrIter=10)
```

In the R code, "group" indicates the assigned latent group membership from Step 1. Three DIF detection methods show that the following items are group-invariant items: Items 11, 14, 16, 17, and 22 according to Lord's test, Items 11, 14, 16, 17, and 22 according to Raju's test, and Items 9, 11, 14, 16, and 22 according to LRT. Items were designated as group-invariance items (anchor items) if they were not detected by at least two DIF detection methods. With this criterion, Items 11, 14, 16, 17, and 22 were considered group-invariant items, and the rest of the items were tested using the Wald test for latent DIF with a two-group mixture Rasch model using Mplus (see Appendix A). Model fit with the group-invariant items was as good as the model without group-invariant items (as shown in Table 2), and there was 99% consistency in the assigned group membership between the two models (following the two criteria described in Cho & Paek, 2014). For the group-variant items (i.e., DIF candidate items), there were statistically significant differences between the item location estimates of the two latent groups except Items 9 and 21 (as indicated by * in DIF magnitudes in Table 3) using a Wald test. The detected DIF items were used to characterize latent groups in Step 3.

Step 3: *Explain DIF using item information and person characteristics*. Table 3 shows the item design of the verbal aggression test and item location estimates. Referring to the item design, the item location estimates for Latent Group 2 were higher than those of Latent Group 1 for all "Shout" items, whereas the item location estimate for Latent Group 1 was higher than those of Latent Group 2 for all "Curse" and "Scold" items. There was one exception for this location pattern for Item 20 ("Scold" Item) for which the DIF magnitude is barely statistically significant based on Wald test. This item location pattern indicates that persons in Latent Group 1 are likely to show "Shout" behavior, whereas persons in Latent Group 2 are likely to show "Curse" and "Scold" behaviors.

According to the proportions for latent groups based on the estimated model, 47% of the persons were classified into Latent Group 1, and 53% of the persons were classified into Latent Group 2 (i.e., $\widehat{\pi}_1 = 0.47$, $\widehat{\pi}_2 = 0.53$). The same proportion was found for the latent group patterns based on classification of individuals using their most likely latent class membership. In marginal MLE, a group-specific continuous latent variable, θ_{jg}, is assumed to follow a normal (N) distribution. The group-specific continuous latent variable for Latent Group 1, θ_{j1}, follows $N(0^*, 1.422^2)$ (the mean was set to 0 to identify the model, as indicated *). The group-specific continuous latent variable for Latent Group 2, θ_{j2}, follows $N(-0.427, 1.629^2)$. The group difference between the two

Table 2. Model Selection

Name	Model Description		Number of Parameters	Log-Likelihood	Information Criteria		
	Number of Dim.	Number of Groups			AIC	BIC	saBIC
Rasch	1	1	25	−4,036.91	8,123.81	8,217.71	8,138.41
Mixture Rasch	1	2	51	−3,925.10	7,954.20	8,143.75	7,981.99
Mixture Rasch*	1	3					
Two-par.	1	1	48	−4,016.43	8,128.86	8,309.14	8,156.89
Mixture Two-par.*	1	2					
Mixture Rasch**	1	2	47	−3,925.73	7,945.46	8,121.98	7,972.91

Note. *The model with convergence problems. **The model with group-invariant items. saBIC indicates a sample-size adjusted BIC.

Table 3. Item Design of 24 Verbal Aggression Items and Item Location Estimates (Standard Errors (SE))

Item	Item Design			Estimates (SE)		DIF	Purified
	Mode	Simulation Type	Behavior	Latent Group 1	Latent Group 2	Magnitude (SE)	Estimates (SE)
1	Want	Other	Curse	−.420 (.289)	−3.099 (.579)	2.679* (.617)	−1.190 (.324)
2	Want	Other	Scold	−.173 (.293)	−1.558 (.390)	1.384* (.470)	-.510 (.313)
3	Want	Other	Shout	−.781 (.314)	.330 (.471)	−1.112* (.599)	-.753 (.279)
4	Want	Other	Curse	−1.345 (.352)	−2.947 (.489)	1.602* (.665)	−1.738 (.322)
5	Want	Other	Scold	−.407 (.345)	−1.590 (.412)	1.184* (.584)	−.658 (.313)
6	Want	Other	Shout	−.876 (.312)	.614 (.524)	−1.490* (.628)	-.665 (.282)
7	Want	Self	Curse	.314 (.333)	−2.180 (.541)	2.494* (.671)	-.473 (.300)
8	Want	Self	Scold	1.339 (.281)	−.388 (.436)	1.726* (.506)	.790 (.318)
9	Want	Self	Shout	1.193 (.321)	1.691 (.552)	−.498 (.722)	1.210 (.262)
10	Want	Self	Curse	−.564 (.307)	−2.352 (.430)	1.787* (.534)	-1.046 (.313)
11	Want	Self	Scold	.169 (.231)	.169 (.231)	.000	.439 (.313)
12	Want	Self	Shout	.083 (.287)	2.332 (.793)	−2.249* (.876)	.639 (.255)
13	Do	Other	Curse	−1.054 (.266)	−1.976 (.350)	.923* (.370)	−1.190 (.311)
14	Do	Other	Scold	−.603 (.238)	−.603 (.238)	.000	-.328 (.316)
15	Do	Other	Shout	.110 (.305)	1.641 (.448)	−1.531* (.535)	.431 (.263)
16	Do	Other	Curse	−1.108 (.238)	−1.108 (.238)	.000	-.829 (.295)
17	Do	Other	Scold	−.136 (.236)	−.136 (.236)	.000	.135 (.303)
18	Do	Other	Shout	.729 (.294)	2.478 (.476)	−1.749* (.538)	1.157 (.274)
19	Do	Self	Curse	.691 (.263)	−.768 (.447)	1.459* (.542)	.295 (.302)
20	Do	Self	Scold	1.915 (.312)	.779 (.454)	1.136* (.577)	1.645 (.324)
21	Do	Self	Shout	2.515 (.363)	3.652 (.733)	−1.136 (.907)	2.911 (.323)
22	Do	Self	Curse	−.936 (.238)	−.936 (.238)	.000	−.658 (.295)
23	Do	Self	Scold	−.045 (.279)	.529 (.308)	−.574* (.322)	.475 (.305)
24	Do	Self	Shout	1.066 (.270)	4.142 (1.631)	−3.076* (1.660)	1.764 (.280)

Note. *Statistical significance at $\alpha = .05$.

latent groups (called impact in DIF literature), −0.427, was not statistically significant ($SE = .373$, z-value $= -1.146$, $p = .252$). The mean of a measure of uncertainty (e_j) across persons was .10 for both models with and without person characteristics, which indicates that the classification is satisfactory.

The association between gender and the assigned latent group membership was not statistically significant (Pearson's chi-square statistic = .0001, $p = .991$). According to the t-test with unequal variance, there was no difference between the two latent groups in trait anger scores. In a multinomial logistic regression model to explain or predict mixing proportion, π_g, gender and trait anger scores were used as predictors. Type 2 label switching occurred between two runs, mixture Rasch models without and with person characteristics (i.e., using the logistic regression model). See Self-Test Item 3 and its answer for the detail and its treatment. As the results of the logistic regression model, the effect of the trait anger scores was barely significant at the $\alpha = .05$ (estimate= .068 [after relabeling due to Type 2 label switching] with Latent Group 2 as the base group, $SE = .033$, $p = .040$), which indicates that the ratio of the odds of being in Latent Group 2 versus Latent Group 1 is $exp(0.068) = 1.070$ when one unit of the trait anger scores changes. The effect of gender was not significant (estimate= .033 [after relabeling due to Type 2 label switching] with Latent Group 2 as the base group, $SE = .588$, $p = .956$). After relabeling due to label switching, there was 97% consistency in latent group membership assignment between the two mixture Rasch models with and without modeling person characteristics.

Step 4: *Discuss the treatment of (detected and explained) DIF item(s)*. In the application here, we modeled DIF to obtain the "purified" primary dimension. Evidence of secondary dimension (i.e., two latent groups) was found using the mixture Rasch model. Assume that one may be interested in extracting the purified target dimension separated from the secondary dimension due to the "Shout" items. Following an approach described in De Boeck et al. (2011), item parameter estimates from the purified dimension were obtained using a (confirmatory) multidimensional mixture item response model

(specifically, the multidimensional mixture Rasch model in our application). In the model, the reference latent group has a primary dimension as an interaction between all items and persons in that group, whereas the focal latent group has two dimensions where a primary dimension is applied for all items and the nuisance or secondary dimension is expressed as an interaction between the "Shout" items (i.e., DIF items) and persons. Here, the DIF magnitudes for "Shout" items are explained by the nuisance dimension. Item estimates from the reference latent group are reported in Table 3, as purified item estimates. Mplus code to fit the (confirmatory) multidimensional mixture item response model is shown in Appendix B.

Concluding Remarks

Mixture item response models have been applied to various problems in education and psychology. This module has described and illustrated how to use mixture item response models for latent DIF analysis. The latent DIF approach is different from the manifest DIF approach in that the grouping variable for DIF analysis is not known a priori. We also showed that the DIF detection methods developed for manifest DIF can be applied for latent DIF after latent group membership is identified. The following includes concluding remarks for latent DIF analysis and mixture item response models.

Uses of Latent DIF

It is more common to use manifest DIF analysis than to use latent DIF analysis. We expect that researchers may need to justify to test users why the latent DIF analysis is implemented over the manifest DIF analysis. For this justification, one may refer to the four a priori reasons for using the latent DIF approach over the manifest DIF approach (see the subsection of "Manifest DIF vs. Latent DIF" for the four reasons). In our empirical example, latent DIF analysis was chosen over manifest DIF analysis in the application for validity purposes. We found that there was no evidence of an association between gender and the latent group memberships. This result indicates that gender is not homogeneous in nature in verbal aggression contexts.

As we described earlier regarding our analysis purpose, the main advantage of the latent DIF approach is the ability to investigate DIF based on the homogeneous groups detected from the data and to explain DIF by understanding the homogenous groups using person and item characteristics. In this regard, it may be more important to obtain person and item information in the latent DIF approach than in the manifest DIF approach. However, in practice, it is not always the case that such information is available. For example, Cho and Cohen (2010) could not use item information because the test content was not available for analysis on account of a test security issue. Even if item information is available, one should check if such information is reliable and valid before using the information.

When policy and legislation dictate DIF analyses, groups for DIF are determined by the policy or legislation and are often manifest, such as gender and ethnicity groups (e.g., Zumbo, 2007). The discrepancy between manifest and latent DIF results increases as the overlap between the manifest and latent group memberships decreases. DIF explanation and treatment based on latent DIF analysis can be challenging for informing policy or legislation when the results from the two analyses differ.

Latent DIF Detection Using Manifest DIF Detection Methods

DIF items in latent DIF applications can be detected using manifest DIF detection methods after the latent group memberships are obtained. However, the number of DIF items detected with latent DIF analysis can be larger than what one would expect from manifest DIF analysis. This occurs because mixture modeling maximizes differences among latent groups, resulting in a higher number of DIF items and DIF magnitudes among latent groups (Samuelsen, 2005). This result was also consistent with previous research using latent DIF analysis (e.g., Cohen & Bolt, 2005; Cohen, Gregg, & Deng, 2005). With this finding, the performance of manifest DIF detection methods in the context of latent DIF is questioned (e.g., more DIF items, high magnitudes). For example, the DIF detection procedures may not perform well when there is a large portion of DIF items and high magnitudes of DIF like in latent DIF analysis (Stark, Chernyshenko, Drasgow & Williams, 2006; Wang & Yeh, 2003). Future research is required to investigate the performance of manifest DIF detection methods for studying latent DIF analysis.

Other Uses of Mixture Item Response Models

In this ITEMS module, we restrict the use of mixture item responses models to latent DIF analysis when a test is unintentionally multidimensional. However, this does not mean that the mixture item response models cannot be used for intentionally multidimensional tests. Multidimensionality to be measured can be presented by different item profiles across latent groups in mixture item response models. In contrast, multidimensionality is presented by different person score profiles across item types in multidimensional item response models. Thus, compared to multidimensional item response models, it may be useful to use mixture item response models to deal with multidimensionality when research questions of interest involve person classification.

Recommended Reading

The following edited books are recommended for reviews and applications of mixture item response models: von Davier and Carstensen (2007) and Hancock and Samuelsen (2007). Heinen (1996), and Sterba (2013) are recommended for understanding mixture item response models with other modeling frameworks. Specifically, Heinen (1996) discusses differences and similarities between mixture modeling and item response modeling, with which the features of mixture item response models can be evident as the combined two modeling frameworks. Sterba (2013) presents an integrative understanding of shared aspects of different kinds of mixture models including mixture item response modeling.

Self-Test

1. (Manifest DIF vs. Latent DIF Analysis) Contrast manifest DIF analysis with latent DIF analysis regarding analysis purpose and DIF procedures.
2. (Latent Group Characterization) Characterize latent groups by referring to the following Q-matrix and item profiles: Taking an example from Cho et al. (2010), a math test that measures computation ability has the item design summarized as a form of Q-matrix (Tatsuoka, 1983). The following table is reproduced

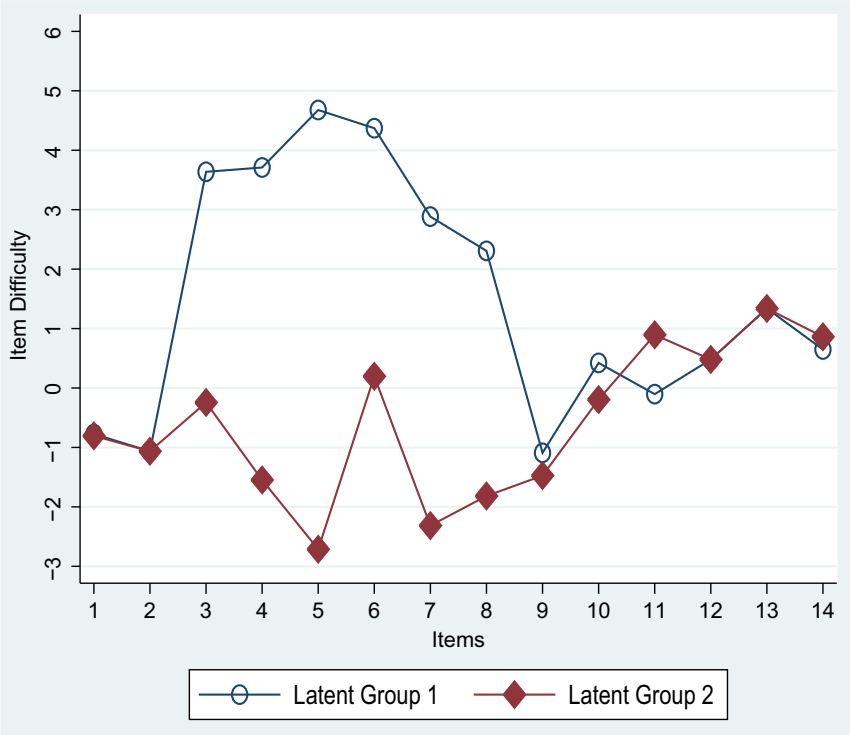

FIGURE 1. Item profiles. (Adapted from Cho et al., 2010.).

from Table 7 in Cho et al. (2010) and shows a Q-matrix, where 1 indicates the particular mathematics skills that are required to solve each item and 0 otherwise.

	Cognitive Skill		
Item	Number Operation	Measurement	Representation
1	1	0	0
2	1	0	0
3	1	1	0
4	1	1	0
5	1	1	0
6	1	1	0
7	1	1	0
8	1	1	0
9	0	1	0
10	1	1	1
11	1	1	1
12	1	0	0
13	1	0	0
14	1	0	0

Cho et al. (2010) found two latent groups based on BIC with mixture Rasch models having the different number of latent groups. They showed item profiles between the two latent groups with a figure where the x-axis is for item numbering and y-axis is for item difficulty estimates. Figure 1 is adapted from Figure 1 in Cho et al. (2010). It presents item profiles for the two latent groups with Items 2, 12, and 13 as group-invariant items:

3. (Label Switching) Two separate runs were considered to interpret latent groups using person characteristics, gender, and trait anger scores, in the verbal aggression data. The first run was to fit a mixture Rasch model without the person characteristics. The second run was to fit a mixture Rasch model with a logistic regression model with person characteristics. Item parameter estimates for the two runs were as follows:

	Item Design			Run 1: Estimates		Run 2: Estimates	
Item	Mode	Simulation Type	Behavior	Latent Group 1	Latent Group 2	Latent Group 1	Latent Group 2
1	Want	Other	Curse	−.490	−3.046	−2.989	−.495
2	Want	Other	Scold	−.243	−1.505	−1.508	−.237
3	Want	Other	Shout	−.851	.383	.362	−.859
4	Want	Other	Curse	−1.415	−2.894	−2.845	−1.428
5	Want	Other	Scold	−.477	−1.537	−1.520	−.491
6	Want	Other	Shout	−.946	.667	.659	−.963
7	Want	Self	Curse	.244	−2.127	−2.081	.227
8	Want	Self	Scold	1.269	−.335	−.359	1.277
9	Want	Self	Shout	1.123	1.744	1.670	1.163
10	Want	Self	Curse	−.634	−2.299	−2.249	−.656
11	Want	Self	Scold	.099	.222	.216	.091
12	Want	Self	Shout	.013	2.385	2.399	.01
13	Do	Other	Curse	−1.124	−1.923	−1.907	−1.138
14	Do	Other	Scold	−.673	−.550	−.556	−.681
15	Do	Other	Shout	.040	1.694	1.696	.032
16	Do	Other	Curse	−1.178	−1.055	−1.062	−1.187
17	Do	Other	Scold	−.206	−.083	−.090	−.215
18	Do	Other	Shout	.659	2.531	2.577	.647
19	Do	Self	Curse	.621	−.715	−.746	.642
20	Do	Self	Scold	1.845	.832	.771	1.881
21	Do	Self	Shout	2.445	3.705	3.565	2.505
22	Do	Self	Curse	−1.006	−.883	−.889	−1.014
23	Do	Self	Scold	−.115	.582	.547	−.103
24	Do	Self	Shout	.996	4.195	4.346	1.002

For comparisons of the estimates between the two runs and between the latent groups, group-specific item estimates were equated using a mean-sigma method. Item parameter estimates were equated based on the estimates reported in Table 3.

Explain why different item profiles were obtained between the two different runs and discuss the possible solution for dealing with this problem.

4. (Scale Comparability) Rost (1990) used the model constraint $\sum_i^I \beta_{ig} = 0$, indicating the summation of group-specific item location parameters over items is 0 within a latent group, and compared item profiles to characterize latent groups. Discuss if $\sum_i^I \beta_{ig} = 0$ is sufficient to establish a common scale across the latent groups for real data analyses.
5. (Parameter Estimation) Assume that there are 15 group-variant items and 5 group-invariant items and the first latent group (Latent Group 1) is set to be a reference latent group. Count the total number of parameters to be estimated for each model below in using a marginal MLE method where a group-specific continuous latent variable is assumed to follow a normal distribution. In addition to constraints for scale comparability, consider model identification constraints, too.

(1) Two-group unidimensional mixture Rasch model,
(2) Three-group unidimensional two-parameter mixture item response model,
(3) Two-group unidimensional three-parameter mixture item response model.

Self-Test Answers

1. The purpose of manifest DIF analysis is to identify items that function differently (i.e., to detect DIF items) across manifest groups being compared, whereas the purpose of latent DIF analysis is not only to detect DIF items but also to explain the potential causes of DIF using item information and person characteristics. Manifest DIF procedures for detecting DIF can be used in latent DIF analysis. However, latent DIF analysis requires additional steps: determining the number of latent groups and assigning latent group membership to each person.
2. Item 1 and Items 3 to 10 were clearly more difficult for the members of Latent Group 1, and Items 11 and 14 were more difficult for the members of Latent Group 2. As shown in Q-matrix, Items 3 through 8, 10, and 11 required two or more different skills for the correct answer, and thus they were considered complex items. Items 1, 2, 9, 12, 13, and 14 were not considered complex items as they required only a single skill for the correct answer. Accordingly, Latent Group 1 can be characterized as a "One-Skill Group," whereas Latent Group 2 can be characterized as a "Two-Skill Group."
3. The item profiles of Latent Group 1 in Run 1 are similar to those of Latent Group 2 in Run 2, whereas the item profiles of Latent Group 2 in Run 1 are similar to those of Latent Group 1 in Run 2. It is suspected label switching occurs across runs (Type 2 label switching). For the comparability across runs, one of the runs can be set as a base run, and the labels in another run can be relabeled. Specifically, setting Run 1 as a base run, Latent Group 2 in Run 2 can be relabelled as Latent Group 1 in Run 2 and Latent Group 1 in Run 2 can be relabeled as Latent Group 2 in Run 2.
4. The constraint can be used for scale comparability only when there is no mean difference in a continuous latent variable. Because we do not know the "true" mean difference in real data sets, the constraint cannot be sufficient for all empirical cases. Refer to Paek and Cho (2014) for a detailed discussion and examples.
5. (1) The mean of a continuous latent variable for Latent Group 1 is set to 0 for the model identification. The total number of parameters is 39: 1 (2 latent groups − 1) mixing proportion, 35 location parameters (30 [=15 × 2] group-variant item location parameters + 5 group-invariant item location parameters), 1 mean of a continuous latent variable for Latent Group 2, and 2 variances of continuous latent variables.

(2) The mean and variance of a continuous latent variable for Latent Group 1 are set to 0 and 1, respectively, for model identification. The total number of parameters is 106: 2 (3 latent groups − 1) mixing proportion, 50 location parameters (45 [=15 × 3] group-variant item location parameters + 5 group-invariant item location parameters), 50 discrimination parameters (45 [=15 × 3] group-variant item discrimination parameters + 5 group-invariant item discrimination parameters), 2 means of continuous latent variables for Latent Groups 2 and 3, and 2 variances of continuous latent variables for Latent Groups 2 and 3.

(3) The mean and variance of a continuous latent variable for Latent Group 1 are set to 0 and 1, respectively, for model identification. The total number of parameters is 108: 1 (2 latent groups − 1) mixing proportion, 35 location parameters (30 [=15 × 2] group-variant item location parameters + 5 group-invariant item location parameters), 35 discrimination parameters (30 [=15 × 2] group-variant item discrimination parameters + 5 group-invariant item discrimination parameters), 35 guessing parameters (30 [=15 × 2] group-variant item guessing parameters + 5 group-invariant item guessing parameters), 1 mean of a continuous latent variable for Latent Group 2, and 1 variance of a continuous latent variable for Latent Group 2.

Acknowledgments

We are grateful to Holmes Finch (Ball State University), an anonymous reviewer, Inhee Choi (University of California, Berkeley), Paul De Boeck (Ohio State University), Harrison Kell (Educational Testing Service), Insu Paek (Florida State University), and Sonya Sterba (Vanderbilt University) for helpful comments on earlier versions of this module.

Appendix A: Mplus Code to Fit a Mixture Rasch Model for Latent DIF

```
TITLE: Rasch_1_2
DATA: FILE IS vadata.txt;
VARIABLE: NAMES ARE y1-y24 anger gender;
        USEVARIABLES ARE y1-y24;
          CATEGORICAL = y1-y24;
        CLASSES = c(2);
ANALYSIS: TYPE = MIXTURE;
          STARTS = 100 10;
```

```
          PROCESS = 2 (STARTS);
          ESTIMATOR = MLR;
          ALGORITHM=INTEGRATION;
MODEL: %OVERALL%
        f1 BY y1-y24@1;
        c ON anger gender; !For
!multinomial regression model
        !Equality constraints for scale
!comparability: (b1) - (b5)
    %c#1% !For Latent Group 1
        [y1$1- y10$1] (c1-c10);
        [y11$1] (b1);
        [y12$1 - y13$1] (c12-c13);
        [y14$1] (b2);
        [y15$1] (c15);
        [y16$1] (b3);
        [y17$1] (b4);
        [y18$1 - y21$1] (c18-c21);
        [y22$1] (b5);
        [y23$1 - y24$1] (c23 - c24);
        [f1@0]; f1;
        %c#2% !For Latent Group 2
        [y1$1- y10$1] (d1-d10);
        [y11$1] (b1);
        [y12$1 - y13$1] (d12-d13);
        [y14$1] (b2);
        [y15$1] (d15);
        [y16$1] (b3);
        [y17$1] (b4);
        [y18$1 - y21$1] (d18-d21);
        [y22$1] (b5);
        [y23$1 - y24$1] (d23 - d24);
        [f1]; f1;
MODEL CONSTRAINT: !Wald's test for DIF
  !magnitudes
NEW(dif1 dif2 dif3 dif4 dif5 dif6 dif7
  dif8 dif9 dif10);
NEW(dif12 dif13 dif15 dif18 dif19 dif20
  dif21 dif23 dif24);
dif1=c1-d1;dif2=c2-d2;
dif3=c3-d3;dif4=c4-d4;
dif5=c5-d5;dif6=c6-d6;
dif7=c7-d7;dif8=c8-d8;
dif7=c7-d7;dif8=c8-d8;
dif9=c9-d9;dif10=c10-d10;
dif12=c12-d12;dif13=c13-d13;
dif15=c15-d15;dif18=c18-d18;
dif19=c19-d19;dif20=c20-d20;
dif21=c21-d21;dif23=c23-d23;
dif24=c24-d24;
OUTPUT: TECH1 TECH8;
        STANDARDIZED;
SAVEDATA: FILE IS Rasch2_gmem_anchor.dat;
SAVE=CPROBABILITIES;
```

Appendix B: Mplus Code to Fit a Multidimensional Mixture Rasch Model for Purifying DIF

```
TITLE: Rasch_1_2
DATA: FILE IS vadata.txt;
VARIABLE: NAMES ARE y1-y24 anger gender;
        USEVARIABLES ARE y1-y24;
          CATEGORICAL = y1-y24;
        CLASSES = c(2);
ANALYSIS: TYPE = MIXTURE;
          STARTS = 100 10;
          PROCESS = 2 (STARTS);
          ESTIMATOR = MLR;
ALGORITHM=INTEGRATION;
MODEL: %OVERALL%
        f1 BY y1-y24@1;
        f2 BY y1-y24;
    %c#1%
        f2 BY y1-y24@0;
        [y1$1- y24$1] (b1-b24);
        [f1@0]; f1;
        [f2@0]; f2@0;
        f1 BY f2@0;
        %c#2%
        f2 BY y1@0;
        f2 BY y2@0;
        f2 BY y3@1;
        f2 BY y4@0;
        f2 BY y5@0;
        f2 BY y6@1;
        f2 BY y7@0;
        f2 BY y8@0;
        f2 BY y9@1;
        f2 BY y10@0;
        f2 BY y11@0;
        f2 BY y12@1;
        f2 BY y13@0;
        f2 BY y14@0;
        f2 BY y15@1;
        f2 BY y16@0;
        f2 BY y17@0;
        f2 BY y18@1;
        f2 BY y19@0;
        f2 BY y20@0;
        f2 BY y21@1;
        f2 BY y22@0;
        f2 BY y23@0;
        f2 BY y24@1;
        [y1$1- y24$1] (b1-b24);
        [f1]; f1;
        [f2]; f2;
        f1 WITH f2@0;
OUTPUT: TECH1 TECH8;
STANDARDIZED;
SAVEDATA: FILE IS gmem_puri.dat;
SAVE=CPROBABILITIES;
```

Notes

[1]The focal group refers to the particular group of interest, whereas the reference group refers to the group with whom the focal group is to be compared.

[2]It is also called a *latent class* or *mixture* in the literature. We chose the term *latent group* to contrast with the manifest group.

[3]They are also known as factor mixture models in a factor analytic modeling framework (e.g., Lubke & Neale, 2008; Muthén & Shedden, 1999).

[4]In our survey, we did not classify papers into two categories, applications for unintentionally multidimensional tests and applications for intentionally multidimensional tests, because our survey purpose is for readers to be aware of examples of mixture item response models.

[5]Cho and Paek (2014) is available from the first author upon request.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*, 67–91.

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

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Dekker.

Bennink, M., Croon, M. A., Keuning, J., & Vermunt, J. K. (2014). Measuring student ability, classifying schools, and detecting item bias at school level, based on student-level dichotomous items. *Journal of Educational and Behavioral Statistics*, *39*, 180–202.

Bock, D. R., & Zimowski, M. F. (1997). The multiple groups IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York, NY: Springer-Verlag.

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.

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.

Boughton, K. A., & Yamamoto, K. (2007). A hybrid model for test speededness. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 147–156). New York, NY: Springer.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370.

Cho, S.-J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics*, *35*, 336–370.

Cho, S.-J., Cohen, A. S., & Bottge, B. (2013). Detecting intervention effects using a multilevel latent transition analysis with a mixture IRT model. *Psychometrika*, *78*, 576–600.

Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2013). Markov chain Monte Carlo estimation of a mixture item response theory model. *Journal of Statistical Computation and Simulation*, *83*, 278–306.

Cho, S.-J., Cohen, A. S., Kim, S.-H., & Bottge, B. (2010). Latent transition analysis with a mixture IRT measurement model. *Applied Psychological Measurement*, *34*, 483–504.

Cho, S.-J., & Paek, I. (2014). Criteria of evaluating class invariant items in mixture item response modeling. (Unpublished manuscript).

Choi, I.-H., & Wilson, M. (2014). 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.

Cohen, A. S., Gregg, N., & Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities Research and Practice*, *20*(4), 225–233.

Congdon, P. (2003). *Applied Bayesian modelling*. Hoboken, NJ: Wiley.

Dai, Y. (2013). A mixture Rasch model with a covariate: A simulation study via Bayesian Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *37*, 375–396.

Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, *83*(401), 173–178.

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

de Ayala, R. J., Kim, S.-H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing*, *2*, 243–276.

De Boeck, P., Cho, S.-J., & Wilson, M. (2011). Explanatory secondary dimension modeling of latent differential item functioning. *Applied Psychological Measurement*, *35*, 583–603.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer-Verlag.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. New York, NY: John Wiley.

Fieuws, S., Spiessens, B., & Draney, K. (2004). Mixture models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 317–340). New York, NY: Springer-Verlag.

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.

Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement*, *47*, 432–457.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models: Modeling and applications to random processes*. New York, NY: Springer.

Hancock, G. R., & Samuelsen, K. M. (2007). *Advances in latent variable mixture models*. Charlotte, NC: Information Age Publishing.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Thousand Oaks, CA: Sage.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Hong, S., & Min, S.-Y. (2007). Mixed Rasch modeling of the self-rating depression scale incorporating latent class and Rasch rating scale models. *Educational and Psychological Measurement*, *67*, 280–299.

Jiao, H., Macready, G., Liu, J., & Cho, Y. (2012). A mixture Rasch model-based computerized adaptive test for latent class identification. *Applied Psychological Measurement*, *36*, 469–493.

Kadengye, D. T., Ceulemans, E., & van den Noortgate, W. (2014). A generalized longitudinal mixture IRT model for measuring differential growth in learning environments. *Behavior Research Methods*, *46*, 823–840.

Kim, J.-S., & Bolt, D. M. (2007). Estimating item response theory models using Markov Chain Monte Carlo. *Educational Measurement: Issues and Practice*, *26*(4), 38–51.

Leite, W. L., & Cooper, L. A. (2010). Detecting social desirability bias using factor mixture models. *Multivariate Behavioral Research*, *45*, 271–293.

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lubke, G., & Neale, M. (2008). Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. *Multivariate Behavioral Research*, *43*, 592–620.

Magis, D., Béland, S., & Raîche, G. (2013). *Collection of methods to detect dichotomous differential item functioning (DIF) in psychometrics*. Retrieved from: http://cran.r-project.org/web/packages/difR/difR.pdf

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

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2010). Improvement in detection of differential item functioning using a mixture item response theory model. *Multivariate Behavioral Research*, *45*, 975–999.

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

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational and Behavioral Statistics*, *7*, 105–118.
Meulders, M., & Xie, Y. (2004). Person-by-item predictors, In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 213–234). New York, NY: Springer-Verlag.
Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, *34*, 521–538.
Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.
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.
Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, *55*, 463–469.
Muthén, L. K., & Muthén, B. O. (1998–2013). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.
Oshima, T., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests. *Journal of Educational Measurement*, *34*, 253–272.
Paek, I., & Cho, S.-J. (2014). A note on parameter estimate comparability across latent classes in mixture IRT modeling. *Applied Psychological Measurement*, *39*, 135–143.
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.
Rijmen, F., & De Boeck, P. (2005). A relation between a between-item multidimensional IRT model and the mixture Rasch model. *Psychometrika*, *70*, 481–496.
Rijmen, F., De Boeck, P., & van der Maas, H. L. (2005). An IRT model with a parameter-driven process for change. *Psychometrika*, *70*, 651–669.
Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282.
Rost, J., & von Davier, M. (1995). Mixture distribution Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.) *models: Foundations, recent developments and applications* (pp. 257–268). New York, NY: Springer.
Samuelsen, K. (2005). *Examining differential item functioning from a latent class perspective* (Unpublished doctoral dissertation). University of Maryland, College Park, MD.
Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement*, *16*, 143–152.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.
Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*, 333–343.
Smit, A., Kelderman, H., & van der Flier, H. (2000). The mixed Birnbaum model: Estimation using collateral information. *Methods of Psychological Research Online*, *5*(4), 31–43.
Spiegelhalter, D., Thomas, A., & Best, N. (2000). WinBUGS Version 1.3.
Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, *91*(1), 25–39.
Sterba, S. K. (2013). Understanding linkages among mixture models. *Multivariate Behavioral Research*, *48*, 775–815.
Suh, Y., Cho, S.-J., & Wollack, J. A. (2012). A comparison of item calibration procedures in the presence of test speededness. *Journal of Educational Measurement*, *49*, 285–311.
Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354.
Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Erlbaum.
Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum.
Vansteelandt, K. (2000). *Formal models for contextualized personality psychology* (Unpublished doctoral dissertation). K.U. Leuven, Leuven Belgium.
Vermunt, J. K. (1997). *LEM 1.0: A general program for the analysis of categorical data*. Tilburg, The Netherlands: Tilburg University.
Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, *17*, 33–51.
Vermunt, J. K., & Magidson, J. (2005). *Technical guide for Latent GOLD 4.0: Basic and advanced*. Belmont, MA: Statistical Innovations Inc.
von Davier, M. (2001). *WINMIRA 2001*. St. Paul, MN: Assessment Systems Corporation.
von Davier, M. (2005). *mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models*. Princeton, NJ: Educational Testing Service.
von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307.
von Davier, M., & Carstensen, C. H. (2007). *Multivariate and mixture distribution Rasch models: Extensions and applications*. New York, NY: Springer.
von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, *76*, 318–336.
von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement*, *28*, 389–406.
Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*, 479–498.
Willse, J. (2009). *mixRasch: Mixture Rasch Models with JMLE*. Retrieved from http://cran.r-project.org/web/packages/mixRasch/mixRasch.pdf
Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin*, *105*(2), 276–289.
Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement*, *40*, 307–330.
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* (pp. 89–98). New York, NY: Waxmann.
Zumbo, B. (2007). Three generation of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*, 223–233.