IRT Modeling in the Presence of Zero-Inflation With Application to Psychiatric Disorder Severity

Applied Psychological Measurement

2015, Vol. 39(8) 583–597

© The Author(s) 2015

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621615588184

apm.sagepub.com

Melanie M. Wall^{1,2}, Jung Yeon Park¹, and
Irin Moustaki³

Abstract

Item response theory (IRT) has been increasingly utilized in psychiatry for the purpose of describing the relationship among items in psychiatric disorder symptom batteries hypothesized to be indicators of an underlying latent continuous trait representing the severity of the psychiatric disorder. It is common to find zero-inflated (ZI) data such that a large proportion of the sample has none of the symptoms. It has been argued that standard IRT models of psychiatric disorder symptoms may be problematic due to the unipolar nature of many clinical traits. In the current article, the authors propose to address this by using a mixture model to approximate the unknown latent trait distribution in the IRT model while allowing for the presence of a non-pathological subgroup. The basic idea is that instead of assuming normality for the underlying trait, the latent trait will be allowed to follow a mixture of normals including a degenerate component that is fixed to represent a non-pathological group for whom the psychiatric symptoms simply are not relevant and hence are all expected to be zero. The authors demonstrate how the ZI mixture IRT method can be implemented in Mplus and present a simulation study comparing its performance with a standard IRT model assuming normality under different scenarios representative of psychiatric disorder symptom batteries. The model incorrectly assuming normality is shown to have biased discrimination and severity estimates. An application further illustrates the method using data from an alcohol use disorder criteria battery.

Keywords

mixture IRT model, latent trait, symptom batteries

¹Columbia University, New York, NY, USA

²New York State Psychiatric Institute, New York City, USA

³London School of Economics and Political Science, UK

Corresponding Author:

Irin Moustaki, Department of Statistics, London School of Economics, Houghton Street, London WC2A 2AE, UK.

Email: i.moustaki@lse.ac.uk

Introduction

Item response theory (IRT) models the relationship between an individual's underlying latent trait(s) and their responses to different measurable items. The most common method of estimation for parameters in this model is maximum likelihood (ML) assuming that the latent trait is distributed normally (Bock & Aitkin, 1981; Embretson & Reise, 2000). In the educational testing context where IRT has been used extensively, it is often reasonable to assume the latent trait, for example, mathematics ability, follows a normal distribution. More recently, IRT has been utilized in psychiatry for the purpose of describing the relationship among psychiatric symptoms and criteria for diagnosis of psychiatric disorders (e.g., Aggen, Neale, & Kendler, 2005; Chan, Orlando, Ghosh-Dastidar, Duan, & Sherbourne, 2004; Krueger & Finger, 2001). However, unlike traits measured in education, the a priori expectation of the distribution of a psychiatric disorder trait is typically that it is right skewed. For example, when measuring the severity of a psychiatric disorder in a general community sample, it is expected that most people are at the non-pathological end of the trait while a small number of individuals are spread out across the mild, moderate, and high severity end of the disorder. In psychiatric symptom batteries, it is common to have floor effects with many people reporting no symptoms (Delucchi & Bostrom, 2004). Nevertheless, the estimation of IRT model parameters (i.e., discrimination and severity parameters) with psychiatric disorder data is almost exclusively obtained under the assumption that the underlying trait is normally distributed.

It has been argued (Reise & Waller, 2009) that standard IRT models of psychiatric disorder symptoms may be problematic due to the unipolar nature of many clinical traits where one end of the trait represents increasing severity and the other end represents absence of the disorder rather than decreasing severity. Reise and Waller (2009) further pointed out that standard IRT modeling of clinical assessments often leads to surprisingly high item discrimination parameters (i.e., greater than 2.5 on the logistic metric). As an example, they referred to Chan et al. (2004) who reported item discriminations of 4.43 ("could not shake the blues") and 4.14 ("felt sad") on a popular depression measure. Reise and Waller (2009) suggested that the cause of these high discrimination parameter estimates is conceptually narrow constructs and consequently homogeneous item content that leads to highly inter-correlated items and consequently high item discriminations. But, as will be demonstrated in the present article, the high item discriminations are likely an artifact of improperly modeling the unipolar nature (i.e., fundamentally non-normal nature) of the latent trait with a normal distribution.

In the current article, the authors propose to address these issues by using a mixture model to approximate the unknown latent trait distribution in the IRT model while allowing for the presence of a non-pathological subgroup. The basic idea is that instead of assuming normality for the underlying trait in the IRT model, they will allow the latent trait to follow a mixture of normals with the number of components empirically determined by the data including a degenerate component that is fixed to represent a non-pathological group for whom the psychiatric symptoms simply are not relevant and hence are all expected to be zero. By incorporating the possibility of a non-pathological group, the model thus provides estimates of discrimination and severity parameters for the subset of the population for whom the symptom battery is relevant. A practical strength of this mixture approach is that it is readily implementable in existing software. In particular, the code for how Mplus can be used to implement this type of mixture of normals with a degenerate component into an IRT modeling framework will be demonstrated and provided. A simulation study motivated by typical psychiatric disorder criteria applications, including highly skewed responses with number of items equal to 10 (typical length of symptom batteries in psychiatric research), will be presented. This simulation adds to the literature by demonstrating the bias that can occur in the standard IRT model assuming normality and how

the zero-inflated (ZI) mixture IRT model easily resolves this bias. Furthermore, the method on a real application of alcohol use disorder criteria obtained from a large national survey will be demonstrated, and the results obtained treating the underlying trait as normal will be contrasted.

Review of the Related Methodological Literature

The method described in this article for IRT modeling with ZI data intersects two related methodological literatures, one focused on flexibly estimating latent trait distributions, and the other focused on identifying latent subgroups with atypical response patterns. Here, relevant work is reviewed.

Previous literature has examined the impact of wrongly assuming normality for the latent trait on item parameter estimation (Reise & Yu, 1990; Seong, 1990; Stone, 1992; Wall, Guo, & Amemiya, 2012) and on person parameters (or latent trait estimation; Sass, Schmitt, & Walker, 2008). Both Seong (1990) and Stone (1992) concluded that bias in item parameters increases with greater departure from normality in the latent trait distribution (e.g., more severe skew) but that bias diminishes as test length increased. Wall et al. (2012) showed that when the latent trait is highly skewed, for a fixed number of items (they considered less than 10 items), bias in the discrimination parameters relating the latent trait to a dichotomous outcome does not diminish with increased sample size. The simulation study by Sass et al. (2008) also showed that when examinees were simulated to follow a positively skewed distribution, estimation of the underlying latent trait using an IRT model (assuming normality for the trait) was highly error-prone.

Several methods are available for weakening the normal distributional assumption for the underlying latent trait. Andersen and Madsen (1977) demonstrated a method for estimating the parameters of a non-normal latent trait distribution by marginal ML given that the item parameters were estimated first with conditional ML which is possible in the Rasch model. Bock and Aitkin (1981) within the two-parameter logistic (2PL) IRT model proposed the Expectation-Maximization (EM) algorithm for estimating the item parameters assuming the latent distribution is approximated by means of a finite number of equally spaced points. This assumption for the latent distribution has now come to be called the non-parametric factor distribution; see, for example, Muthén (2007) and Knott and Tzamourani (2007). Mislevy (1984) proposed a two-stage procedure to estimate population parameters of various parametric latent distributions such as log beta and mixtures of normals. Cagnone and Viroli (2012) for the 2PL IRT model and Wall et al. (2012) for a latent factor model with mixed continuous and categorical indicators used a mixture of normals to approximate the latent trait distribution and performed ML estimation. The use of mixtures of normals has become common in the context of growth mixture models (similar in structure to IRT models) because they are easy to use for model comparison and readily implementable in software regardless of whether the latent subgroups are of interest (Bauer, 2007; Bauer & Curran, 2003, 2004). Woods & Thissen (2006) and Woods (2008) used spline-based density estimation to approximate the latent distribution within a 2PL IRT model, which is the so-called Ramsay-Curve IRT. All of these methods are designed to flexibly model the latent distribution while assuming the functional relationship between the latent trait and the responses are the same across the distribution of the trait.

Another line of literature has utilized mixture IRT models to identifying subsets of the population (latent classes) that are characterized by different response patterns (Cao & Stokes, 2008; Cohen & Bolt, 2005; De Ayala, Kim, Stapleton, & Dayton, 2002; Li, Cohen, Kim, & Cho, 2009; Lubke et al., 2007; Mislevy & Verhelst, 1990; Moustaki & Knott, 2014; Von Davier & Rost, 2006). In this literature, the relationship between the latent trait and the response pattern is assumed to be different in every component of the mixture. Psychometricians have used this type of mixture IRT model to detect and characterize differential item functioning or atypical

response patterns. A special case of the mixture IRT model (Finkelman, Green, Gruber, & Zaslavsky, 2011; Muthén & Asparouhov, 2006) is the ZI IRT model where two components are considered such that one has response probabilities fixed to be zero thus incorporating ZI responses, and the other is assumed to be normally distributed.

As will be detailed in the next section, the ZI mixture IRT model considered in the present article builds upon the ZI IRT model (Finkelman et al., 2011; Muthén & Asparouhov, 2006) by extending the non-zero component to be modeled with a flexible latent distribution (i.e., a mixture of normal). Hence, the proposed method combines the idea of identifying differential response patterns (i.e., zero responses) with a method for weakening the normal distributional assumption for the latent trait.

Model and Estimation

2PL IRT Model

Let y_{ij} be the dichotomous response on the j th item for i th individual where $P(y_{ij} = 1 | \theta_i) = P_j(\theta_i)$ stands for the probability that i th individual with latent trait θ_i responds to the j th item in the affirmative, where $j = 1, \dots, J$. The authors consider the 2PL IRT model (Birnbaum, 1968)

$$\text{logit } P_j(\theta_i) = \alpha_j(\theta_i - \beta_j), \quad (1)$$

where α_j and β_j are the item discrimination and severity parameters, respectively. In addition to Equation 1, the IRT model further assumes conditional independence for two different items y_{ij} and $y_{ij'}$ such that y_{ij} is independent of $y_{ij'}$ conditional on θ_i . That is known as the measurement part of the model. The structural part of the model, or the model for the latent trait, is then often assumed to be normally distributed, $\theta_i \sim N(0, 1)$. Hence, given a set of independent and identically distributed (i.i.d.) observations, the marginal likelihood can be easily derived and estimation performed via ML.

ZI Mixture IRT Model for the Latent Trait and Estimation

The basic idea of the method described here is to replace the normality assumption for θ_i with a mixture of normals. It is assumed that the population is made up of K subgroups each coming from a normal distribution with different mean μ_k and variance σ_k^2 mixed at random in proportion to the relative group sizes $\eta_1, \eta_2, \dots, \eta_K$. Because the subgroups are unobserved (or may not truly exist), the authors specify the mixture model in terms of a latent categorical variable (latent class) c_i (Hagenaars & McCuthcheon, 2002), which can take on any one of the K different values for each individual i with $P(c_i = k) = \eta_k$, such that $\sum_{k=1}^K \eta_k = 1$. Then the mixture model for the latent θ is

$$\theta_i \sim \sum_{k=1}^K \eta_k N(\mu_k, \sigma_k^2). \quad (2)$$

Floor effects are common in psychiatric symptom batteries, such that many individuals indicate no to all items, what is referred to as zero-inflation. A possible explanation for this is these individuals are coming from a non-pathological group, that is, a group of individuals who do not have any of the symptoms because they simply do not have the disorder and, thus, do not exhibit any of the trait. These individuals are fundamentally different from individuals who have some level of severity of disorder for which the battery of items was designed to measure. Moreover, the authors hypothesize that the non-pathological group is different from individuals

who have the trait but did not endorse any of the symptoms. That is, the model allows for the possibility that there may be people who are pathological but who do not exhibit symptoms. One way to consider incorporating the non-pathological group into the mixture model (Equation 2) is by introducing a mixture component that has a fixed mean at an extreme negative value and a zero variance, that is

$$\mu_K = -100 \text{ and } \sigma_K^2 = 0. \quad (3)$$

In other words, this component has a degenerate distribution with mass 1 at the value of the trait equal to -100 . By fixing μ_K at an arbitrarily large negative value and the variance to zero, this implies $\theta_i = -100$ for all individuals in this class, which then ensures practically zero probability of endorsement to each item.

The IRT model (Equation 1) with a mixture model (Equation 2) that includes a fixed component (Equation 3) will be referred to as the ZI mixture IRT model. An identifiability problem arises with this type of specification related to the arbitrary nature of latent variable means and variances. As is true in any latent trait model, it is necessary to fix the mean and variance of the latent trait, most commonly to 0 and 1, respectively. For the mixture model (Equation 2), the mean and variance of θ_i are

$$\mu = E(\theta_i) = \eta_1 \mu_1 + \eta_2 \mu_2 + \dots + \eta_K \mu_K, \quad (4)$$

$$\sigma^2 = \text{Variance}(\theta_i) = \eta_1 (\sigma_1^2 + \mu_1^2) + \eta_2 (\sigma_2^2 + \mu_2^2) + \dots + \eta_K (\sigma_K^2 + \mu_K^2) - \mu^2. \quad (5)$$

When the mean and variance are fixed, that is, $\mu = 0$ and $\sigma^2 = 1$ in Equations 4 and 5, it is seen that the choice of what μ_K is fixed at (e.g., -100 vs. -1000) creates indeterminacy in the other parameters. To avoid this indeterminacy, the non-pathological group is not included in the scaling of the latent trait. That is, the mean and variance of the latent trait formed are fixed by only the $K - 1$ components, that is, $\theta_i \sim \sum_{k=1}^{K-1} \eta_k N(\mu_k, \sigma_k^2)$ to be 0 and 1, respectively. Thus, the discrimination and severity parameters are estimated and scaled only for the pathological group, that is, the group who are hypothesized to have some level of severity of the trait under study.

ML via the EM algorithm with numerical integration in the E-step to approximate the integral over the latent trait can be used for parameter estimation. Specific syntax for how to implement the model in Mplus is shown in the supplementary online material. Similar syntax could be developed using the Latent Gold software (Vermunt & Magidson, 2013). To determine the optimal number of classes K , the authors recommend fitting increasing number of classes and comparing model fit using information criteria that penalize for the complexity, that is, Akaike information criterion (AIC) and Bayesian information criterion (BIC). The BIC has been demonstrated to show good performance determining the optimal number of classes in mixture models (Nylund, Asparouhov, & Muthén, 2007).

Simulation Study

The authors simulate data mimicking typical response patterns coming from a psychiatric disorder symptom battery administered in a general community population. The number of symptoms is taken to be 10 as this represents a common length of psychiatric symptom batteries; for example, there are 11 substance use symptom criteria that make up the current *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.; *DSM-5*; American Psychiatric Association [APA], 2013) definition of alcohol use disorder. A moderate sample size for a psychiatric epidemiologic study of $n = 5,000$ is considered.

The response patterns are generated such that one subgroup in the population did not have any probability of having any of the symptoms (i.e., the “non-pathological group”) and another subgroup (i.e., the “pathological group”) where latent traits for individuals were drawn from a continuous distribution and symptoms occurred according to the IRT model (Equation 1). Three scenarios were considered. The first two generated the continuous latent trait distribution for the pathological group from a normal $N(0, 1)$ and differed in the mixing proportion η_K that governs the proportion of non-pathological individuals fixed at 25% or 75%. Because psychiatric symptom batteries are often concerned with identifying individuals at the more severe end of the trait, in these two scenarios, the severity parameters are fixed to be equally spaced between 0.5 and 3.5, thus resulting in items with relatively low prevalence from the normally distributed pathological group. True discrimination parameters were all fixed to be one, also consistent with psychiatric symptom batteries where the assumption is usually that items can be simply added with equal weight to obtain a score. The third scenario generated the continuous latent trait distribution for the pathological group from a non-normal distribution, specifically a mixture of two-normals with one mode at the low end and one at the high end of the trait both with equal probability. The bimodal latent trait was scaled to have mean 0 and variance 1. In the third scenario, the proportion of non-pathological individuals, η_K , was fixed at 50% and the severity parameters were fixed to be equally spaced between -1.5 and 1.5 to represent a battery intended to measure the spectrum from mild to severe disorder. The discrimination parameters were again fixed to be 1 in the data generating model. Hence under the first two scenarios, the true data generating mechanism for θ_i is a ZI mixture (Equations 2 and 3) with $K = 2$, and under the third scenario, it is the same but with $K = 3$. Under each scenario, 100 data sets are generated and model (Equation 1) is fitted assuming for the latent trait (a) a normal distribution, (b) a two-class ZI mixture, (c) a three-class ZI mixture, and for the third scenario, a four-class ZI mixture.

Estimated severity and discrimination parameters, $\hat{\alpha}_j$ and $\hat{\beta}_j$, for each item $j = 1 \dots 10$ are summarized in terms of bias and mean square error (MSE) across each estimation method. To compare estimators in terms of item bias aggregating across items, $1/J \sum_{j=1}^J |\bar{\alpha}_j - \alpha_j|$ and $1/J \sum_{j=1}^J |\bar{\beta}_j - \beta_j|$ are used, where $J = 10$ and $\bar{\alpha}_j$ and $\bar{\beta}_j$ are the means of the respective estimates across the 100 simulated data sets specifically, $\bar{\alpha}_j = 1/B \sum_{b=1}^B \hat{\alpha}_{bj}$ and $\bar{\beta}_j = 1/B \sum_{b=1}^B \hat{\beta}_{bj}$, where $B = 100$. To compare the estimators in terms of MSE, the MSE for each item is first obtained, $MSE_j(\alpha) = \text{variance}(\hat{\alpha}_j) + \text{bias}(\hat{\alpha}_j)^2$ and $MSE_j(\beta) = \text{variance}(\hat{\beta}_j) + \text{bias}(\hat{\beta}_j)^2$, then the MSEs are averaged across the $J = 10$ items, that is, $1/J \sum_{j=1}^J MSE_j(\alpha)$ and $1/J \sum_{j=1}^J MSE_j(\beta)$.

Results When Truth Is ZI $K = 2$ Mixture

Figure 1 presents the empirical means ($\bar{\alpha}_j$ and $\bar{\beta}_j$) from the three different models fit in the first two scenarios. When the true distribution of the latent trait has 75% people who are non-pathological (left side of Figure 1), it is seen that discrimination estimates from the model that assumes the underlying distribution is normal are dramatically overestimated with values ranging from 2.5 to 3.5 while the true value is 1. Similarly, though to a less extent, the normal model overestimates the discrimination when only 25% of the true sample is non-pathological (right side of Figure 1) with values ranging from 1.3 to 1.7. Moreover, for the severity parameters with true values ranging from 0.5 to 3.5, in the case where 75% of the population is non-pathological, the normal model overestimates the severity on the low end ($M = 1.5$ rather than 0.5) but underestimates the severity on the high end ($M = 2.8$ rather than 3.5). The bias for the severity parameter is much smaller in the case of 25% non-pathological population but exhibits the same phenomena of underestimation at the higher values of true severity.

Figure 1. Truth is ZI $K = 2$ mixture with 75% (left) or 25% (right) of the population non-pathological. Note. Empirical mean (and median for the three-class ZI mixture) item parameter estimates (discrimination on top, severity on bottom) across 100 simulated data sets estimated with the three different models. ZI = zero-inflated.

Table 1. Truth is ZI K = 2 Mixture With 75% or 25% of the Population Non-Pathological.

Model	Discrimination parameters				Severity parameters			
	75% non-pathological		25% non-pathological		75% non-pathological		25% non-pathological	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
Normal IRT	1.763	3.249	0.397	0.173	0.456	0.293	0.177	0.053
Two-class ZI mixture IRT	0.012	0.026	0.007	0.009	0.033	0.139	0.013	0.035
Three-class ZI mixture IRT	0.306	0.691	0.177	0.281	0.076	0.443	0.037	0.193

Note. Empirical absolute bias and MSE across 100 simulated data sets of size $n = 5,000$ aggregated across all 10 discrimination and 10 severity parameters. ZI = zero-inflated; MSE = mean square error; IRT = item response theory.

Table 2. Truth is ZI K = 2 Mixture With 75% or 25% of the Population Non-Pathological.

Model	Number of parameter	75% non-pathological		25% non-pathological	
		AIC	BIC	AIC	BIC
Normal	20	14050.06	14180.41	33919.22	34049.57
Two-class ZI mixture	21	13880.62 ^a	14017.48 ^a	33828.74 ^a	33965.60 ^a
Three-class ZI mixture	23	13893.37	14043.26	33843.67	33993.56

Note. Model comparison with AIC and BIC averaged over 100 simulated data sets. ZI = zero-inflated.

^aTwo-class ZI mixture resulted in best fit (i.e., lowest AIC and BIC) for 100% of simulated data sets.

In comparison, the two-class ZI mixture IRT is unbiased for both the discrimination and severity parameters in every case. The two-class ZI mixture IRT model matches the data generating mechanism and has the smallest MSE across all estimators (see Table 1).

The three-class ZI mixture model exhibits some upward mean bias for the discrimination parameters. Closer inspection of the empirical distribution of the estimates of the discrimination across the models (Figure 2) shows that the three-class ZI mixture model estimates are right skewed. Of note, the median of the estimator appears to be unbiased demonstrated also in Figure 1 where the median discrimination parameter estimates (across 100 data sets) are all nearly one. Indeed, the three-class mixture model is overly complicated for the data that were generated with just two classes, and for a large proportion of the data sets, computational warnings are indicated by Mplus, that is, stating the procedure has reached a saddle point. Moreover, the authors found that in nearly all the cases where the discrimination parameter estimate was larger than 1.5, one of the three classes was estimated to have nearly zero (i.e., <0.01) proportion of the population. This indicates that the poorly estimated discrimination values found in the three-class ZI model come from data sets where the three-class ZI model would not be chosen anyway because the third class had zero probability. Furthermore, this estimation difficulty for the three-class ZI mixture IRT model is found to be a result of overfitting which is also avoided by noting the AIC and BIC both suggests choosing the simpler two-class model (Table 2). Specifically for every one of the 100 data sets generated, the two-class model was chosen by BIC.

Figure 3 compares the factor score estimates (posterior predicted mean of the underlying latent trait given observed data) from the normal model and the two-class ZI mixture model.

Figure 2. Truth is ZI $K = 2$ mixture with 75% non-pathological.

Note. Side-by-side boxplots of 100 estimates of all 10 of the discrimination parameter β_1 - β_{10} (true value = 1) for each of the three models. ZI = zero-inflated.

Figure 3. Truth is ZI $K = 2$ mixture with 75% (left) or 25% (right) of the population non-pathological.

Note. Factor score estimates compared with true factor score. Top shows results using the two-class zero-inflated mixture, bottom shows the normal model. Factor scores estimates for individuals with all responses equal to zero are not included. ZI = zero-inflated.

Table 3. Truth is ZI K = 3 Mixture With 50% of the Population Non-Pathological.

Model	Number of parameter	50% non-pathological	
		AIC	BIC
Normal	20	37290.02	37420.38
Two-class ZI mixture	21	36126.21	36263.07
Three-class ZI mixture	23	36092.22 ^a	36242.12 ^a
Four-class ZI mixture	25	36105.58	36268.51

Note. Model comparison with AIC and BIC averaged over 100 simulated data sets. ZI = zero-inflated.

^aThree-class ZI mixture resulted in best fit (i.e., lowest AIC and BIC) for 80% of simulated data sets, and two-class model was best fit for 20%. Neither the normal nor the four-class ZI mixture was ever the best fit.

Figure 4. Truth is ZI K = 3 mixture with 50% of the population non-pathological.

Note. Empirical mean (and median for the four-class ZI mixture) item parameter estimates (discrimination on left, severity on right) across 100 simulated data sets estimated with the four different models. ZI = zero-inflated.

The two-class ZI model factor scores are found to be better correlated with the true scores. The normal model factor scores are shrunk together more, not providing the same level of separation of individuals that the two-class ZI model does. This corresponds to the biased severity parameters found for the normal model where the more extreme items were biased downward and the less extreme items were biased upward.

Results When Truth Is ZI K = 3 Mixture

This scenario provides the opportunity to investigate how the ZI mixture IRT model performs when the truth is that in addition to there being a non-pathological group, the pathological group is non-normally distributed. In Figure 4, large upward bias is found again in the discrimination parameters when incorrectly assuming the normal distribution for the underlying trait of the population, with estimates ranging between 2 and 4 rather than the true value of 1. Also, because of the non-pathological group that has all zero responses for at least 50% of the population, the normal model estimates the severity parameters all to be on the high end of the trait, ranging from

Table 4. NESARC Alcohol Disorder Criteria Data Model Comparison With AIC and BIC.

Model	Number of parameter	AIC	BIC
Normal	22	186864.013	187044.448
Two-class ZI mixture	23	186293.301	186481.937
Three-class ZI mixture	25	186295.470	186500.510

Note. NESARC = National Epidemiologic Survey on Alcohol and Related Conditions; ZI = zero-inflated.

0.4 to 1.5, which is inflated from the true values ranging from -1.5 to 1.5 . From Figure 4, it can be seen that the two-class ZI mixture IRT model shows improvement in terms of bias of the parameters, but because the pathological group is non-normally distributed, some bias remains. The three-class ZI model (the true model in this scenario) shows overall good estimation of the item parameters. Moreover, in Table 3, the three-class model is found to outperform the other models in terms of AIC and BIC. Of note, the four-class ZI mixture model showed similar right skew in its parameter estimates as was seen in the first two scenarios where the three-class model was overfitting. But because it was never chosen as the best model, this bias is less of a concern.

IRT Modeling of Alcohol Use Disorder Criteria

The authors demonstrate the ZI mixture IRT model on an alcohol use disorder data from the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) of the National Institute on Alcohol Abuse and Alcoholism (NIAAA). In this data set, the abuse and dependence criteria of the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM-IV*; APA, 1994) define an alcohol use disorder continuum using a large ($N = 43,093$) nationally representative sample of the U.S. population. For the purpose of the present analysis, the sample is restricted to $n = 26,946$ respondents who indicated they are current drinkers. Alcohol use disorder criteria were previously analyzed with a normal IRT model using a similar subset of the NESARC (Saha, Chou, & Grant, 2006). The respondents were asked about lifetime problems with 11 different alcohol disorder diagnostic criteria. For example, one of the criteria is problems with “Tolerance” as defined by a need for markedly increased amounts of alcohol to achieve intoxication or desired effect, and another criteria is “Time Spent” which corresponds to agreement with a great deal of time being spent in activities necessary to obtain alcohol, use alcohol, or recover from its effect. All responses for the 11 diagnostic criteria are scored as binary (yes/no). The distribution of sum scores of the 11 items is right skewed with 48% of the sample endorsing none of the criteria, and 13%, 10%, 8%, 6%, 4%, and 10% endorsing 1, 2, 3, 4, 5, and 6 or more criteria, respectively.

The normal IRT model and ZI mixture IRT models with two and three classes were fit to the 11 criteria also adjusting for stratification, clustering, and weighting due to the complex sampling design of the NESARC. Figure 5 shows the resulting discrimination and severity parameters for all three models (the two- and three-class ZI mixture IRT model estimates are practically identical). The AIC and BIC for the two-class ZI mixture IRT model is found to be smallest, indicating it is the best (Table 4). Notice that the discrimination parameters are all much higher using the normal model. Based on the similarity of the NESARC data to the data the authors generated in the simulation and the results they found there, they conclude that the normal model IRT discrimination estimates for the alcohol symptoms are biased upward, and the two-class mixture model has removed the bias. Moreover, the severity parameters for the normal model are higher than those for the ZI mixture models. Part of the decrease in the severity parameters in the ZI mixture models is due to the extraction of the non-pathological class.

Figure 5. NESARC alcohol disorder criteria data: Discrimination (top) and severity (bottom) estimates. Note. The two- and three-class ZI mixture model estimates are practically identical. NESARC = National Epidemiologic Survey on Alcohol and Related Conditions; ZI = zero-inflated; LL = larger/longer; HU = hazardous use; T = tolerance; CD = cut down; W = withdrawal; TS = time spent; PP = physical or psychological problem; SI = social/interpersonal problem; NR = neglect role; L = legal problem; AGP = activities given up.

That is, in the ZI mixture IRT model, the severity parameters are representative of the subset of the population who have some non-zero probability of having a criteria. Indeed, the two-class mixture IRT model estimated that 31% of the population comes from a non-pathological subgroup with the remaining 69% from a pathological subgroup. Hence, the prevalence of criteria is higher in the pathological group and thus severity goes down.

Discussion

The authors have proposed the ZI mixture IRT model to address estimation of IRT parameters in typical psychiatric data where a non-trivial proportion of the sample is likely to be non-pathological and hence not exhibiting the trait under study. The model combines the idea of using latent classes to identify a group with a unique response patterns (i.e., zero responses) with a method for weakening the normal distributional assumption for the latent trait through a mixture of normals. The authors have shown that there is bias in the discrimination and severity parameters when the standard IRT model assuming normality is used when the truth is that there is a non-pathological group within the sample. They have shown that the bias in the IRT parameters can be alleviated with the ZI mixture IRT. This approach (a) is straightforward to implement in existing software, (b) allows for simple model choice using AIC or BIC to determine how flexible (i.e., how many classes) the mixture model needs to be, (c) provides an

estimate of the proportion of the population that is non-pathological, and (d) can easily be extended to accommodate the inclusion of covariates.

One idea that may arise when considering ways to deal with ZI data is to simply drop all the individuals in the sample with all zeros and analyze the remaining data. Although tempting in its simplicity, this method wrongfully throws out a proportion of the sample who are actually pathological but who do not meet any of the criteria. In the alcohol example, 48% of the sample had all zeros, while the ZI mixture IRT model estimated the non-pathological group to only be 31% of the population. Throwing out all the zeros is the same as selecting who to fit the IRT model to conditional on their scores and this has been shown to lead to biased estimates (Muthén, 1989; Muthén & Jöreskog, 1983). A strength of the ZI mixture IRT model is that the model implicitly identifies what proportion to exclude and include when estimating the item parameters.

Recently, a new type of model motivated by the diffusion model from mathematical psychology has been proposed for “positive abilities” by van der Maas, Molenaar, Maris, Kievit, and Borsboom (2011) and Lucke (2014). Meant to address measurement of unipolar traits (or abilities), this type of modeling while potentially promising is still in early stages without standard software for estimation or evidence that it outperforms or even is comparable with other methods.

Finally, the authors have presented the ZI mixture IRT model for a unidimensional IRT model (i.e., one underlying latent trait). Future work may consider expanding the mixture model to allow for multidimensional underlying traits. While in theory this should be straightforward, computational and identifiability difficulties may arise that require special attention.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Aggen, S. H., Neale, M. C., & Kendler, K. S. (2005). DSM criteria for major depression: Evaluating symptom patterns using latent-trait item response models. *Psychological Medicine*, 35, 475-487.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.
- Andersen, E. B., & Madsen, M. (1977). Estimating the parameters of a latent population distribution. *Psychometrika*, 42, 357-374.
- Bauer, D. J. (2007). Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research*, 42, 757-786.
- Bauer, D. J., & Curran, P. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, 8, 338-363.
- Bauer, D. J., & Curran, P. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods*, 9, 3-29.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.

- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-459.
- Cagnone, S., & Viroli, C. (2012). A factor mixture analysis model for multivariate binary data. *Statistical Modelling*, 12, 257-277.
- Cao, J., & Stokes, L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, 73, 209-230.
- Chan, K. S., Orlando, M., Ghosh-Dastidar, B., Duan, N., & Sherbourne, C. D. (2004). The interview mode effect on the Center for Epidemiological Studies Depression (CES-D) Scale: An item response theory analysis. *Medical Care*, 42, 281-289.
- Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, 42, 133-148.
- 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.
- Delucchi, K. L., & Bostrom, A. (2004). Methods for analysis of skewed data distributions in psychiatric clinical studies: Working with many zero values. *American Journal of Psychiatry*, 161, 1159-1168.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Finkelman, M. D., Green, J. G., Gruber, M. J., & Zaslavsky, A. M. (2011). A zero- and K-inflated mixture model for health questionnaire data. *Statistics in Medicine*, 30, 1028-1043.
- Hagenaars, J., & McCutcheon, A. (2002). *Applied latent class analysis*. Cambridge, UK: Cambridge University Press.
- Knott, M., & Tzamourani, T. (2007). Bootstrapping the estimated latent distribution of the two-parameter latent trait model. *British Journal of Mathematical and Statistical Psychology*, 60, 175-191.
- Krueger, R. F., & Finger, M. S. (2001). Using item response theory to understand comorbidity among anxiety and unipolar mood disorders. *Psychological Assessment*, 13, 140-151.
- 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.
- Lubke, G., Muthén, B., Moilanen, I., McGough, J., Loo, S., Swanson, J., & . . . Smalley, S. (2007). Subtypes versus severity differences in the attention-deficit/hyperactivity disorder in the northern Finnish birth cohort. *Journal of the American Academy of Child & Adolescent Psychiatry*, 46, 1584-1593.
- Lucke, J. F. (2014). Positive trait item response models. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology: Presentations from the 77th Annual Psychometric Society Meeting* (pp. 199-213). New York, NY: Springer.
- Mislevy, R. J. (1984). Estimating latent distribution. *Psychometrika*, 49, 359-381.
- Mislevy, R. J., & Verhelst, N. D. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, 55, 195-215.
- Moustaki, I., & Knott, M. (2014). Latent variable models that account for atypical responses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63, 343-360.
- Muthén, B. (1989). Factor structure in groups selected on observed scores. *British Journal of Mathematical and Statistical Psychology*, 42, 81-90.
- Muthén, B. (2007). Latent variable hybrids: Overview of old and new models. In G. R. Hancock & K. M. Samuelson (Eds.), *Advances in latent variable mixture models* (pp. 1-24). Charlotte, NC: Information Age.
- Muthén, B., & Asparouhov, T. (2006). Item response mixture modeling: Application to tobacco dependence criteria. *Addictive Behaviors*, 31, 1050-1066.
- Muthén, B., & Jöreskog, K. (1983). Selectivity problems in quasi-experimental studies. *Evaluation Review*, 7, 139-173.
- Nylund, K., Asparouhov, T., & Muthén, B. (2007). Deciding on the number of classes in LCA and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14, 535-569.
- Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, 5, 27-28.

- Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133-144.
- Saha, T. D., Chou, S. P., & Grant, B. F. (2006). Toward an alcohol use disorder continuum using item response theory: Results from the National Epidemiologic Survey on Alcohol and Related Conditions. *Psychological Medicine, 36*, 931-941.
- Sass, D. A., Schmitt, T. A., & Walker, C. M. (2008). Estimating non-normal latent trait distribution within item response theory using true and estimated item parameters. *Applied Measurement in Education, 21*, 65-88.
- Seong, T. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14*, 299-311.
- Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*, 1-16.
- van der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review, 118*, 339-356.
- Vermunt, J. K., & Magidson, J. (2013). *LG-Syntax users guide: Manual for Latent GOLD 5.0 Syntax Module*. Belmont, MA: Statistical Innovations.
- Von Davier, M., & Rost, J. (2006). Mixture distribution item response models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics: Vol. 26. Psychometrics* (pp. 643-661). Amsterdam, The Netherlands: Elsevier.
- Wall, M. M., Guo, J., & Amemiya, Y. (2012). Mixture factor analysis for approximating a non-normally distributed continuous latent factor with continuous and dichotomous observed variables. *Multivariate Behavioral Research, 47*, 276-313.
- Woods, C. M. (2008). Ramsay-curve item response theory for the 3PL item response model. *Applied Psychological Measurement, 32*, 447-465.
- Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika, 71*, 281-301.