10-1-2017

Item Response Modeling of Multivariate Count Data With Zero Inflation, Maximum Inflation, and Heaping

Brooke E. Magnus
Marquette University, brooke.magnus@marquette.edu

David M. Thissen
University of North Carolina - Chapel Hill

Accepted version. *Journal of Educational and Behavioral Statistics*, Vol. 42, No. 5 (October 1, 2017): 531-558. DOI.

Journal of Educational and Behavioral Statistics, Vol. 42, No. 5 (2017): 531-558. DOI.

Item Response Modeling of Multivariate Count Data With Zero Inflation, Maximum Inflation, and Heaping

Brooke E. Magnus
Marquette University
David Thissen
University of North Carolina

Abstract

Questionnaires that include items eliciting count responses are becoming increasingly common in psychology. This study proposes methodological techniques to overcome some of the challenges associated with analyzing multivariate item response data that exhibit zero inflation, maximum inflation, and heaping at preferred digits. The modeling framework combines approaches from three literatures: item response theory (IRT) models for multivariate count data, latent variable models for heaping and extreme responding, and mixture IRT models. Data from the Behavioral Risk Factor Surveillance System are used as a motivating example. Practical implications are discussed, and recommendations are provided for researchers who may wish to use count items on questionnaires.

Keywords
count data, zero inflation, heaping, item response theory

Count data are prevalent in the behavioral and health sciences. Statistical methods for the analysis of univariate count outcomes have existed for several decades and are commonly variants of the log-linear model, including Poisson regression (e.g., Agresti, 2002; Cameron & Trivedi, 2013; McCullagh & Nelder, 1989), negative binomial regression (e.g., Hilbe, 2011), and their zero-inflated extensions (e.g., Lambert, 1992). These models have widespread application in fields such as psychology (e.g., Lewis et al., 2010), medicine (e.g., Roberts & Brewer, 2010), economics (e.g., Deb & Trivedi, 1997), among others. Loeys, Moerkerke, De Smet, and Buysse (2012) recently published a review of some of the current challenges and proposed solutions to modeling univariate count outcomes in psychological research.

Questionnaires that comprise multiple items eliciting count responses are becoming increasingly common, particularly in public health. Often, these surveys are designed to assess the severity of symptoms, asking respondents to recall the frequency of various thoughts or behaviors over a specified period of time. For example, a survey measuring alcohol dependence may ask the respondent to report the number of drinks he or she consumes during a typical week. While statistical methods for the analysis of a single count outcome are widely available, methods for modeling multivariate count outcomes on questionnaires are considerably less well-developed. One approach might be to modify traditional item response theory (IRT) techniques, invoking a log link function in place of the usual logit link and a Poisson distribution in place of a Bernoulli or multinomial conditional response distribution. However, if one examines most self-report count data more closely, a number of additional challenges surface that require a more complex methodological approach.

Figure 1 shows a histogram of responses to 4 items on the Behavioral Risk Factor Surveillance System (BRFSS; Centers for Disease Control and Prevention, 1984–present). These items ask respondents to report the number of days in the past 30 days they have experienced a specific symptom, thought, or behavior. It is clear from the histograms that the observed responses do not follow a standard count distribution (e.g., Poisson, negative binomial). Not only is there a very large proportion of respondents reporting 0 days, much larger than would be expected from a standard count distribution, but there is also a substantial proportion of respondents reporting the maximum of 30 days. Further, there is a noticeable inflation at days that are multiples of 5, an example of a phenomenon known as heaping in the biostatistics literature (e.g., H. Wang & Heitjan, 2008). Simple modifications to a conventional IRT model are not likely to account for the potential subpopulations and individual differences that manifest in Figure 1. This research attempts to address the challenges of modeling multivariate count data with inflation and heaping by combining methodological approaches from three distinct but related literatures: IRT models for multivariate count data (L. Wang, 2010), latent variable models for heaping (H. Wang & Heitjan, 2008) and extreme responding (Böckenholt, 2012; Bolt & Johnson, 2009; De Boeck & Partchev, 2012; Thissen-Roe & Thissen, 2013), and mixture IRT models (Finch & Pierson, 2011; Finkelman, Green, Gruber, & Zaslavsky, 2011; Sawatzky, Ratner, Kopec, & Zumbo, 2012; Wall, Park, & Moustaki, 2015).

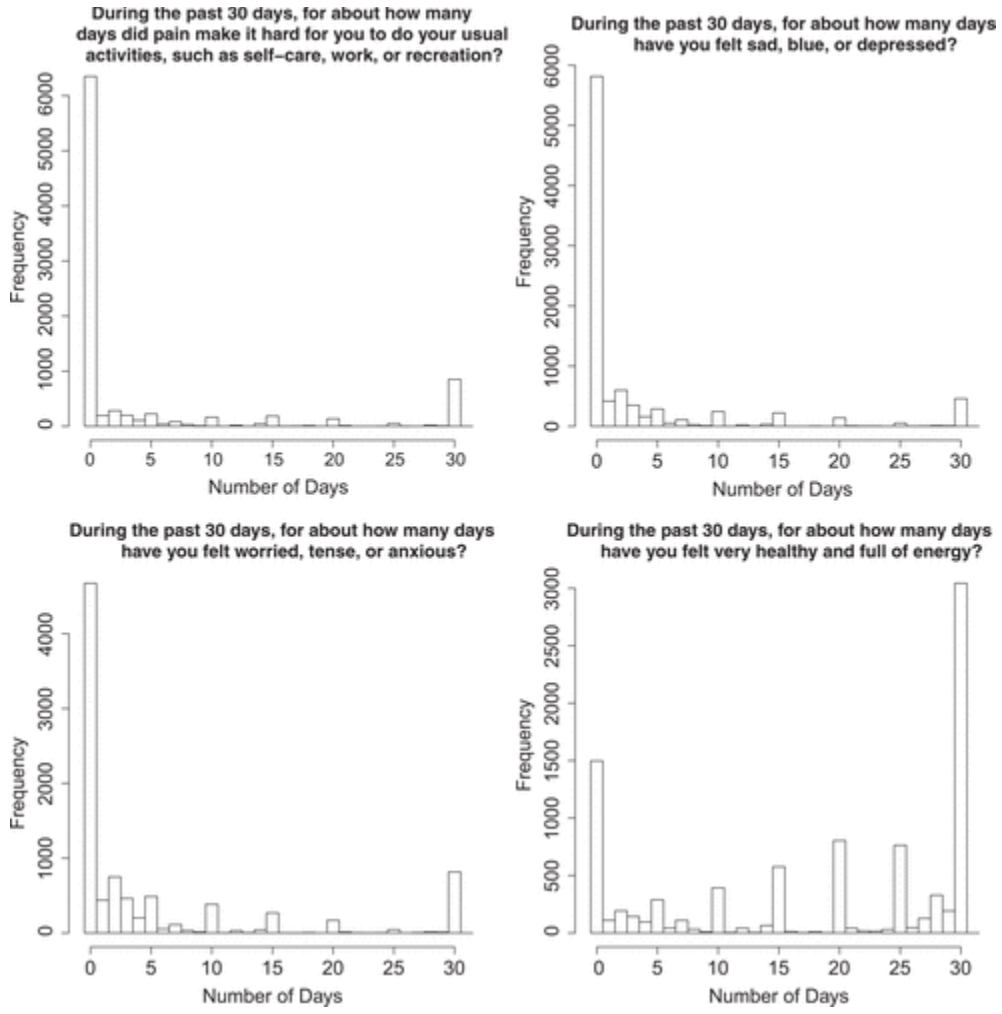

Figure 1. Frequency histograms for four general emotional health items from the Behavioral Risk Factor Surveillance System (2014) eliciting count responses.

Item Response Models for Multivariate Count Data

Several researchers have proposed psychometric models for multivariate count data (Böckenholt, Kamakura, & Wedel, 2003; L. Wang, 2010; Wedel, Böckenholt, & Kamakura, 2003). Perhaps most relevant for data like those from the BRFSS, L. Wang (2010) developed an IRT model for zero-inflated count data, highlighting its application to the analysis of substance use frequencies. Such data often exhibit zero inflation because a subset of respondents abstain from substance use and endorse zero for every item. Based on Lambert's (1992) original zero-inflated Poisson regression model, Wang's zero-inflated Poisson-IRT (ZIP-IRT) model is a latent variable mixture model that accounts for two different response processes: the zero process that relates to whether the event (i.e., alcohol consumption) has a chance of occurring at all and the Poisson process that relates to the expected count (i.e., number of alcoholic drinks consumed), given that the event has a chance of occurring. L. Wang (2010) uses a logit link function and Bernoulli conditional response distribution to model the probability of being in the Poisson process, and a log link function and Poisson conditional response distribution to model the expected count, given that someone is in the Poisson process. Each model component has a set of item discrimination and location parameters as well as a single latent variable that represents substance dependency.

While Wang's model provides an item response modeling framework for analyzing the psychometric properties of zero-inflated multivariate count data, it assumes that the nonperfect zero state is a Poisson process. Real-world data analysis suggests that the Poisson distribution rarely describes observed count responses; this is especially true of retrospective self-report data in which heaping is present (e.g., H. Wang & Heitjan, 2008), as illustrated in Figure 1. For this reason, a more flexible modeling framework may be useful.

Heaping and Response Style

Modeling heaping in univariate count outcomes has been a subject of interest in the biostatistics literature (Heitjan & Rubin, 1990; Ridout & Morgan, 1991; H. Wang & Heitjan, 2008). To account for heaping and digit rounding in self-reported counts of cigarette use, H. Wang and Heitjan (2008) introduced a model in which the observed cigarette count is a function of both an unobserved true cigarette count and a latent class "heaping behavior" variable. Others have also developed models for heaping and data coarsening in univariate outcomes, such as rounding in self-reported age (Heitjan & Rubin, 1990), digit preference in retrospective reporting of women's number of menstrual cycles before a positive pregnancy (Ridout & Morgan, 1991), and rounding in clinician-reported measurements from ultrasound images (Wright & Bray, 2003). These models have applications to univariate count outcomes; however, review of the biostatistics literature has not uncovered methods of accounting for heaping in multivariate count outcomes.

While the psychometrics literature does not include specific models for heaping in multivariate count data, research on extreme responding on surveys comprising Likert-type items addresses a similar concept within an IRT framework. Some have used multidimensional IRT models to account for an "extreme response style" (ERS) latent variable (Bolt & Johnson, 2009; Bolt & Newton, 2011). Bolt and Newton (2011) describe a multidimensional nominal response model (NRM) in which there is a latent variable related to the construct of interest as well as a latent variable for ERS. Others have approached the topic from a decision tree perspective, in which the observed responses manifest from a sequence of internal decisions (Böckenholt, 2012; De Boeck & Partchev, 2012; Thissen-Roe & Thissen, 2013). Böckenholt (2012) and De Boeck and Partchev (2012) propose a tree structure for capturing individual differences in response style, arguing that it is possible that more than one response process is at play when someone responds to a Likert-type questionnaire item. Thissen-Roe and Thissen (2013) adopt a similar view, positing that response direction and response intensity to Likert-type items can be modeled by two stage-like processes. Biostatistical models for heaping and psychometric models for extreme responding developed from different methodological frameworks; however, both approaches converge on the idea of a latent variable underlying individual differences in response style. The count response methodology from the biostatistics literature, coupled with the extreme responding methodology from the psychometrics literature, lay the groundwork for the development of an IRT model that can accommodate multivariate count responses that also exhibit heaping.

Mixture IRT

Unlike traditional IRT models, mixture IRT models assume that the observed responses are sampled from a population that has a number of distinct subpopulations (Rost, 1990, 1997; von Davier & Rost, 2006). Under the assumption of local independence, the marginal mixture distribution of the observed item responses $\boldsymbol{U} = (u_1, \ldots, u_J)$ is expressed

(1)

$$P(u_1,\ldots,u_J)=\sum_{g=1}^{G}\ \pi_g\left(\int_\theta \prod_j\ P_{gj}(u_j|\theta)\varphi(\theta|g)d\theta\right),$$

where $\int_\theta\ \prod_j\ P_{gj}(u_j|\theta)\varphi(\theta|g)d\theta$ is the conditional probability of observing response pattern $(u_1,\ldots,u_J)$ in subpopulation g. Observed responses and latent variable densities are conditional on the subpopulation, with πg denoting the proportion of the population belonging to subpopulation g. Both the latent variable θ and class membership g are treated as unobserved variables and are estimated as part of the model.

Mixture IRT has been applied to the analysis of substance use or risky behavior scales, where many of the items may not be applicable to substantial proportions of the population (Finch & Pierson, 2011; Finkelman et al., 2011; Muthen & Asparouhov, 2006; Sawatzky et al., 2012; Wall et al., 2015). Respondents belonging to one of the latent classes—for example, a subgroup of people who abstain from drinking but are nonetheless asked a series of questions relating to symptoms of alcohol dependence—may not engage with the items the same way as other subgroups in the population, and the mixture IRT modeling can help to account for both individual- and group-level differences.

When many of the respondents in the population possess none or very low levels of the construct being measured, such as a group of individuals not endorsing any of the criteria on a symptoms checklist, it is plausible that the latent variable follows a mixture distribution with a zero-inflated component. Such constructs are often referred to as “unipolar” (Reise & Waller, 2009; Wall et al., 2015). In clinical assessment, there may also be a smaller subset of respondents who are extreme at the other end of the latent variable, endorsing all possible symptoms on a checklist. Finkelman, Green, Gruber, and Zaslavsky (2011) referred to the high frequency of respondents with the maximum observed score as *K* inflation. To circumvent the challenges associated with measuring low-prevalence psychiatric disorders, Finkelman et al. (2011) developed a latent class IRT model to account for extreme subpopulations. One latent class describes individuals with no symptoms; the IRT model for this class is degenerate. A second latent class describes the people exhibiting all of the symptoms; the IRT model for this class is also degenerate. The remaining latent class, labeled the graded class, describes people along the severity continuum that is implied by a traditional IRT model with a normal population density. More recently, Wall, Park, and Moustaki (2015) proposed a similar model for measuring psychiatric disorder severity in a zero-inflated population. In both applications, the presence of respondents from a potentially heterogeneous population requires mixture IRT in place of an IRT model that assumes a normal population density.

The Proposed Latent Class IRT Model

A review of the literature suggests three methodological approaches for solving three distinct problems in measurement: zero-inflated Poisson psychometric models for the analysis of multivariate zero-inflated count data (L. Wang, 2010), latent variable models to account for heaping in univariate count outcomes (H. Wang & Heitjan, 2008) and extreme responding in multivariate Likert-type items (Böckenholt, 2012; Bolt & Johnson, 2009; Bolt & Newton, 2011; De Boeck & Partchev, 2012; Thissen-Roe & Thissen, 2013), and mixture IRT models for clinical assessment in heterogeneous populations (Finch & Pierson, 2011; Finkelman et al., 2011; Sawatzky et al., 2012; Wall et al., 2015). All three methods have utility in particular scenarios; however, we are unaware of any existing unifying framework for the analysis of multivariate count outcomes that also accounts for zero inflation,

maximum inflation (*K* inflation), and heaping. This research borrows elements from all three methodological approaches in developing a latent class IRT model for multivariate zero-inflated count data that are sampled from a potentially heterogeneous population. To make the presentation concrete, we use as an illustration the responses to the four count items from the BRFSS that are shown in Figure 1.

According to the general latent class model, the unconditional probability of observed response u_j to item *j* can be expressed

(2)

$$P_j = \sum_{g=1}^{G} \pi_g P_{gj}(U_j = u_j|\theta),$$

in which *g* denotes latent class membership, π_g specifies the probability of belonging to latent class *g*, and $P_{gj}(U_j = u_j|\theta)$ is the conditional probability of observing response u_j from someone in latent class *g*, where $\sum_{g=1}^{G} \pi_g = 1$ (Hagenaars & McCuthcheon, 2002). To reflect the potential subpopulations in data like those in Figure 1, we propose four mutually exclusive latent classes and two IRT models. We express the conditional probabilities $P_{gj}\left(U_j = u_j\middle|\theta\right)$ that are given by the two IRT models as IRT trace lines, $T_j(U_j = u_j|\theta)$.

One latent class, with probability π0, describes some, perhaps many, of the people who respond 0 days to all 4 items with response vector $\mathbf{U} = 0 = (0,0,0,0)$. This class may represent people who are at a floor level of the latent variable, or it may represent a subset of individuals for whom the items do not apply. Similarly, a second latent class, with probability π30, describes some, perhaps many, of the people who respond 30 days to all 4 items with response vector $\mathbf{U} = 30 = (30,30,30,30)$. These two latent classes are referred to as the zero class and the maximum class, respectively. The IRT models for the zero and maximum classes are degenerate in the sense that they do not depend on individual differences in the latent variable θ.

In addition to the zero and maximum classes, we propose two "graded" latent classes that describe people falling along the continuum of the latent variable. One graded class with probability π_e is referred to as the exact count class and comprises the subset of people whose responses follow a standard count distribution. These are individuals who possess some level of the latent variable and report the exact number of days they experience a symptom, regardless of whether the counts are multiples of 5. To model the conditional probability of count response u_j from members of the exact count class, we use a negative binomial IRT model to allow for overdispersion:

(3)

$$T_j(U_j = u_j|\theta) = \left(\frac{\Gamma(u_j + \delta_j^{-1})}{\Gamma(u_j + 1)\Gamma(\delta_j^{-1})}\right)\left(\frac{\delta_j^{-1}}{\delta_j^{-1} + \exp(a_j\theta + c_j)}\right)^{\delta_j^{-1}}\left(\frac{\exp(a_j\theta + c_j)}{\delta_j^{-1} + \exp(a_j\theta + c_j)}\right)^{u_j},$$

in which a_j, c_j, and δ_j are the discrimination, intercept, and overdispersion parameters for item *j*, respectively, and ϑ is the latent variable being measured by the 4 items. The larger the *a* parameter,

the more effective the item is in separating individuals on the latent variable. The *c* parameter is the expected count for someone at the mean level of the latent variable ($\vartheta = 0$).

The remaining latent class with probability π_r is referred to as the rounding/selected response class; it includes individuals who respond only with counts that are multiples of 5. Instead of treating the item as having an open-ended count scale, these people treat the item as having a smaller, fixed number of response categories: {0,5,10,15,20,25,30}. To model the conditional probability of observing count response u_j from members of the rounding/selected response class, an NRM can be used (Bock, 1972; Thissen, Cai, & Bock, 2010). The trace line for the NRM is expressed

(4)

$$T_j(U_j = k_j|\theta) = \frac{\exp(a_{jk}\theta + c_{jk})}{\sum_{m=1}^{M} \exp(a_{jm}\theta + c_{jm})},$$

in which k_j corresponds to the response category for item *j*. To avoid confusion with count response u_j that can take on any nonnegative integer value up to 30, we adopt the alternative notation k_j such that $k_j = u_j \in \{0,5,10,15,20,25,30\}$; that is, k_j can only take on values of u_j that are multiples of 5. In Equation 4, a_{jk} is the slope parameter and c_{jk} is the intercept parameter, both for response category *k*, and *M* is the total number of response alternatives. For model identification, the constraints $a_1 = c_1 = 0$ are imposed. An advantage of the NRM over other IRT models for polytomous data is that the ordering of response categories can be examined empirically after fitting the NRM to the data. Figure 2 depicts a tree diagram of the proposed latent response processes that result in each of the 31 possible observed counts for a single item.

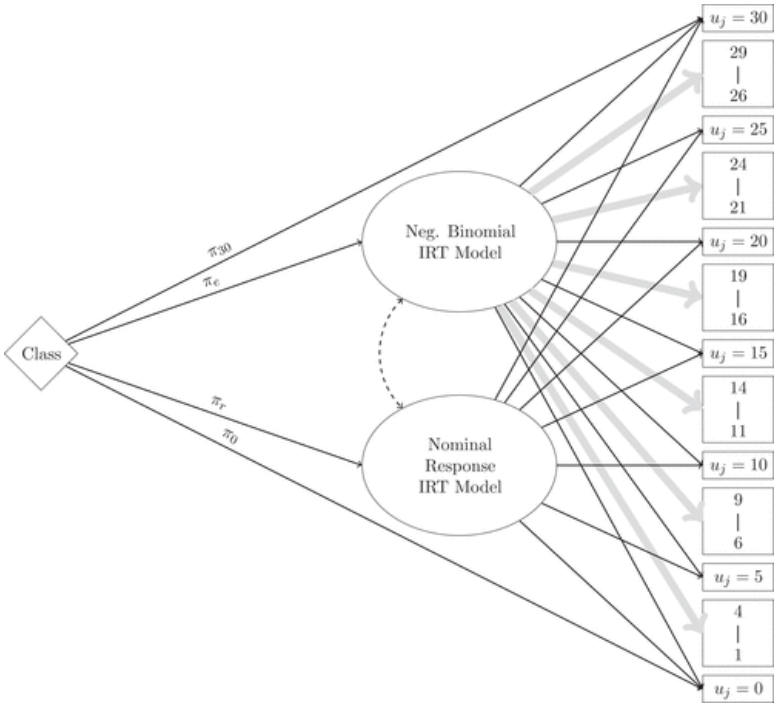

Figure 2. Tree diagram showing the proposed response processes. Consider an observed count of 0. The respondent may be a member of the zero class and select 0 days for every item; this option is represented by the direct path from the item to a zero response, without passing through either of the two item response theory (IRT) models. It is also possible that the person is a member of the exact count class and reports 0 days. This option is represented by the indirect path to the zero response through the negative binomial IRT model. It is also possible that the respondent is a member of the rounding/selected response class and reports 0 days. Instead of passing through the negative binomial IRT model, these individuals arrive at a zero response via the nominal response IRT model. Multiple potential internal processes also underlie the nonzero responses.

The Full Latent Class IRT Model

Let I0, Ie, Ir, and I30 be indicator variables denoting membership in the zero, exact count, rounding/selected response, and maximum classes, respectively, with probabilities π0, πe, πr, and $\pi_{30} = 1 - \pi_0 - \pi_e - \pi_r$. Assuming these four mutually exclusive latent classes, the general latent class model in Equation 2 can be written

(5)

$$\begin{aligned} P_j = \pi_0[P_{0j}(U_j = 0) = 1; P_{0j}(U_j \neq 0) = 0] \\ +\pi_e T_{ej}(U_j = u_j|\theta, I_e = 1; a_j, c_j, \delta_j) \\ +\pi_r T_{rj}(U_j = k_j|\theta, I_r = 1; a_{jk}, c_{jk}) \\ +\pi_{30}[P_{30j}(U_j = 30) = 1; P_{30j}(U_j \neq 30) = 0], \end{aligned}$$

where $u_j = \{0,1,\ldots,30\}$and $k_j = u_j \in \{0,5,10,15,20,25,30\}$. Assuming four count items, let $N_{\mathbf{0}}$ be the number of people with response pattern **U** = **0** and $N_{\mathbf{3}}$ be the number of people with response pattern **U** = **30**. Let N_u be the number of people with response pattern **U** = **u**. For notational simplicity, let **e** be any response pattern that includes at least one non-0 and non-30 exact count and *k* be any response pattern that includes only multiples of 5, excluding response patterns **0** and **30**. Given response patterns **U** = **u**, the log likelihood of item parameters $\boldsymbol{\alpha} = \{\mathbf{a}_j, \mathbf{c}_j, \mathbf{a}_{jk}, \mathbf{c}_{jk}, \delta_j\}$ for items *j* = 1, …, *J*, as well as the latent class proportions π_0, π_e, π_r, and π_{30}, can be expressed

(6)

$$\begin{aligned} \log L(\boldsymbol{\alpha}, \pi_0, \pi_e, \pi_r, \pi_{30}; \{\mathbf{u}\}_1^J) = N_0 \log[\pi_0 + \pi_e T(\mathbf{U} = 0|\boldsymbol{\alpha}, \theta, I_e = 1) + \pi_r T(\mathbf{U} = 0|\boldsymbol{\alpha}, \theta, I_r \\ = 1)] + N_{30} \log[\pi_{30} + \pi_e T(\mathbf{U} = 30|\boldsymbol{\alpha}, \theta, I_e = 1) + \pi_r T(\mathbf{U} = 30|\boldsymbol{\alpha}, \theta, I_r \\ = 1)] + \sum_{\mathbf{u} \notin \{\mathbf{0},\mathbf{30},\mathbf{e}\}} N_u \log[\pi_r T(\mathbf{U} = \boldsymbol{k}|\boldsymbol{\alpha}, \theta, I_r = 1) + \pi_e T(\mathbf{U} = \boldsymbol{k}|\boldsymbol{\alpha}, \theta, I_e \\ = 1)] + \sum_{\mathbf{u} \notin \{0,30,\boldsymbol{k}\}} N_{\boldsymbol{u}} \log[\pi_e T(\mathbf{U} = \mathbf{u}|\boldsymbol{\alpha}, \theta, I_e = 1)]. \end{aligned}$$

In Equation 6, $T(\mathbf{U} = \mathbf{k}|\boldsymbol{\alpha}, \theta, I_r = 1)$ traces the conditional probability of observing response pattern $\mathbf{k} = \mathbf{u}$: $u_j \in \{0,5,10,15,20,25,30\}$ for someone in the rounding/selected response class influenced by latent variable ϑ, $T(\mathbf{U} = \mathbf{k}|\boldsymbol{\alpha}, \theta, I_e = 1)$ traces the conditional probability of observing response pattern $\mathbf{k} = \mathbf{u}$: $u_j \in \{0,5,10,15,20,25,30\}$ for someone in the exact count

class, and $T(\mathbf{U} = \mathbf{u}|\boldsymbol{\alpha}, \theta, I_e = 1)$ traces the conditional probability of observing response pattern with only exact counts for someone in the exact count class. The proportions of people in each of the four latent classes are estimated as part of the model.

Parameter estimation for the latent class IRT model can done with maximum likelihood using nlm, R's nonlinear optimizer that directly minimizes a user-specified function using a Newton-type algorithm; to implement maximum likelihood, we used nlm to minimize $(-1) \times \log L$ in Equation 6. A total of 63 parameters were estimated: $\{\mathbf{a}_{jk}, \mathbf{c}_{jk}, \mathbf{a}_j, \mathbf{c}_j, \delta_j, \pi_0, \pi_{30}, \pi_r\}$. Class membership proportions were estimated as logits, and standard errors were computed using the delta method. R code is available upon request from the first author.

Simulation

To test the software implementation, we simulated 10,000 responses to four hypothetical count items with a 0 to 30 response scale. First, each of 10,000 observations was assigned to one of the four latent classes. Then, depending on the latent class membership, item response data were generated: For members of the zero and maximum classes, response patterns of **U** = (0, 0, 0, 0) and **U** = (30, 30, 30, 30) were produced, respectively. For members of the exact class, response patterns were generated from a negative binomial IRT model; for members of the rounding/selected response class, response patterns were generated from a nominal response IRT model. To evaluate the implementation of parameter estimation in R, we then fit the proposed latent class IRT model to the simulated data. The data generating parameters are shown in Table 1. The model converged after 285 iterations, requiring approximately 13 hr on a desktop computer with a quad-core 2.4 gigahertz Intel Core processor and 4 gigabyte of RAM. The parameter estimates plotted against the data generating parameters are shown in Figure 3. This figure shows that when the proposed model is fit to simulated data with known population parameters, the R program recovers both the IRT parameters and the proportions of respondents in each latent class.

TABLE 1.
The 63 Data Generating Parameters Used for the Test of Software Implementation (a_1 and c_1 Were Fixed to 0 for All Items)

Item Parameter	Item #1	Item #2	Item #3	Item #4
Nominal response item parameters				
a_1	0.00	0.00	0.00	0.00
a_2	1.00	0.90	0.85	1.95
a_3	1.00	1.25	1.10	1.65
a_4	1.50	1.50	1.15	1.00
a_5	1.50	1.00	1.85	1.65
a_6	2.00	1.40	2.25	0.90
a_7	1.25	1.35	1.75	1.25
c_1	0.00	0.00	0.00	0.00
c_2	0.30	−0.35	−0.15	−0.75
c_3	−0.75	−0.75	−0.25	−0.30
c_4	−0.50	−1.25	−0.45	−1.25
c_5	−0.25	−0.55	−0.75	−0.80
c_6	−0.80	−0.80	−0.90	−1.00
c_7	−1.00	−1.20	−0.95	−1.25
Negative binomial item parameters				
a	1.15	1.15	1.30	0.80
c	1.10	0.00	1.20	1.30
δ	0.40	0.50	0.48	0.34
Latent Classes	Zero	Exact Count	Rounding	Maximum
Proportion	0.30	0.20	0.40	0.10

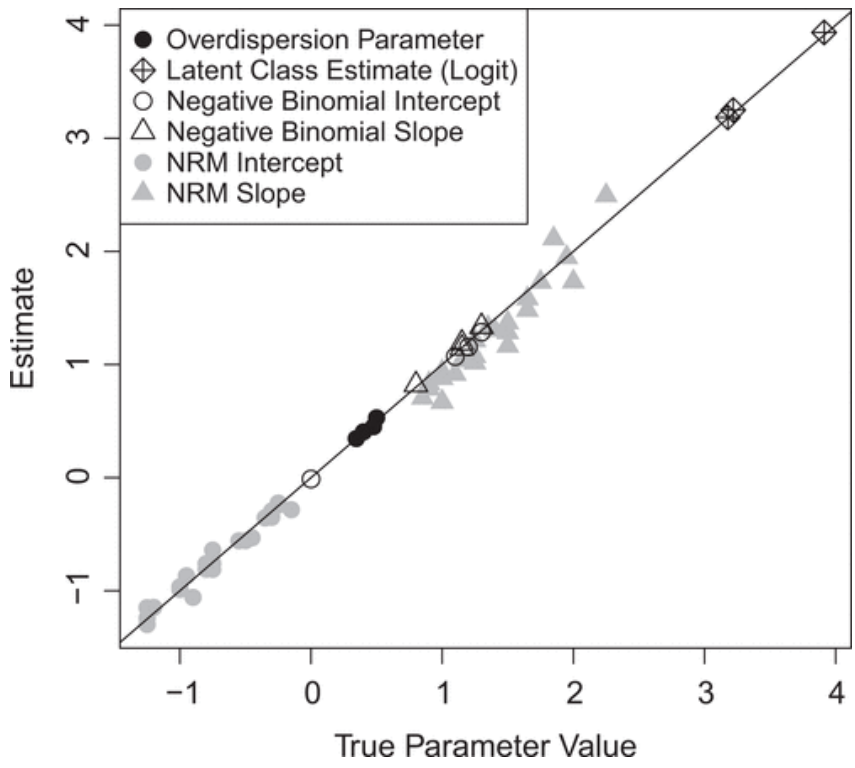

Figure 3. Parameter estimates from the latent class item response theory model (*y*-axis) relative to the true/data generating parameters (*x*-axis).

Empirical Analysis of the BRFSS Scale

The analytic sample consisted of 9,042 individuals who responded to a set of four count items from the 2014 version of the BRFSS. The 4 items form a subscale measuring emotional health status, with each item response reported on a 0- to 30-day scale. For the pain, depression, and anxiety items, an increasing number of days suggest worse health; for the energy item, an increasing number of days suggest better health. To maintain consistency of scale direction, the energy item was reverse coded. The names “pain,” “depressed,” “anxious,” and (reversed) “energy” are used to refer to each of the items.

When fit to empirical data, the latent class IRT model converged after 472 iterations, requiring approximately 16.5 hr. Estimates of the proportions of individuals belonging to each of the four latent classes are in the last row of Table 2. Nearly one third of respondents in the sample either (a) treated the items as multiple-choice questions instead of open-ended counts or (b) rounded their answers to the nearest multiple of 5. Only about half of the respondents used the full range of the open-ended count scale in selecting nonmultiple of 5 counts. While 24% of the sample endorsed 0 for every item, only 16% of respondents were estimated to belong to the zero class; such a decomposition indicates that someone endorsing zero for all 4 items has approximately 67% probability of belonging to the zero class and 33% probability of belonging to one of the two graded classes. Of the people who endorsed 30 for every item, 68% were estimated to belong to the maximum class. According to the model, the remaining 45 people with an all-30 response pattern fall along the continuum of the latent variable measured by the scale: poor emotional health.

TABLE 2.
Parameter Estimates (With Standard Errors) of the Latent Class Item Response Theory Model Fit to the Behavioral Risk Factor Surveillance System Data

Item Parameter	# Days Pain	# Days Depressed	# Days Anxious	(30-) # Days Energy
Nominal response item parameters				
a_1	0.00 (—)	0.00 (—)	0.00 (—)	0.00 (—)
a_2	0.64 (0.13)	6.75 (0.17)	1.33 (0.13)	−0.04 (0.08)
a_3	0.90 (0.12)	18.02 (0.12)	2.35 (0.10)	0.53 (0.08)
a_4	1.24 (0.09)	24.07 (0.08)	3.18 (0.08)	0.91 (0.08)
a_5	1.43 (0.10)	27.63 (0.08)	4.03 (0.07)	1.65 (0.08)
a_6	1.61 (0.14)	30.48 (0.13)	5.36 (0.11)	1.86 (0.08)
a_7	1.08 (0.07)	24.28 (0.09)	3.37 (0.07)	1.45 (0.06)
c_1	0.00 (—)	0.00 (—)	0.00 (—)	0.00 (—)
c_2	−2.91 (0.10)	−4.35 (0.10)	−1.46 (0.07)	−0.45 (0.06)
c_3	−3.25 (0.12)	−13.55 (0.12)	−2.16 (0.08)	−0.26 (0.06)
c_4	−3.24 (0.10)	−20.05 (0.01)	−2.90 (0.08)	−0.38 (0.06)
c_5	−3.64 (0.12)	−26.00 (0.14)	−4.12 (0.10)	−1.24 (0.08)
c_6	−4.80 (0.19)	−32.76 (0.28)	−7.78 (0.22)	−1.69 (0.09)
c_7	−1.95 (0.06)	−20.43 (0.12)	−2.51 (0.07)	0.08 (0.05)
Negative binomial item parameters				
a	1.26 (0.07)	1.74 (0.03)	1.33 (0.03)	0.72 (0.03)
c	1.24 (0.05)	0.15 (0.03)	1.17 (0.02)	2.55 (0.02)
δ	6.51 (0.23)	0.97 (0.06)	0.85 (0.04)	1.24 (0.04)

Latent Classes	Zero	Exact Count	Rounding	Maximum
Proportion	0.16 (0.04)	0.52 (0.02)	0.31 (0.03)	0.01 (0.10)

Note. $N = 9{,}042$.

IRT parameter estimates can also be found in Table 2. To examine how closely the IRT parameters estimated from the model reproduce the empirical response distributions, we simulated 9,042 responses to the 4 items based on the estimates in Table 2. While comparison of the empirical and model-implied distributions does not allow an analysis of model fit at the response pattern level, comparing the two distributions can help inform whether the IRT models are appropriate for the data. The empirical response distributions for pain, depressed, and anxious—shown in the first three columns of Figure 4—are reproduced fairly well, suggesting that the negative binomial and nominal response IRT models are appropriate model choices for these 3 items; however, the empirical response distribution for (reversed) energy, which is shown in the rightmost column of Figure 4, is not reproduced nearly as well. This is because the negative binomial distribution is unable to account for the increasing number of people reporting counts toward the upper limit of the 0 to 30 count range. Recommendations for alternative model specifications that may better accommodate the energy item are described in the Discussion section.

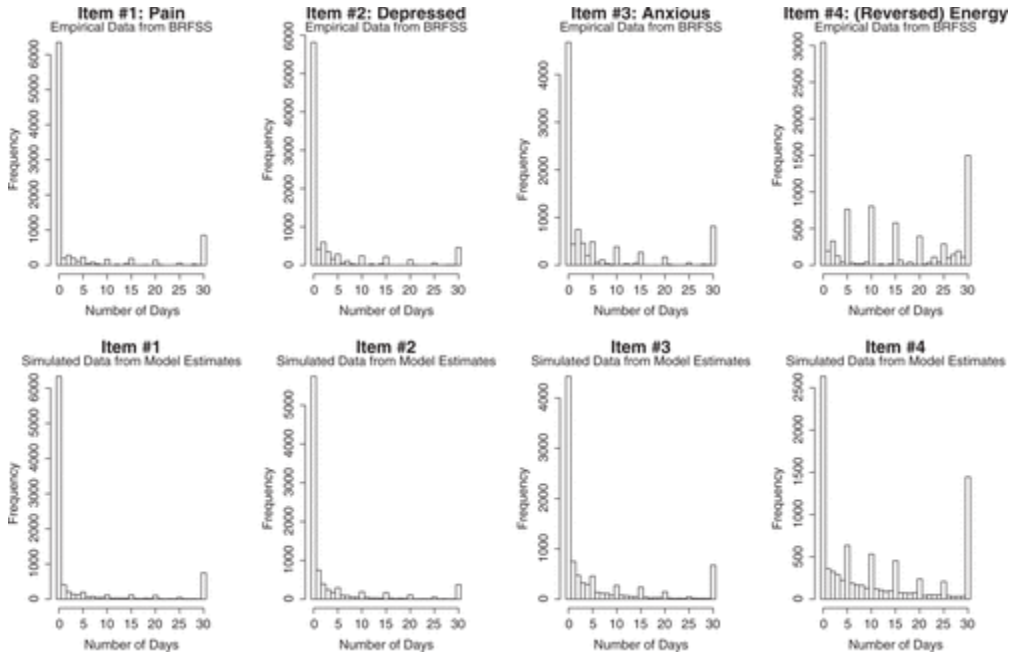

Figure 4. Upper panel: histograms of the four count items on the Behavioral Risk Factor Surveillance System (*N* = 9,042). Lower panel: histograms of responses simulated from the estimated item response and latent class parameters in Table 2 (*N* = 9,042).

Interpretation of the Rounding/Selected Response Class

Because the NRM was used to parameterize the response process for members of the rounding/selected response class, the ordering of response categories could be examined empirically. The NRM item parameters show a relatively linear trend: As the number of days category increases, the *a* parameters tend to increase and *c* parameters tend to decrease nearly linearly, with the exception of the *a* and *c* parameters corresponding to the 30-day category. Anxious and depressed are most discriminating, suggesting that these 2 items are more strongly related to the poor emotional health latent variable than either pain or (reversed) energy. The discriminating ability of these items can also be seen in the trace lines in Figure 5. The trace lines for anxious and depressed tend to function more similarly to each other than either of the other items, likely because the poor emotional health latent variable is really defined by these 2 items, with pain and energy acting as ancillary items.

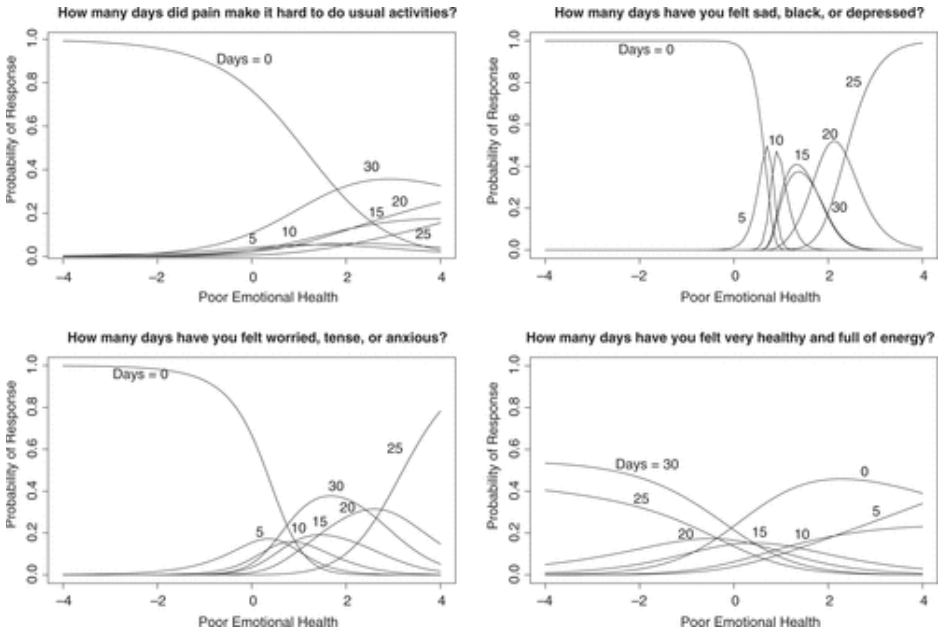

Figure 5. Nominal response trace lines for the rounding/selected response class.

The trace lines in Figure 5 reveal several other item characteristics. First, the 0- and 30-day response categories tend to be associated with the greatest probabilities of endorsement, regardless of someone's level of the latent variable. This phenomenon is particularly salient in examining the trace lines for the pain and energy items, where at every point along the latent variable, 0 days or 30 days always has a higher probability of endorsement than any of the other response categories. Second, Figure 5 shows that the response categories tend to be in an increasing order for 5 days through 25 days: As one's level of poor emotional health increases, so does the probability of endorsing a response category that represents a greater number of days (or a lesser number of days for the energy item). The increasing order does not hold for the 30-day response option, however; this is most clearly seen in the trace lines for depressed and anxious, where for people at high levels of poor emotional health, a response of 25 days is actually associated with higher probability of endorsement than a response of 30 days.

The pain, depressed, and anxious items tend to discriminate only among individuals who fall at or above average levels of poor emotional health. This is seen in Figure 5, where for these 3 items, it is not until $\vartheta \geq 0$ that the trace lines cross. The relatively flat trace lines for the energy item suggest that this item is only weakly related to the latent variable; people tend to endorse a smaller number of days for this item regardless of their level of poor emotional health. Even for someone at low levels of ϑ, the probability of endorsing 25 days for energy (or equivalently, 5 days for the reverse-coded energy item) is greater than 0.3. For someone who is at the average level of poor emotional health, there is near-equal probability of endorsing 0 days or 30 days for the energy item. Compared to the other 3 items, the endorsement of lower counts is common for the energy item.

Interpretation of the Exact Count Class

The negative binomial item parameter estimates for the exact count class are also shown in Table 2. The item discrimination for a count item can be interpreted as the log expected change in the number of days associated with a 1 standard deviation (*SD*) increase in poor emotional health; this value can then be exponentiated to be placed on a more interpretable scale. For a 1 *SD* increase in poor

emotional health, one expects an additional 3.53 days for pain, 5.70 days for depressed, and 3.78 days for anxious; one expects 2.05 fewer days for energy. The *c* parameter, which is the item intercept, is interpreted as the log expected number of days of a particular symptom for someone who is at the average level of poor emotional health. For someone who is at the average level of poor emotional health, one expects 3.46 days for the pain item, 1.16 days for the depressed item, 3.22 days for the anxious item, and 17.19 days for the energy item (or equivalently, 12.81 days for the reversed energy item).

In fitting a count IRT model to item response data, there are multiple ways to graphically display the item parameters. One option is to plot the trace lines associated with each count IRT model; the negative binomial trace lines for these four count items are shown in the upper two rows of Figure 6. Within a particular plot, each curve corresponds to 1 of the 31 possible open-ended counts, where increasing levels of poor emotional health indicate greater probabilities of endorsing higher counts. Similar to the NRM trace lines, flatter trace lines suggest a weaker relationship between the item and the latent variable, and the location of the trace lines indicates where on the latent variable continuum the item is most discriminating. One can choose a level of the latent variable and find the endorsement probability that corresponds to each of the 31 counts. For the pain item, for example, someone with ϑ = 0 has roughly a .60 probability of endorsing 0 days, a .1 probability of endorsing 1 day, a .05 probability of endorsing 2 days, and a near-zero probability of endorsing any of the counts greater than 3 days.

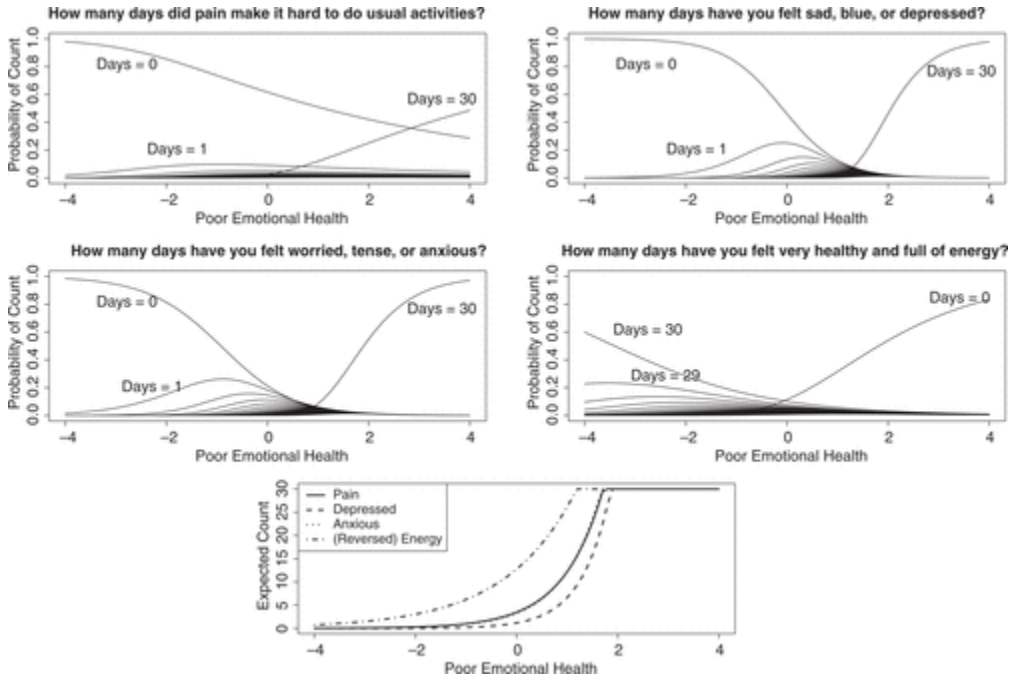

Figure 6. First two rows: negative binomial trace lines describing probability of count endorsement for members of the exact count class. Last row: expected endorsed count as a function of the latent variable for members of the exact count class.

Perhaps a more intuitive approach to visualizing the relationship between the latent variable and the item responses is to plot the expected counts for each item as a function of the latent variable. Such a plot is shown in the lowermost panel of Figure 6. The low discriminating power of the energy item compared to the depressed and anxious items is shown with its flatter slope. The relatively high

expected counts for (reversed) energy—even for people with low levels of poor emotional health—is also shown in Figure 6. For someone who is 2 *SD*s below average on poor emotional health, the expected number of days for pain, depressed, and anxious is approximately 0; for someone at this same level of poor emotional health, the expected number of days for (reversed) energy is 3.

Overall, both pain and energy do a poor job of separating individuals on the latent variable. As was the case for the rounding/selected response class, the 0- and 30-day response options dominate the trace line plots: Across all levels of poor emotional health, these are almost always the response categories with the highest probabilities of endorsement. Figure 6 also suggests that lower counts are more easily endorsed for the energy item, even for individuals at low levels of the latent variable. For example, someone at low levels of poor emotional health has only a modeled 40% to 60% probability of reporting 30 days for the energy item (or equivalently, 0 days for the reverse–coded energy item); for the other 3 items, however, individuals at low levels of poor emotional health have a nearly 100% probability of endorsing the 0-days response option. One explanation is that the energy item is not as strongly related to poor emotional health as the other 3 items. Alternatively, it may just be that the anxious and depressed items are so strongly related to each other that they define the construct that is being measured by the scale, making the other 2 items appear less relevant.

Scale Scores

According to the latent class IRT model, scale scores depend not only on response patterns but also on latent class membership, and one of the complexities involved in scoring is the uncertainty of latent class membership. Because latent class membership is not known *a priori*, in most cases, it is not possible to directly classify individuals. For many response patterns, there exist two possible class memberships and consequently two plausible scale scores.

Only members of the exact count and rounding/selected response classes fall along the latent variable continuum—this means that only approximately 83% of the population, or 7,498 of the 9,042 people in the sample, should receive scores that are estimates of ϑ values. To account for the 16% and 1% of the population belonging to the zero and maximum classes, respectively, 1,451 all-0 response patterns and 93 all-30 response patterns were removed, resulting in a scoring sample of 7,498 respondents. Removing a proportion of the all-0 and all-30 response patterns yields score distributions that represent those that would be observed in the population of people belonging to one of the two graded classes. Because it is not possible to identify the specific individuals belonging to the zero and maximum classes, scores are discussed only at the population level and not at the individual level.

For a given response pattern $\mathbf{U} = \mathbf{u} = (u_1, u_2, u_3, u_4)$ in the scoring sample, scale scores $\hat{\theta}_{\mathrm{EAP}}$ were computed as the mean of the posterior distribution of ϑ; the posterior distribution is simply the product of the trace lines for each response *u* to Item *j* and the prior density—in this case, a standard normal density (Thissen & Wainer, 2001). We approximated the mean of the posterior distribution using rectangular quadrature,

(7)

$$\hat{\theta}_{\text{EAP}} \approx \frac{\sum_{1}^{q} \prod_{j=1}^{4} T_{jq}(u_j)\theta_q \mathrm{d}\theta}{\sum_{1}^{q} \prod_{j=1}^{4} T_{jq}(u_j)\mathrm{d}\theta},$$

where $T_j(u_j)$ is the trace line for item j. Standard errors of scale scores were computed as the SD of the posterior distribution of ϑ,

(8)

$$SD(\hat{\theta}_{\text{EAP}}) \approx \sqrt{\frac{\sum_{1}^{q} \prod_{i=1}^{4} T_j(u_j)(\theta_q - \hat{\theta}_{\text{EAP}})^2 \mathrm{d}\theta}{\sum_{i}^{q} \prod_{i=1}^{4} T_j(u_j)\mathrm{d}\theta}}.$$

For individuals known to be members of the exact count class (i.e., their response patterns include at least one count that is not a multiple of 5), a single scale score was computed using the estimated negative binomial trace lines for $T_j(u_j)$ in Equation 7. For individuals who could belong to either the exact count or the rounding/selected response class (i.e., their response patterns contained only multiples of 5), two plausible scale scores were computed. The first assumes the person is a member of the exact count class and was computed using the negative binomial trace lines; the second assumes the person is a member of the rounding/selected response class and was computed using the NRM trace lines. Figure 7 displays histograms of the scale scores for members of the exact and rounding/selected response classes.

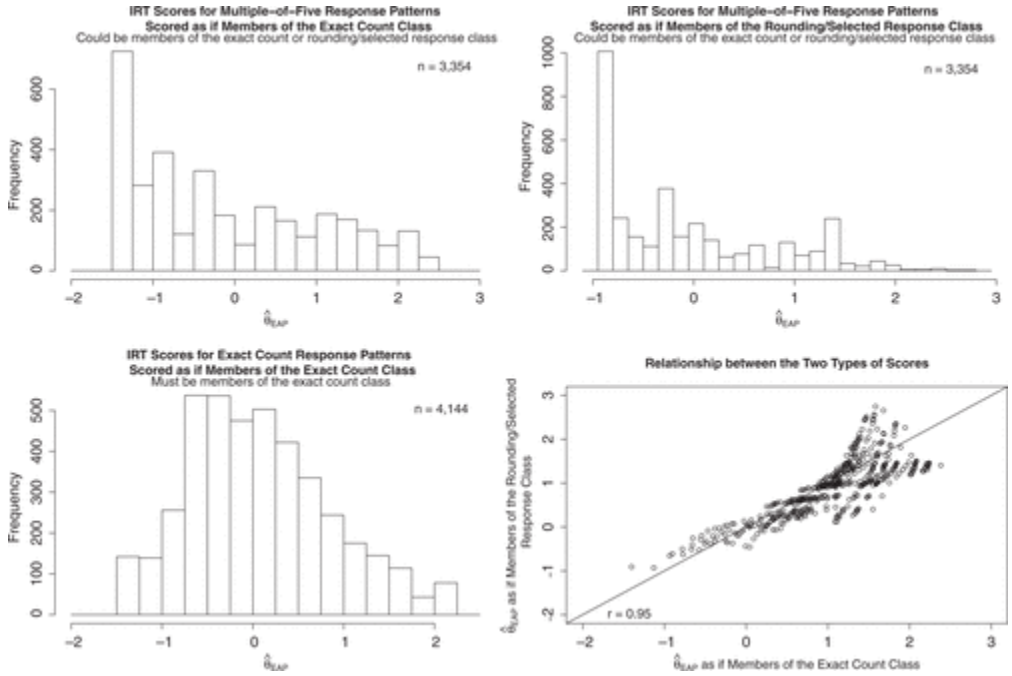

Figure 7. Item response theory scale scores (response pattern expected a posteriori) for members of the exact count and rounding/selected response classes. Upper left: scale scores for multiple of 5 response patterns scored according to the negative binomial trace lines. Upper right: scale scores for multiple of 5 response patterns scored according to the nominal response model trace lines. Lower left: scale scores for nonmultiple of 5 response patterns scored according to the negative binomial trace lines. Lower right: scatterplot showing the relationship between two types of scoring methods for multiple of 5 response patterns.

The lower left panel of Figure 7 shows the scale scores for individuals who must belong to the exact count class. Scale scores range from $\hat{\theta}_{\mathrm{EAP}} = -1.34$, corresponding to response pattern **U** = (0, 0, 0, 1), to $\hat{\theta}_{\mathrm{EAP}} = 2.24$, corresponding to response pattern **U** = (29, 29, 29, 30). The remaining 45% of the scoring sample exhibited response patterns that included only multiples of 5; because this type of response pattern can manifest from either an exact count or rounding/selected response process, two different scores are plausible. The upper left panel of Figure 7 shows a histogram when scores are computed as though the individuals belong to the exact count class; the upper right panel shows a histogram when scores are computed as though the individuals belong to the rounding/selected response class. Both plots exhibit peakedness due to the high frequency of all-0 response patterns. Even though 16% of the people estimated to belong to the zero class were removed from the scoring sample, there are still many people with all-0 response patterns that belong to one of the two graded classes. Table 3 shows scale scores for different response patterns when assumed to belong to the exact count class versus the rounding/selected response class. The discrepancy between the associated scale scores is due to the nonlinear ordering of response categories for the NRM. Consequently, within the rounding/selected response class, **U** = (0, 0, 0, 0) and **U** = (30, 30, 30, 30) are not representative of the most extreme levels of the latent variable. Rather, it is a response pattern of **U** = (0, 0, 0, 5) that is associated with the lowest scale score, $\hat{\theta}_{\mathrm{EAP}} = -0.93$, and a response pattern of **U** = (25, 25, 25, 25) that is associated with the highest scale score, $\hat{\theta}_{\mathrm{EAP}} = 2.87$.

TABLE 3.
Expected Scale Scores and Posterior Standard Deviations Associated With Different Response Patterns: Comparison of the Exact Count and Rounding/Selected Response Classes

Response Pattern **U**	Exact Count $\hat{\theta}_{\mathrm{EAP}}(SD)$	Selected Response $\hat{\theta}_{\mathrm{EAP}}(SD)$
(0, 0, 0, 0)	−1.40 (0.75)	−0.91 (0.72)
(5, 5, 5, 5)	0.52 (0.41)	0.57 (0.21)
(10, 10, 10, 10)	0.95 (0.38)	0.94 (0.17)
(15, 15, 15, 15)	1.19 (0.36)	1.33 (0.30)
(20, 20, 20, 20)	1.36 (0.35)	2.00 (0.39)
(25, 25, 25, 25)	1.50 (0.35)	2.87 (0.48)
(30, 30, 30, 30)	2.38 (0.52)	1.41 (0.32)

Note. *SD* = standard deviation.

A researcher planning to use these scale scores in secondary statistical analyses may be interested in the correlation between the two types of scores: Does it matter which type of score is computed for people who could belong to either the exact or rounding/selected response class? The correlation between the two sets of scores is 0.95, with a scatterplot shown in the lower right panel of Figure 7. The scatterplot reveals a funnel-shaped relationship, in which scores are not as highly correlated at the extreme positive end of the latent variable. Table 3 further highlights this trend. The scale score estimates that correspond to response patterns with higher counts, such as **U** = (25, 25, 25, 25) and **U** = (30, 30, 30, 30), show greater differences between the exact count and rounding/selected response scoring methods than scale scores representing response patterns with lower counts. According to the negative binomial IRT model, expected counts are a monotonically increasing function of the latent variable, whereas according to the NRM, expected counts do not always increase with the latent variable. A side effect of this nonmonotonic relationship is that for some items, counts of 30 are sometimes associated with better health than counts of 25. When this is true, scale scores that are computed from response patterns including 25 days are greater than scale scores based on response patterns including 30 days. While the two types of scale scores are highly correlated overall, the scores are essentially uncorrelated for individuals with poor emotional health that is more than 1 *SD* above average.

Posterior *SD*s are plotted in Figure 8. As tends to be true of IRT scores, the posterior *SD*s are larger at the extreme ends of the latent variable and smaller near values of the latent variable where the items are most discriminating. It is worth noting that the posterior *SD*s for the exact count class are almost never smaller than those for the rounding/selected response class; they are only lower when scale scores drop below average $(\hat{\theta}_{\text{EAP}} < 0)$. Because scale scores below the average belong to relatively healthy individuals, these scale scores are the product of response patterns with low counts, suggesting that when counts are low, poor emotional health is measured with greater precision in the exact count class than in the rounding/selected response class; for higher counts, the opposite is true.

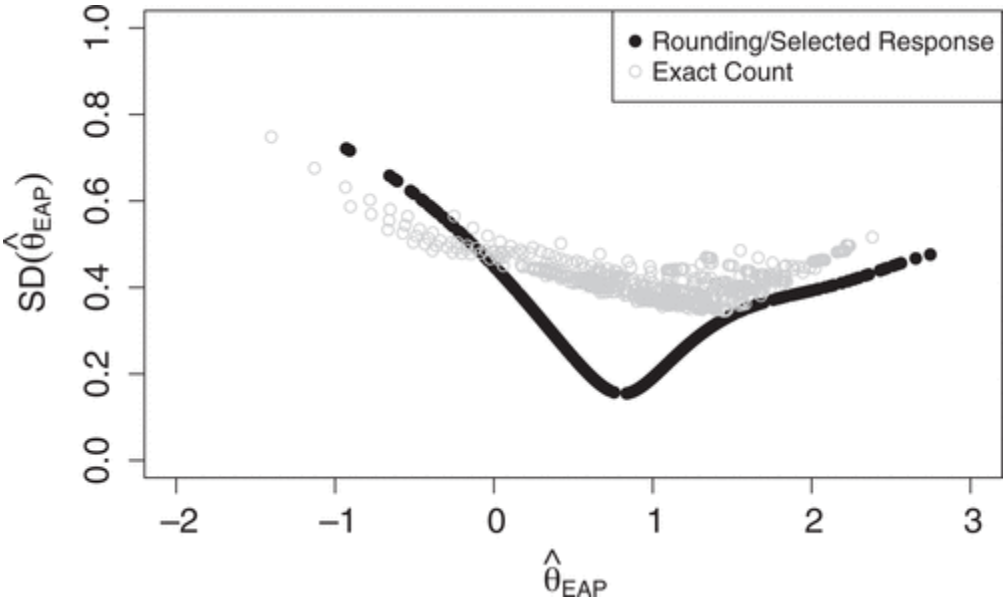

Figure 8. Posterior standard deviations (*SD*s) as a function of scale scores for the exact count and rounding/selected response classes. The *x*-axis displays the scale scores observed in the sample; the *y*-axis shows the posterior standard deviation associated with each of those scale scores.

Discussion

The goal of this research was to develop a latent class IRT model that could account for zero inflation, maximum inflation, and heaping in multivariate open-ended count item response data. Results of the empirical analysis suggest that a latent class IRT model that uses a negative binomial IRT model for the count process, a nominal response IRT model for heaping at preferred digits, and two degenerate IRT models for the zero and maximum classes may be a good approximation for the underlying response process that produces the observed count distributions. The results provide evidence that this modeling approach is useful in analyzing data from scales with multiple open-ended count items—in particular, the four count items from the BRFSS.

The results reveal a peculiarity in the way some people may respond to specific types of open-ended count items on questionnaires, and this peculiarity may have implications for scale development. When retrospectively responding to open-ended count items, a sizable proportion of individuals may not treat the item as an open-ended count but instead may treat it as a selected response item with a smaller number of response categories. While this phenomenon may not be characteristic of all types of count data, it is present within this subscale of the BRFSS. Further, within the rounding/selected response latent class, the seven response categories are not in an increasing order with respect to the latent variable. Specifically, the ordering of response categories suggests that people with the highest levels of poor emotional health are more likely to endorse 25 days than 30 days. Within the rounding/selected response class, someone who endorses 30 days may really mean some large quantity of days (i.e., more than 15). Choosing 30 days with this meaning does not require the respondent to engage in any type of count process. On the other hand, because selecting 25 days reflects the use of some type of count process and not just choosing the maximum response as a shortcut, someone who endorses 25 days likely really means a number around 25 days. This finding is counterintuitive to the inherent ordering of counts that one may expect in designing scales with open-ended count items, and it has implications for researchers who wish to draw conclusions about an individual's level of poor emotional health from a response pattern that include counts: Higher counts may not always indicate higher levels of the latent variable.

The results also suggest that it is important to account for population heterogeneity in item response data, not only in considering differences in response style but also in recognizing that the scale may not be measuring the same latent variable for all individuals. Specifically, the IRT analyses of the four count items on the BRFSS suggest that 16% of respondents belong to a zero class and 1% of respondents belong to a maximum class. These respondents may be at some floor or ceiling level of the latent variable or, in some sense, not at any value of the latent variable ϑ at all. For example, someone in the maximum class may fall at such severe levels of poor emotional health that this particular scale should not be used to assess that person—perhaps a different scale that provides more nuanced measurement at the extreme levels of the latent variable should be used instead. Further assessment of people who may belong to the zero or maximum classes is a logical next step. Another reason someone may be a member of the zero or maximum class is that the items may not be relevant to the respondent. Wall et al. (2015) and Finkelman et al. (2011) describe the unipolar nature of many clinical traits, such that a substantial proportion of the sample does not exhibit any of the symptoms or behaviors that are referenced in the items. In describing her zero-inflated Poisson IRT model, L. Wang (2010) explains a similar phenomenon that is commonly observed on questionnaires about substance use, in which many people report zero units of alcohol, cigarettes, and marijuana because they abstain from substance use and the items are not applicable. While the analog to emotional health symptoms

is not as intuitive, it is possible that some respondents are so low on psychopathology that they view the items as irrelevant. These individuals who belong to the zero class may be viewed similarly to substance use abstainers. Whatever the reason, ignoring the zero and maximum classes may lead to a biased representation of the latent variable in the population.

Recommendations

The model developed as part of this research is computationally complex. Not only does parameter estimation require several hours of computing time, but to our knowledge, user-friendly software that can implement these types of latent class IRT models is limited or perhaps nonexistent. At minimum, researchers need to directly specify the model log likelihood and use an optimizer such as R's nlm; more complicated models—for example, those designed for scales with a larger number of items—may require more sophisticated programming knowledge. Researchers can obviate such complex modeling techniques by not including items that elicit a retrospective count response on their scales and questionnaires. Alternative methods of framing the question can simplify item-level analyses.

Perhaps the simplest strategy to avoid eliciting open-ended count responses is to bin the response options before administering the questionnaire. Binning eliminates the issue of individual differences in response style, such that the exact count and rounding/selected response classes are no longer needed. It may reduce recall error. Framing the question with binned counts is less taxing on the respondent's memory than asking for a cumulative raw frequency. While binning may eliminate some of the issues associated with retrospectively reported open-ended count data, it can introduce other statistical modeling challenges (McGinley, Curran, & Hedeker, 2015). If one wishes to preserve raw frequencies, a more reliable approach may be to use a daily diary response format instead of retrospective counts. The daily counts could then be tallied to obtain a more accurate total frequency. One advantage of this approach is that because respondents are not retrospectively reporting a cumulative frequency, heaping is much less likely to be present in the observed item response data. Without heaping, a rounding/selected response class is no longer needed to account for digit preference, and a simple count IRT model may be sufficient. This reduction in parameters would greatly reduce estimation time as well as the complexities involved in scoring individuals when class membership is unknown. Because the results suggest that items eliciting raw count responses can substantially reduce the *SD*s of scale scores, the daily diary approach may be preferable to binning the counts, as the true count nature of the data is preserved and more information is available in each item response. Future research could investigate this topic.

Limitations

One of the major limitations of using open-ended count IRT models is the difficulty of assessing absolute model fit. As the number of count items increases, the number of possible response patterns becomes unmanageably large, producing multiway contingency tables with extreme sparseness. Such sparseness creates challenges in the development of goodness-of-fit statistics that compare observed and expected response pattern frequencies. Currently, examination of fit is limited to measures of relative model fit that can be computed from the model log likelihood such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC).

The negative binomial IRT model assumes that the observed counts are unbounded. For most of the count items, this IRT models serve as reasonable approximations for the empirical response distributions—very high counts are rarely observed in the data, and most of the 30s are manifestations of a rounding or selected response process, not a count process. However, an IRT model that uses a

bounded conditional count response distribution, such as a beta-binomial distribution, is likely more appropriate for 30-day recall items in which observations toward the upper limit are frequent. Future extensions of this work could include beta-binomial IRT models for the exact count class to accommodate bounded count responses.

Conclusions

The goal of this research was to develop an IRT model that could address some of the challenges that commonly arise in analyzing multivariate count data from questionnaires. The proposed model integrates elements from three different methodological approaches rooted in psychometrics and biostatistics: IRT models for zero-inflated count data, latent variable models for heaping and response style, and latent class IRT. While not without limitations, the latent class IRT model is able to address many of the issues involved in analyzing the multivariate open-ended count items that are becoming more common in clinical assessment, and we believe that they show promise of wider applicability in the field of psychological measurement.

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

Agresti, A. (2002). *Categorical data analysis*. New York, NY: Wiley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika*, 37, 29–51. doi:10.1007/BF02291411

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, 17, 665–678. doi:10.1037/a0028111

Böckenholt, U., Kamakura, W. A., Wedel, M. (2003). The structure of self-reported emotional experiences: A mixed-effects Poisson factor model. *British Journal of Mathematical and Statistical Psychology*, 56, 215–229. doi:10.1348/000711003770480011

Bolt, D. M., Johnson, T. R. (2009). Applications of MIRT model to self-report measures: Addressing score bias and DIF due to individual differences in response style. *Applied Psychological Measurement*, 33, 335–352. doi:10.1177/0146621608329891

Bolt, D. M., Newton, J. R. (2011). Multiscale measurement of extreme response style. *Applied Psychological Measurement*, 71, 814–833. doi:10.1177/0013164410388411

Cameron, C., Trivedi, P. K. (2013). Regression analysis of count data, *Econometric Society Monograph* No. 53 (2nd ed.). Cambridge, England: Cambridge University Press.

Centers for Disease Control and Prevention. (1984–present). *Behavioral risk factor surveillance system survey data*. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

Deb, P., Trivedi, P. K. (1997). Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics*, 12, 313–336. doi:10.1002/(SICI)1099-1255(199705)12:33.0.CO;2-G

De Boeck, P., Partchev, I. (2012). IRTress: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, 48, 1–28. doi:10.18637/jss.v048.c01

Finch, W. H., Pierson, E. E. (2011). A mixture IRT analysis of risky youth behavior. *Frontiers in Psychology*, 2, 1–10. doi:10.3389/fpsyg.2011.00098

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. doi:10.1002/sim.4217

Hagenaars, J., McCuthcheon, A. (2002). *Applied latent class analysis*. Cambridge, England: Cambridge University Press.

Heitjan, D. F., Rubin, D. B. (1990). Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association*, 85, 304–314. doi:10.1080/01621459.1990.10476202

Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed.). Cambridge, England: Cambridge University Press.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to the defects in manufacturing. *Technometrics*, 34, 1–14. doi:10.1080/00401706.1992.10485228

Lewis, M. A., Neighbors, C., Geisner, I. M., Lee, C. M., Kilmer, J. R., Atkins, D. C. (2010). Examining the associations among severity of injunctive drinking norms, alcohol consumption, and alcohol-related negative consequences: The moderating roles of alcohol consumption and identity. *Journal of Addictive Behaviors*, 24, 177–189. doi:10.1037/a0018302

Loeys, T., Moerkerke, B., De Smet, O., Buysse, A. (2012). The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. *British Journal of Mathematical and Statistical Psychology*, 65, 163–180. doi:10.1111/j.2044-8317.2011.02031.x

McCullagh, P., Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Boca Raton, FL: Chapman & Hall.

McGinley, J. S., Curran, P. J., Hedeker, D. (2015). A novel modeling framework for ordinal data defined by collapsed counts. *Statistics in Medicine*, 34, 2312–2324. doi:10.1002/sim.6495

Muthen, B., Asparouhov, T. (2006). Item response mixture modeling: Application to tobacco dependence criteria. *Addictive Behaviors*, 31, 1050–1066. doi:10.1016/j.addbeh.2006.03.026

Reise, S. P., Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, 5, 27–48. doi:10.1146/annurev.clinpsy.032408.153553

Ridout, M. S., Morgan, B. J. T. (1991). Modelling digit preference in fecundability studies. *Biometrics*, 47, 1423–1433. doi:10.2307/2532396

Roberts, J. M., Brewer, D. D. (2010). Measures and tests of heaping in discrete quantitative distributions. *Journal of Applied Statistics*, 28, 887–896. doi:10.1080/02664760120074960

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14, 271–282. doi:10.1177/014662169001400305

Rost, J. (1997). Logistic mixture models. In van der Linden, W. J., Hambleton, R. K. (Eds.), *Handbook of modern item response theory* (pp. 449–463). New York, NY: Springer.

Sawatzky, R., Ratner, P. A., Kopec, J. A., Zumbo, B. D. (2012). Latent variable mixture models: A promising approach for the validation of patient reported outcomes. *Quality of Life Research*, 21, 637–650. doi:10.1007/s11136-011-9976-6

Thissen, D., Cai, L., Bock, R. D. (2010). The nominal categories item response model. In Nering, M. L., Ostini, R. (Eds.), *Handbook of polytomous item response theory* models (pp. 43–76). New York, NY: Routledge.

Thissen, D., Wainer, H. (2001). *Test scoring*. New York, NY: Routledge.

Thissen-Roe, A., Thissen, D. (2013). A two-decision model for responses to Likert-type items. *Journal of Educational and Behavioral Statistics*, 38, 522–547. doi:10.3102/1076998613481500

von Davier, M., Rost, J. (2006). Mixture distribution item response models. In Rao, C. R., Sinharay, S. (Eds.), *Handbook of statistics*, Vol. 26: Psychometrics (pp. 643–661). Amsterdam, the Netherlands: North Holland.

Wall, M. M., Park, J. Y., Moustaki, I. (2015). IRT modeling in the presence of zero-inflation with application to psychiatric disorder severity. *Applied Psychological Measurement*, 39, 583–597. doi:10.1177/0146621615588184

Wang, H., Heitjan, D. F. (2008). Modeling heaping in self-reported cigarette counts. *Statistics in Medicine*, 27, 3789–3804. doi:10.1002/sim.3281

Wang, L. (2010). IRT-ZIP modeling for multivariate zero-inflated count data. *Journal of Educational and Behavioral Statistics*, 35, 671–692. doi:10.3102/1076998610375838

Wedel, M., Böckenholt, U., Kamakura, W. A. (2003). Factor models for multivariate count data. *Journal of Multivariate Analysis,* 87, 356–369. doi:10.1016/S0047-259X(03)00020-4

Wright, D. E., Bray, I. (2003). A mixture model for rounded data. *The Statistician*, 52, 3–13. doi:10.1111/1467-9884.00338