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

Brooke Erin Magnus

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology (Quantitative).

Chapel Hill
2016

Approved by:

David Thissen

Laura Castro-Schilo

Patrick J. Curran

John S. Preisser

Eric A. Youngstrom

Abstract

Brooke Erin Magnus: Item Response Modeling of Multivariate Count Data with Zero-Inflation, Maximum Inflation, and Heaping
(Under the direction of David Thissen)

Questionnaires that include items eliciting count responses are becoming increasingly common in psychology and health research. Item response data from these types of questionnaires pose analytic challenges, including inflation at zero and the maximum, as well as heaping at preferred digits; such data complexities are not well-suited for conventional IRT modeling approaches and software. This research proposes methodological techniques to overcome those challenges by combining approaches from three related but distinct literatures: IRT models for multivariate count data, latent variable models for heaping and extreme responding, and mixture IRT models. Scales from the Behavioral Risk Factor Surveillance System are used as motivating examples in addressing three questions. First, what are some methods of addressing inflation and heaping in multivariate count item response data? Second, are complex models really needed, or can heaping and inflation in count data be ignored? And finally, what value do count items add to scales?

The results suggest that count item response data can be modeled within a latent class IRT framework. The proposed latent class IRT model has a Poisson or negative binomial component for a class of individuals who respond to items according to a strict count process, a nominal response component for a class of individuals who respond to items according to a multiple choice or rounding process, and two degenerate models to describe some of the individuals who always endorse the minimum or maximum counts. A comparison of the full model with more parsimonious models reveals that all four latent classes are needed to describe the empirical item response distributions. Methods of computing scale scores are described. The results also provide evidence that including count items on scales may improve measurement precision, but the degree of improvement is dependent on latent class membership. Count items are likely to

be most informative when respondents engage in a true count process. The results also support the idea that if count items are to be used on scales, it is advisable to include more than one. Practical implications are discussed and recommendations are provided for researchers who may wish to use count items on questionnaires.

Acknowledgements

First and foremost, I thank my advisor, Dave Thissen. Thank you for your continued guidance and patience throughout my time at UNC. You've challenged me to think harder than I realized I was capable of, and I have truly grown as a researcher because of your mentorship. I am fortunate to have had the opportunity to work with you. I also sincerely thank the members of my dissertation committee: Laura Castro-Schilo, Patrick Curran, John Preisser, and Eric Youngstrom. Laura, thank you for being such a wonderful mentor and role model. I am consistently in awe of all that you do, and I look up to you in so many ways. Patrick, you've provided me with invaluable feedback on projects over the years, and I will not forget the times you've come to bat for me. John, thank you for venturing into my world of psychometrics and providing your biostatistical expertise on my dissertation. Your categorical course is one of the reasons I became interested in this topic. And Eric, I have enjoyed having your clinical perspective to remind me of the broader implications of measurement research. I'd also like to express my gratitude to the many other people who have served as my mentors during graduate school, including Bryce Reeve, Abigail Panter, and the faculty of the Quantitative Psychology program.

I owe countless thanks to Yang Liu, former officemate, continued friend and colleague. My graduate school experience was tremendously enhanced by your presence, and I know I will continue to learn from you throughout my career. Thank you also to the former and current students of the L. L. Thurstone Psychometric Lab – Corinne, Cara, Veronica, Michael, Nathan, Zack, Stephanie, Sierra, Jason, Noah, and Teague – for your never-ending ideas and encouragement, especially this past year. I will sorely miss our Thursday afternoons at Tru.

Outside of Davie Hall, I am forever indebted to my core support system in North Carolina. Ellie, Rachel, Abby, and Jordy: Our weekly dinners and daily laughs have kept me going for the last few years. You are some of the kindest and most insightful people I have ever had the pleasure of knowing, and I am so fortunate to have you in my life. Carrie, Laura, Patty, Elise,

Lahnna, Matt, Susan, and Kelly: I can always count on you to brighten my day, and I don't think I would have made it through graduate school, my dissertation, and the perils of the job search without your unwavering support. And finally, I thank my family for standing behind me in everything that I do. I would not have my PhD if not for you.

Table of Contents

List of Tables

List of Figures

Chapter 1: Background and Motivation

Count data are prevalent in the social sciences. In psychology, for example, a researcher may be interested in predicting the number of occurrences of a specific behavior based on a set of covariates, such as using scores on an attachment scale to predict the number of perpetrations of unwanted pursuit behavior in broken up couples (e.g., Loeys, Moerkerke, De Smet, & Buysse, 2012). 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 (number of drinks consumed per week, Lewis, Neighbors, Geisner, Lee, Kilmer, & Atkins, 2010), medicine (number of tumors at time of death, Dunson & Herring, 2005; number of sexual partners, Roberts & Brewer, 2008) and economics (number of hospital stays, Deb & Trivedi, 1997), among others. Loeys et al. (2012) recently published a review of some of the current challenges and proposed solutions to modeling univariate count outcomes in psychological research.

Psychological questionnaires that comprise multiple items eliciting count responses are becoming increasingly common, particularly in the domain of public health. Often, these surveys are designed to assess the severity of symptoms and ask respondents to recall the frequency of various thoughts or behaviors over a pre-specified period of time. For example, a survey that is intended to measure depression may include an item asking the respondent to estimate the number of days he or she has felt sad in the past month; a survey measuring alcohol dependence may ask the respondent to report the number of drinks he or she consumes during a typical week. As an indication of the prevalence of these types of items in survey research, Table 1 lists some examples of published questionnaires including at least one count item.

While statistical methods for the analysis of a single count outcome are widely available (e.g., Agresti, 2002; Cameron & Trivedi, 2013; McCullagh & Nelder, 1989), methods for modeling

Questionnaire	Source/Authors
Behavioral Risk Factor Surveillance System	Centers for Disease Control (1984-present)
National Health and Nutrution Examination Survey	Centers for Disease Control (1971-present)
Youth Risk Behavior Survey	Centers for Disease Control (1991-present)
WHO Disability Assessment Schedule 2.0	World Health Organization (1998-present)
HIV Risk-Taking Behaviour Scale	Ward, Darke, & Hall (1990)
Cognitive Appraisal of Risky Events Scale	Fromme, Katz, & Rivet (1997); Katz, Fromme, & D'Amico (2000)
Inventory of Statements about Self-Injury	Klonsky & Glenn (2008)
Self-Injurious Thoughts and Behaviors Interview	Nock, Holmberg, Photos, & Michel (2007)

Table 1. Questionnaires with at least one item eliciting a count response

multivariate count outcomes, such as responses to sets of count items on questionnaires, are considerably less well-developed. A question that arises from the content of these questionnaires is how one can derive meaningful scores from count responses that reflect the construct that is of interest (e.g., depression or alcohol dependence). Item response theory (IRT), rooted in educational measurement (Thissen & Wainer, 2001), has played an increasing role in the assessment of psychiatric and health outcomes (e.g., Finch & Pierson, 2011; Finkelman, Green, Gruber, & Zaslavsky, 2011; Sawatzky, Ratner, Kopec, & Zumbo, 2012; Wall, Park, & Moustaki, 2015). Educational applications of IRT have direct analogs in psychiatric assessment (Reise & Waller, 2009) – just as a latent ability level is thought to underlie a person's answers to items on an educational test, a latent variable is also believed to influence someone's responses to items on a questionnaire assessing health status. For example, someone with a high level of depression is likely to endorse the more severe response categories on items comprising a depressive symptoms scale, just as someone with high proficiency is expected to select correct responses on an educational test.

The IRT literature is heavily focused on the analysis of binary, ordinal, and nominal item types, likely due to IRT having its origins in educational assessment – one is unlikely to find a count item on a math or reading test. A reasonable approach to modeling multivariate count responses might be to modify traditional IRT techniques, invoking a log link function in place of the usual logit or probit link and a Poisson distribution in place of a Bernoulli or multinomial conditional response distribution. However, if one examines most count data more closely, a number of additional challenges surface that require a more complex methodological approach.

To illustrate the analytic issues, Figure 1 shows histograms of 5,000 randomly selected responses to four items about general health found on the Behavioral Risk Factor Surveillance System (BRFSS; CDC, 1984-present). Each item asks 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 zero 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. Inflation at zero and the maximum may reflect unique subsets of people with either a complete absence or such a severe presence of depression that these respondents do not come from the same populations as the rest of the sample. Further, there is noticeable inflation at days that are multiples of five. For example, it is more common for people to report feeling depressed for five days than four or six days, even though these adjacent values are in theory no less plausible. This type of inflation at preferred digits is commonly referred to as heaping or data coarsening in the biostatistics literature (Heitjan & Rubin, 1990; H. Wang & Heitjan, 2008; Wright & Bray, 2003). Simply modifying a traditional IRT model to invoke a log link function and a Poisson conditional response distribution is not likely to account for the potential subpopulations and individual differences that result in the histograms observed in Figure 1. This research attempts to address this problem by combining methodological approaches from three related but distinct literatures: IRT models for multivariate count data, latent variable models for heaping and extreme responding, and mixture IRT models.

1. Item Response Models for Multivariate Count Data

Over the past decade, the literature on psychometric models for multivariate count data has grown (Bockenholt, Kamakura, & Wedel, 2003; L. Wang, 2010; Wedel, Bockenholt, & Kamakura, 2003); however, it remains quite sparse, especially in comparison with the advancement seen in other areas of IRT. Most recently, L. Wang (2010) developed an item response model for zero-inflated Poisson data (IRT-ZIP) using a non-linear mixed modeling framework. Based on Lambert's (1992) original zero-inflated Poisson regression model, Wang's model is a latent variable mixture model that accounts for two different response processes: the zero process, which relates to whether the event has a chance of occurring at all, and the Poisson process, which

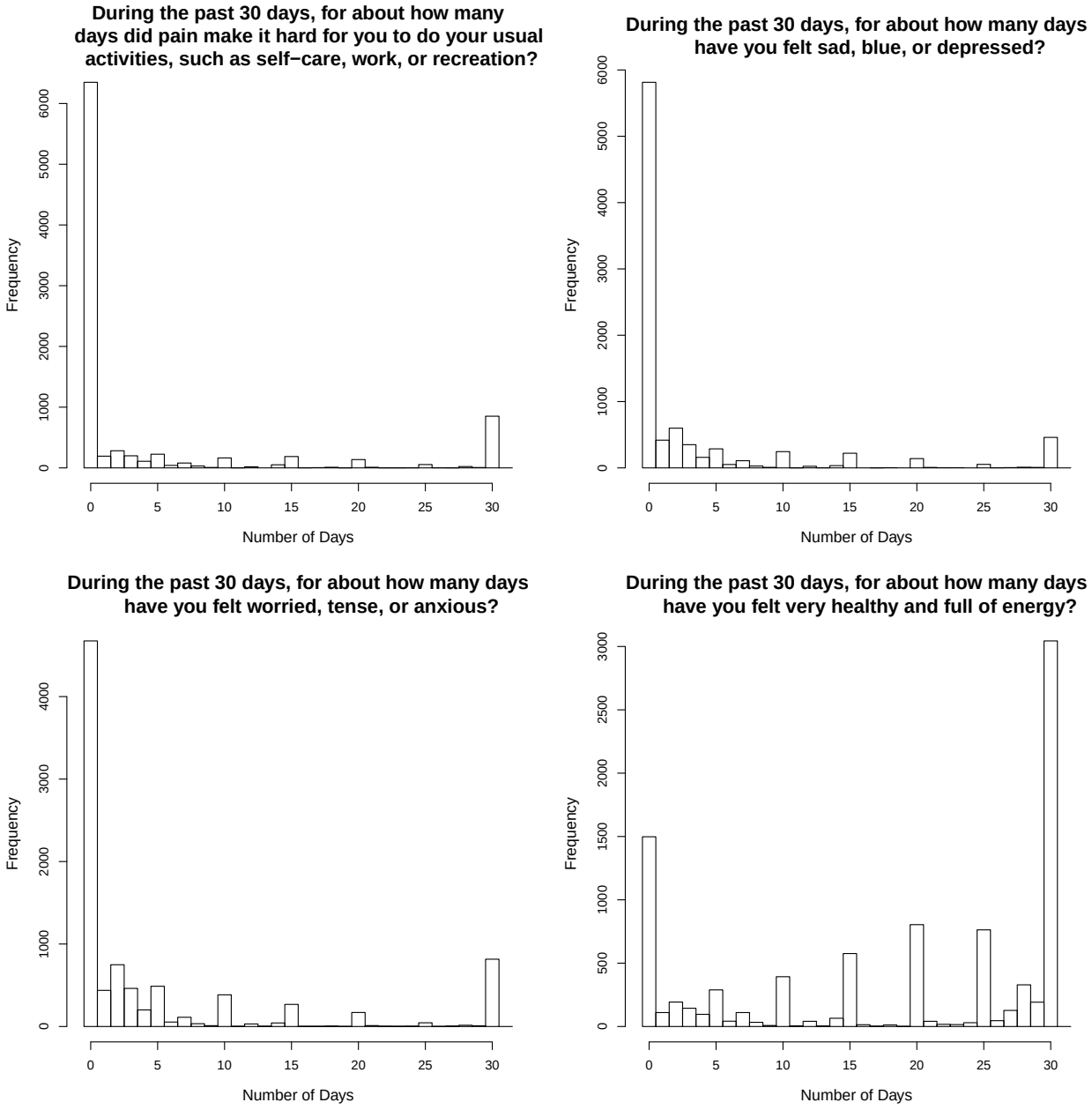


Figure 1. Frequency histograms for four general emotional health items from the BRFSS (2014) eliciting count responses.

relates to the expected event count, given that the event has a chance of occurring. Consider an item that asks respondents to report the number of alcoholic beverages they have consumed in the last week. There are at least two reasons someone might register a response of zero. One reason is that the person may simply not drink alcoholic beverages, being an abstainer, and would thus respond with zero to any question related to the frequency of alcohol consumption. On the other hand, someone may report zero not because that person is an abstainer, but because he or she, either by chance or some other reason, has not consumed any alcoholic beverages in the last week. While both underlying processes result in the same observed response, these two types of zeros are qualitatively different. Zero-inflated Poisson models, including Wang's IRT-ZIP model, can be used to distinguish probabilistically the abstainers from the people who normally drink but may not have consumed any alcoholic beverages in the specified time frame.

According to Wang's IRT-ZIP model, the observed count response U_{ij} for person i on item j is expressed

$$U_{ij} \sim \begin{cases} 0, & \text{with probability } 1 - p_{ij} \\ \text{Poisson}(\lambda_{ij}), & \text{with probability } p_{ij} \end{cases} \tag{1}$$

in which

$$P(U_{ij} = 0) = (1 - p_{ij}) + p_{ij}e^{-\lambda_{ij}} \tag{2}$$

$$P(U_{ij} = u_{ij}) = \frac{p_{ij}e^{-\lambda_{ij}}\lambda_{ij}^{u_{ij}}}{u_{ij}!},\ u_{ij} = 1, 2, ... \tag{3}$$

The probability p_{ij} of person i being in the Poisson process on item j, and λ_{ij}, the expected count for person i on item j given that person i is in the Poisson process, can then be modeled as a function of item parameters $(a_{1j}, a_{2j}, b_{1j}, b_{2j})$ and a latent variable (θ_i):

$$\log(\lambda_{ij}) = a_{1j}(\theta_i - b_{1j}) \tag{4}$$

$$\text{logit}(p_{ij}) = a_{2j}(\theta_i - b_{2j}). \tag{5}$$

For model identification, θ_i is assumed to follow a standard normal distribution. There are two sets of item parameters; one set of parameters is for the Poisson process (Equation 4), while the other corresponds to the zero state (Equation 5). The parameters a_{1j} and a_{2j} are item discriminations for the Poisson process and zero state, respectively; the larger these values, the more discriminating the item is in estimating scores on the latent variable. The parameters b_{1j} and b_{2j} are the location parameters for the Poisson process and zero state, respectively. The larger b_{1j}, the more difficult it is for someone in the Poisson process (i.e, someone who is not an abstainer) with a given latent variable score to reach a high expected count; the larger b_{2j}, the more likely it is for someone with a given latent variable score to be classified as part of the zero-state (abstainers) and not enter the Poisson process. By varying the values of these two sets of parameters, one observes different proportions of zero-inflation and ranges of expected counts. Wang parameterized the IRT-ZIP model as a generalized multilevel model that can be estimated using marginal maximum likelihood. The marginal log likelihood for item parameter sets $\boldsymbol{a}$ and $\boldsymbol{b}$ given observed responses $\boldsymbol{u}$ is written

$$l(\boldsymbol{a}, \boldsymbol{b}|\boldsymbol{u}) = \sum_{i=1}^{N} \log \int \prod_{j=1}^{J} p(u_{ij}|\boldsymbol{a}, \boldsymbol{b}, \theta_i) N(\theta_i) d\theta_i, \tag{6}$$

in which $p(u_{ij}|\boldsymbol{a}, \boldsymbol{b}, \theta_i)$ can be re-expressed based on the conditional probabilities from Equations 2 and 3:

$$p(u_{ij}|\boldsymbol{a}, \boldsymbol{b}, \theta_i) = [(1 - p_{ij}) + p_{ij}e^{-\lambda_{ij}}]^{1-y_{ij}} \left[\frac{p_{ij}e^{-\lambda_{ij}}\lambda_{ij}^{u_{ij}}}{u_{ij}!}\right]^{y_{ij}} \tag{7}$$

where $y_{ij} = 1$ if $u_{ij} \neq 0$ and $y_{ij} = 0$ otherwise.

Wang implemented marginal maximum likelihood in SAS PROC NLMIXED, using adaptive quadrature to approximate the integral in Equation 6. Treating the estimated item parameters as fixed, she then computed latent variable scores using empirical Bayes methods. Wang applied the IRT-ZIP model to data from the National Longitudinal Survey of Youth, combining three items to form a substance use scale and examining trends in substance use over time.

While Wang's IRT-ZIP model provides an item response modeling framework for analyzing the psychometric properties of zero-inflated multivariate count data, it is somewhat restrictive.

Mainly, it assumes that the non-perfect zero state is a Poisson process; however, 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 prevalent (e.g., H. Wang & Heitjan, 2008), as illustrated in Figure 1. For this reason, a more flexible modeling approach may be useful.

2. Heaping and Response Style

H. Wang and Heitjan (2008) examined self-reported counts of cigarette use as the outcome of a clinical trial on smoking cessation. They were interested in evaluating the effect of an antidepressant on smoking abstinence in a sample of smokers intending to quit. At a follow-up assessment after being instructed to quit smoking, participants were asked to retrospectively report the number of cigarettes smoked each day. Frequency distributions of cigarette counts revealed a very large proportion of zeros, potentially representing a subset of respondents who had successfully quit smoking, as well as heaped responses at multiples of five. In particular, the authors noted the unsurprising heaping at 20, corresponding to the number of cigarettes sold in a package.

To account for the large proportion of respondents reporting zero cigarettes, as well as the non-trivial heaping at 5, 10, and 20 cigarettes, H. Wang and Heitjan (2008) introduced a model in which the observed cigarette count was a function of both the unobserved true cigarette count and a latent "heaping behavior" variable. Their discrete latent heaping behavior variable could take on one of four values, depending on the type of rounding behavior: exact count, multiple of 5, multiple of 10, or multiple of 20. They also modeled the potential relationship between the heaping behavior and the underlying count variable according to a proportional odds model, hypothesizing that coarser rounding may be associated with larger true cigarette counts. They used Bayesian methods to estimate model parameters, fitting a series of zero-inflated Poisson and negative binomial models that either ignored or accounted for heaping, and found strong evidence for improved model fit after accounting for heaping. They also found that heaping had a sizable effect on the estimated quit probability and mean cigarette count, with over 40% of the sample exhibiting some type of rounding behavior.

From a psychometric perspective, Wang and Heitjan's model could be considered a multi-dimensional latent variable approach. Not only does the original latent variable of interest –

cigarette addiction – influence the observed cigarette count, but there is an additional individual differences variable influencing people's rounding behavior: Someone at high levels of the latent variable that reflects rounding is more likely to exhibit coarse rounding behavior. In their model, Wang and Heitjan treated rounding behavior as discrete with ordered categories, similar to a latent class variable.

The idea that a self-reported cigarette count is influenced by an underlying variable is not unlike psychometric models, which assume that a latent variable underlies observed responses to questionnaire items; however, Wang and Heitjan's model dealt only with a single count outcome, not the multiple items or measures that are routinely used in psychometric modeling. Other statisticians have also developed models for heaping in univariate outcomes. For example, Heitjan and Rubin (1990) used multiple imputation to model rounding in self-reported age, treating the estimation of true age as a missing data problem. Ridout and Morgan (1991) introduced a model to account for heaping in women's retrospective reports of the number of menstrual cycles before a positive pregnancy, with digit preference often occurring at 6, 12, and 3 cycles. Wright and Bray (2003) used Bayesian mixture modeling techniques to capture the rounding process in clinician-reported measurements from ultrasound images, in which different components of the mixture model represented different levels of rounding. Review of the biostatistics literature, however, 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 addresses a similar concept within an IRT framework. Bolt and Johnson (2009) used a multidimensional nominal response IRT model to account for the individual differences that increase the probability that some respondents select the "strongly disagree" or "strongly agree" options on a rating scale; this tendency is referred to as extreme response style (ERS). Research suggests that ERS can interfere with the response process purportedly being modeled using IRT, and if not accounted for, can result in less precise estimates of the latent variable of interest, biased item parameter estimates, and spurious correlations of latent variable estimates with other variables (Bolt & Newton, 2011; Jin & Wang, 2014; Thissen-Roe & Thissen, 2013). Similar consequences may hold if heaping in count item response data is ignored.

Bolt and Johnson (2009) modeled item responses as a function of both the substantive construct of interest and an ERS latent variable. According to their model, the probability of selecting response category k on item j is expressed

$$P(U_j = k|\theta_1, \theta_2, ..., \theta_d) = \frac{\exp(a_{jk1}\theta_1 + a_{jk2}\theta_2 + ... + a_{jkd}\theta_d + c_{jk})}{\sum_{h=1}^{K} \exp(a_{jk1}\theta_1 + a_{jk2}\theta_2 + ... + a_{jkd}\theta_d + c_{jk})}, \tag{8}$$

where θ_1, θ_2, ..., θ_d are the d latent variables assumed to underlie someone's category selection, and a and c are the slope and intercept parameters for category k on item j. One of the latent variables may reflect ERS. For example, Bolt and Newton (2011) described a multidimensional nominal response model in which there is a latent variable related to the construct of interest, θ_1, and a latent variable for ERS, θ_{ERS},

$$P(U_j = k|\theta_1, \theta_{\text{ERS}}) = \frac{\exp(a_{jk1}\theta_1 + a_{jk2}\theta_{\text{ERS}} + c_{jk})}{\sum_{h=1}^{K} \exp(a_{jk1}\theta_1 + a_{jk2}\theta_{\text{ERS}} + c_{jk})}. \tag{9}$$

θ_{ERS} is identified by a pattern of large a_{jk2} parameters for extreme responses, and smaller a_{jk2} parameters for less extreme responses. These models allow estimation of IRT scores that reflect each latent variable, with scores accounting for the simultaneous influence of the substantive construct and response style on the observed responses.

Other researchers have conceptualized item responses as observed outcomes resulting from a sequence of internal decisions (Böckenholt, 2012; De Boeck & Partchev, 2012; Thissen-Roe & Thissen, 2013). Böckenholt (2012) and De Boeck and Partchev (2012) proposed a tree structure for capturing the ways in which individuals may differ in answering test items. They argued that while IRT models tend to assume a single response process, it is possible that multiple response processes are at play when someone responds to a questionnaire item. Böckenholt (2012) provided an example of a Likert-type item with five categories: *strongly disagree, disagree, neither disagree or agree, agree*, and *strongly agree.* At Process I, the respondent decides whether he or she expresses an opinion. If not, the person selects *neither disagree or agree* and the tree process ends; responses for the remaining processes are coded as missing. If the respondent chooses to express an opinion, the tree branches to Process II where the respondent decides the direction of the opinion: agreement vs. disagreement. Either choice results in branching to Process III, where the intensity of that opinion is reported – the respondent selects *strongly agree*

or *agree* if the opinion is positive, *strongly disagree* or *disagree* if the opinion is negative. The probability of a particular response category can be expressed as the product of the respective branch probabilities. Böckenholt's (2012) model does not assume that the same latent variable underlies each process. As noted in Thissen-Roe and Thissen's (2013) review of the literature, however, Böckenholt's (2012) and De Boeck and Partchev's (2012) tree structure models are members of the generalized linear mixed model (GLMM) family and cannot handle bilinear functions; therefore, item discrimination parameters are not included in any of the response process models. An additional limitation of Böckenholt's (2012) model is that it does not have the feature that more than one branching path can lead to the same observed response.

Using an argument similar to Böckenholt's (2012) and De Boeck and Partchev's (2012), Thissen-Roe and Thissen (2013) developed a two-decision model for responses to Likert-type items. They posited that in responding to each Likert-type item, examinees answer two internal items. The response to the first pseudo-item is the realization of a binary process: Does the respondent agree or disagree? The response to the second pseudo-item describes the strength of the first response: Given that the respondent agrees (or disagrees), how strong is that agreement (or disagreement)? Like Böckenholt's (2012) model, the probability of observing a response category is expressed as the product of the probabilities of the relevant outcome occurring at each stage. At the first stage, the probability is a function of the latent variable(s) the items are designed to measure; at the second stage, the probability is a function of the intended latent variable as well as a different latent variable that reflects the secondary construct of extreme response behavior. Unlike Böckenholt's (2012) and De Boeck and Partchev's (2012) models, Thissen-Roe and Thissen's (2013) model is not a member of the GLMM family and thus can accommodate an item discrimination parameter. The central idea behind these tree structure models is that response intensity is modeled by a separate response process and with a different latent variable than response direction.

Biostatistical models for heaping and psychometric models for extreme responding developed in different fields and from different methodological frameworks; however, both approaches converge on the idea of a latent variable underlying individual differences in the response process. The tree structure psychometric models posit that one latent variable underlies the first decision about presence or absence of an opinion, and a separate but possibly correlated latent variable

underlies the second decision about intensity of opinion. A similar approach could be adopted in modeling zero-inflated count data with heaping. One latent variable may underlie the presence or absence of the chance of a countable behavior; given that the respondent may potentially exhibit a non-zero frequency, a second latent variable could reflect the intensity of the frequency and a third latent variable could represent individual differences in response style (RS) – or in the case of count data, rounding behavior. Some extreme response models treat RS as a continuous latent variable (e.g., Böckenholt, 2012; Bolt & Johnson, 2009; Bolt & Newton, 2011; De Boeck & Partchev, 2012; Thissen-Roe & Thissen, 2013); accordingly, differences in scores on the latent construct represent quantitative differences in response styles. Other models assume that RS is a discrete latent variable, in which categorical latent classes correspond to qualitatively different types of response styles (e.g., Maij-de Meij, Kelderman, & van der Flier, 2008; Moors, 2008; Rost, Cartensen, & Von Davier, 1997). The latter is more similar to the approach taken by biostatisticians including H. Wang and Heitjan (2008), who group people into rounding classes depending on their response style. In addition to accounting for distinct response styles, latent class IRT, commonly referred to as mixture IRT, has potential application to the analysis of multivariate count data.

3. Mixture Item Response Theory

Unlike traditional item response models, mixture item response models assume that the observed responses are sampled from a population that has a number of subgroups or subpopulations (Rost, 1990, 1997; von Davier & Rost, 2006). Item parameters or even the parametric form of the item response model may vary across these subgroups. Under the assumption of local independence, the marginal mixture distribution of the observed item responses $u = (u_1, ..., u_J)$ is expressed

$$p(u_1, ..., u_J) = \sum_{g=1}^{G} \pi(g) \left(\int_\theta \prod_j p_{gj}(u_j|\theta)\phi(\theta|g)d\theta \right) \tag{10}$$

where $\int_\theta \prod_j p_{gi}(u_j|\theta)\phi(\theta|g)d\theta$ is the conditional probability of response pattern $(u_1, ..., u_J)$ in subpopulation g, written $p(u_1, ..., u_J|g)$. Observed responses and latent variable densities are conditional on the subpopulation, with $\pi(g)$ denoting the proportion of the population belonging to subpopulation g. Both the latent variable(s) θ and the class membership g are treated as

unobserved variables and are estimated as part of the model.

von Davier and Rost (2006) reviewed applications of mixture item response modeling to educational measurement; for example, student strategy usage on an educational test can be treated as an unobserved latent class variable. Depending on the strategy used, items may vary in their difficulty parameters, with different test taking strategies making certain items easier to answer correctly (Bolt, Cohen, & Wollack, 2001; Mislevy & Verhelst, 1990; Rost, 1990). Other researchers have used mixture item response modeling to account for test speededness (Bolt, Cohen, & Wollack, 2002; Yamamoto & Everson, 1997). Such models often include a "speeded" class comprising individuals who had insufficient time to answer items at the end of a test, and a "nonspeeded" class comprising the individuals who had enough time to answer all items; potential qualitative differences in response processes are accounted for in mixture item response models. The same rationale can be applied to mixture item response modeling in health outcomes research (Finch & Pierson, 2011; Finkelman et al., 2011; Muthen & Asparouhov, 2006; Sawatzky et al., 2012). Respondents belonging to different 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 in the same way as other subgroups in the population.

Finkelman et al. (2011) proposed a mixture IRT model for binary zero- and K-inflated health questionnaire data. They addressed the measurement of psychiatric disorders with low prevalence, arguing that the normal prior commonly used as the population distribution in IRT is unrealistic when people with high levels of the latent variable are rare. When many of the respondents in the population possess none or very low levels of the construct being measured, such as a large group of people 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. In the psychometrics literature, these types of clinical constructs are sometimes referred to as "unipolar" because it is possible for a respondent to exhibit a complete absence of the latent variable (Reise & Waller, 2009; Wall et al., 2015). Sometimes in clinical assessment, there may also be a small subset of respondents who are extreme at the other end of the latent variable, endorsing all possible symptoms on a checklist. Finkelman et al. (2011) referred to the high frequency of respondents with the maximum observed score as K-inflation. To overcome

the challenges associated with measuring low-prevalence psychiatric disorders, Finkelman et al. (2011) used a latent class item response model to account for extreme subpopulations. One latent class describes the people with no symptoms (the no-symptom group); a second class describes the people exhibiting most or all of the symptoms (the all-symptom group). The remaining latent class, labeled the graded class, describes people along the severity continuum implied by a traditional item response model with a normal prior. It is the presence of respondents from potentially different populations that requires mixture IRT in place of a conventional IRT model.

For a general latent class IRT model, the probability of person i endorsing item j is given by

$$P_j = \sum_{g=1}^{G} \pi_g P_{gj}. \tag{11}$$

In Equation 11, class membership is indexed $g = 1, ..., G$ and P_{gj} is the conditional probability of someone in class g endorsing item j. Finkelman et al. (2011) used indicator variables I_1, I_2, and I_3 to denote membership in the no-symptom class, the graded class, and the all-symptom class, respectively; π_1, π_2, and $\pi_3 = 1 - \pi_1 - \pi_2$ are the proportions of people in the population belonging to the respective classes. The probability of endorsing item j is 0 if $I_1 = 1$ and 1 if $I_3 = 1$. If $I_2 = 1$, the probability of endorsing item j is given by the 2-parameter logistic (2PL) IRT model

$$P_j(U_j = 1|\theta_i) = \frac{\exp\{a_j(\theta_i - b_j)\}}{1 + \exp\{a_j(\theta_i - b_j)\}} \tag{12}$$

in which a_j is the discrimination parameters for item j, b_j is the difficulty parameter for item j, and θ_i is the latent variable for person i. For this graded class, a standard normal prior is used as is conventional in IRT. Because someone in the no-symptom class will never endorse the item, and someone in the all-symptom class will always endorse the item, these two classes have item response models that are degenerate, with $P_{gj} = 0$ for the no-symptom class and $P_{gj} = 1$ for the all-symptom class. Finkelman et al. (2011) used an EM algorithm to estimate IRT model parameters and proportions of respondents in each class. Because the authors used a nationally representative sample in their data example, they used maximum pseudo likelihood estimation

with survey weights.

They let $\boldsymbol{u} = (u_1, ..., u_J)$ be a vector of 0s and 1s representing the response pattern for an individual, with $\mathbf{0} = (0, ..., 0)$ and $\mathbf{1} = (1, ..., 1)$ denoting the response patterns for the no-symptom class and all-symptom class, respectively. They let $N_{\boldsymbol{u}}$ be the sum of the weights of individuals with response pattern $\boldsymbol{u}$, with special cases N_0 and N_1 for individuals with no symptoms and all symptoms, respectively. For a set of item parameters $\boldsymbol{\alpha}$, the log likelihood is expressed

$$\begin{aligned} l(\boldsymbol{\alpha}, \pi_0, \pi_1, \pi_2; \{\boldsymbol{u}_i\}_1^I) = N_0 \log[\pi_0 + \pi_2 P(\boldsymbol{U} = \mathbf{0} | \boldsymbol{\alpha}, \phi(\theta), I_2 = 1)] \\ + N_1 \log[\pi_1 + \pi_2 P(\boldsymbol{U} = \mathbf{1} | \boldsymbol{\alpha}, \phi(\theta), I_2 = 1)] \\ + \sum_{\boldsymbol{u} \notin \{\mathbf{0}, \mathbf{1}\}} N_u \log[\pi_2 P(\boldsymbol{U} = \boldsymbol{u} | \boldsymbol{\alpha}, \phi(\theta), I_2 = 1)] \end{aligned} \tag{13}$$

in which $P(\boldsymbol{U} = \boldsymbol{u} | \boldsymbol{\alpha}, \phi(\theta), I_2 = 1)$ is the probability of observing response pattern $\boldsymbol{u}$ for someone in the graded component of the model. Like zero-inflated count data, a main issue with this type of data is that the people with response pattern $\mathbf{0}$ and in the no-symptom class are indistinguishable from those with response pattern $\mathbf{0}$ who are in the graded class, and the people with response pattern $\mathbf{1}$ and in the all-symptom class are indistinguishable from those with response pattern $\mathbf{1}$ who are in the graded class. That is, the total weights of graded component individuals satisfying $\boldsymbol{u} = \mathbf{0}$ and $\boldsymbol{u} = \mathbf{1}$ are treated as missing data. The observed data are the weights and response patterns of all individuals satisfying $\boldsymbol{u} \notin \{\mathbf{0}, \mathbf{1}\}$ (and therefore in the graded component). Finkelman et al. (2011) used an EM algorithm in M*plus* to estimate the IRT parameters and proportions of people in each class. They demonstrated an empirical example using data from the oppositional defiant disorder (ODD) and conduct disorder (CD) symptom scales of the US National Comorbidity Survey Adolescent supplement. As expected, the two-class and three-class mixture IRT models fit the data better than a unidimensional 2PL model, suggesting subsets of respondents with extreme levels of the latent variable who are not well described by the same model as the general population.

More recently, Wall et al. (2015) proposed a similar model for measuring psychiatric disorder severity in a zero-inflated population. Like Finkelman et al. (2011), Wall et al. (2015) treated

each symptom on a clinical checklist of alcohol dependence as a binary item, with endorsement of the item/symptom suggesting higher levels of the latent variable. As is common in clinical assessment, a large proportion of their sample did not endorse any of the symptoms, potentially representing a fundamentally different subset of individuals from those the assessment was designed to measure; therefore, instead of assuming normality of the latent variable, the authors used a composite population distribution composed of a mixture of normals, with a degenerate component for the subgroup of respondents who did not endorse any of the symptoms. They considered the population to comprise G subgroups, each coming from a normal distribution with mean μ_g and variance σ_g^2, mixed in proportion to the group sizes, η_1, η_2,..., η_G; for the people endorsing none of the symptoms, they assumed a degenerate distribution. Like Finkelman et al. (2011), they used a 2PL IRT model to describe the non-degenerate class, using maximum likelihood with an EM algorithm in M*plus* to estimate model parameters. Results of their simulation study suggested that the assumption of a normal latent variable distribution in the presence of a non-pathological subset of respondents leads to biased item parameter estimates and shrunken IRT scores that provide less separation of individuals.

Mixture IRT can also be conceptualized as testing for differential item functioning (DIF) when the groups are unknown (Cohen & Bolt, 2005; Samuelson, 2008; von Davier & Rost, 2006). Finch and Pierson (2011) used mixture IRT to analyze items on the Youth Risk Behavior Survey (YRBS). They were interested in identifying subgroups of adolescents most at risk for engaging in risky behaviors based on observed response patterns, allowing for potential differences in item parameters depending on latent class membership. Specifically, they examined whether certain items were more or less discriminating or easily endorsed across different at-risk subgroups in the population. Their model was based on the 2PL IRT model with group-specific item parameters and level of the latent variable,

$$P_j(U_j = 1|g, \theta_g) = \frac{\exp\{a_{jg}(\theta_g - b_{jg})\}}{1 + \exp\{a_{jg}(\theta_g - b_{jg})\}}. \tag{14}$$

Each respondent is assigned to a latent class as part of model estimation, and the proportion of respondents belonging to each class (π_g) is estimated under the constraint $\sum_{g=1}^{G} \pi_g = 1$. The authors fit a 2PL mixture IRT model to 14 items from the YRBS relating to engagement

in risky behaviors, such that endorsement of the item suggested greater propensity for risk taking behavior; it is worth nothing, however, that not all items were originally binary. For example, the item "How many times have you smoked marijuana in the previous month" was recoded from its original binned count response to a binary response. While item format was not central to the authors' research goals, it is important to point out that the practice of dichotomizing count responses does occur in psychological research, and while it reduces analytic burden, the loss of information may have unintended consequences. Understanding the potential effects of binning or dichotomizing count data is one of the motivating factors for the current research. Finch and Pierson (2011) found that a four-class mixture IRT solution fit the data best, and that certain items were more discriminating or easily endorsed depending on the latent class. For example, they identified a latent class with an especially high propensity of engaging in risky sexual activities; for this particular class, the items pertaining to sexual behavior were not as discriminating as items relating to other risky behaviors. The authors noted the potential usefulness of mixture IRT modeling in clinical assessment, with group-specific parameters allowing mental health professionals to focus on items and behaviors that are most discriminating in identifying the highest-risk adolescents within specific subgroups of risk-takers.

Sawatzky et al. (2012) applied a similar mixture IRT model to the measurement of patient reported outcomes. Like Finch and Pierson (2011), they were interested in fitting an item response model to data from a potentially heterogeneous population. Their model was based on the graded response model (GRM), in which the probability of responding at or above category k on item j is expressed

$$P_{jk}(U_j \geq k|\theta) = \frac{\exp\{a_j(\theta - b_{jk})\}}{1 + \exp\{a_j(\theta - b_{jk})\}} \tag{15}$$

where b_{jk} is the threshold between the categories of item j and a_i is the discrimination parameter for item j. To account for potential population heterogeneity, the authors introduced an additional subscript, g, for group membership:

$$P_{jkg}(U_j \geq k|\theta, G = g) = \frac{\exp\{a_{jg}(\theta - b_{jkg})\}}{1 + \exp\{a_{jg}(\theta - b_{jkg})\}}. \tag{16}$$

G is the latent class variable with g classes, and $f(x) = \sum_{g=1}^{G} \pi_g f_g(u)$, where f is the mixture of latent class-specific distributions and π_k is the proportion of respondents in each class, with $\sum_{g=1}^{G} \pi_g = 1$. This model is very similar to that of Finch and Pierson (2011), only with the GRM used in place of the 2PL model. The authors used maximum likelihood via an EM algorithm to estimate model parameters. As an empirical example, they analyzed the physical functioning subscale of the SF-36 Health Survey, consisting of 10 items designed to measure the general health status of the US population. Each item asked the respondent to report the degree to which his or her health limits daily activities; response options are "Yes, limited a lot", "Yes, limited a little", or "No, not limited at all". One might hypothesize that the population of interest comprises qualitatively different subgroups. For example, in the general population there is possibly a group of people who experience no physical functioning impairment; this group may interpret the items differently from the rest of the population, making the scale less appropriate for samples from that population. Sawatzky et al. (2012) used mixture IRT to answer this question empirically and arrived at a three-class solution, suggesting a class of people with very few physical limitations and two classes of people with more severe physical functioning limitations. The second class reported less struggle with more routine physical functioning activities such as climbing stairs and walking a block, whereas the third class reported struggle with both routine and more challenging activities. Mixture IRT highlights those items that may be best suited for the assessment of subgroups within clinical populations.

4. Research Questions

A review of the literature suggests three methodological approaches for solving three distinct problems in measurement: Poisson and zero-inflated Poisson psychometric models for the analysis of multivariate zero-inflated count data (L. Wang, 2010); latent variable models to account for heaping and response style, including heaping in univariate count outcomes (e.g., H. Wang & Heitjan, 2008) and extreme responding in multivariate item response data (e.g., Böckenholt, 2012; Bolt & Johnson, 2009; Bolt & Newton, 2011; De Boeck & Partchev, 2012; Moors, 2008; Thissen-Roe & Thissen, 2013); and mixture IRT models for clinical assessment in potentially 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, I am unaware of any existing

studies that provide an item response modeling framework for multivariate count outcomes that also accounts for zero-inflation, inflation at the maximum (K-inflation), and heaping. Finkelman et al.'s (2011) mixture IRT model accounts for inflation due to large proportions of respondents with a complete absence or very severe presence of the latent variable, but their model includes only binary items that comprise a checklist of symptoms. The same limitation applies to Wall et al.'s (2015) mixture IRT model for zero-inflation in the measurement of psychiatric disorders and Finch and Pierson's (2011) mixture IRT model for the analysis of risky youth behavior. Sawatzky et al.'s (2012) mixture IRT model is slightly more flexible, incorporating ordinal rating-scale items into the analysis of patient reported outcomes across distinct classes, but their model still includes only traditional item types that are well suited for conventional IRT models. A more flexible modeling approach that incorporates complex item types – particularly those eliciting inflated count responses – is still needed. Wang's (2010) IRT-ZIP model partially addresses this need, but her model ignores the issue of heaping at preferred digits. Statisticians have developed methods to account for heaping in univariate count data (e.g., H. Wang & Heitjan, 2008), but these models tend to consider only a single count outcome; conversely, the extreme response literature in IRT addresses response behavior in multivariate data but does not consider count responses (e.g., Böckenholt, 2012; Bolt & Johnson, 2009; Bolt & Newton, 2011; De Boeck & Partchev, 2012; Thissen-Roe & Thissen, 2013). The goal of this research is to borrow elements from all three existing methodological approaches in developing a latent class IRT model for multivariate zero-inflated count data that are sampled from a potentially heterogeneous population. The purpose of including latent classes is not only to account for inflation at zero and the maximum, but also to account for differences in response style due to digit or "nice number" preference.

4.1 Primary Aim #1: Addressing Inflation and Heaping in Count IRT Models

In incorporating count items into questionnaires, what are some parsimonious ways to account for zero-inflation, K-inflation, and heaping? This can be viewed as a latent class IRT problem, in which latent classes may reveal subpopulations of respondents. Specifically, one latent class for zero-inflation may represent people who are at a floor level of the latent variable, or people to whom the questions do not apply. Another latent class may represent those who have a very severe presence of the latent variable (K-inflation). Latent classes may also be able

to account for individual differences in "nice number" preferences on open-ended count scales, such as the propensity to select responses that are multiples of five. Depending on the latent class, a different IRT model can be used to describe the conditional probability of a count response. The motivating data for Primary Aim #1 are responses to the four count items from the BRFSS that form an emotional health subscale, as shown in Figure 1.

Scoring Because clinicians are often interested in scoring individuals on psychological assessments, an additional goal of Primary Aim #1 is to compute IRT scores for people based on their observed responses to the four count items. For this example, the IRT scores, which are computed from IRT models that depend on both a latent variable and a latent class membership, represent an emotional health latent variable, where higher scores suggest worse emotional health. The goal of this aim is to compute IRT scores from the latent class IRT model and examine different options for reporting scores at the individual-level.

4.2 Primary Aim #2: Are Complex Models Really Needed?

A pragmatic follow-up question to that posed in Primary Aim #1 is whether such complex modeling techniques are needed, or whether more parsimonious models that omit any of the latent classes describe the data just as well. The latent class IRT model that accounts for zero inflation, maximum inflation, and individual differences in response style is the most complex model, comprising four distinct latent classes; however, simpler models can be obtained by fixing to 0 different components of the log likelihood of that most complex model. For example, is it necessary to account for heaping at multiples of five, or will a model that does not account for heaping describe the data just as well?

4.3 Secondary Aim: What is the value added of including count items on scales?

There is evidence that researchers use open-ended count items on scales, but do these items provide information above and beyond more conventional response formats? Specifically, do open-ended count items contribute to measurement precision? This question is motivated by a specific set of item responses from the BRFSS, shown in Figure 2. The first six items are traditional Likert-type items that can take on any of five response categories; the last item is a count item ranging from 0 to 30 and exhibits the same inflation and heaping properties as the items described in Primary Aim #1. To what extent does including the count item increase measurement precision? The answer to this question may have implications for researchers

considering including open-ended count items on their scales.

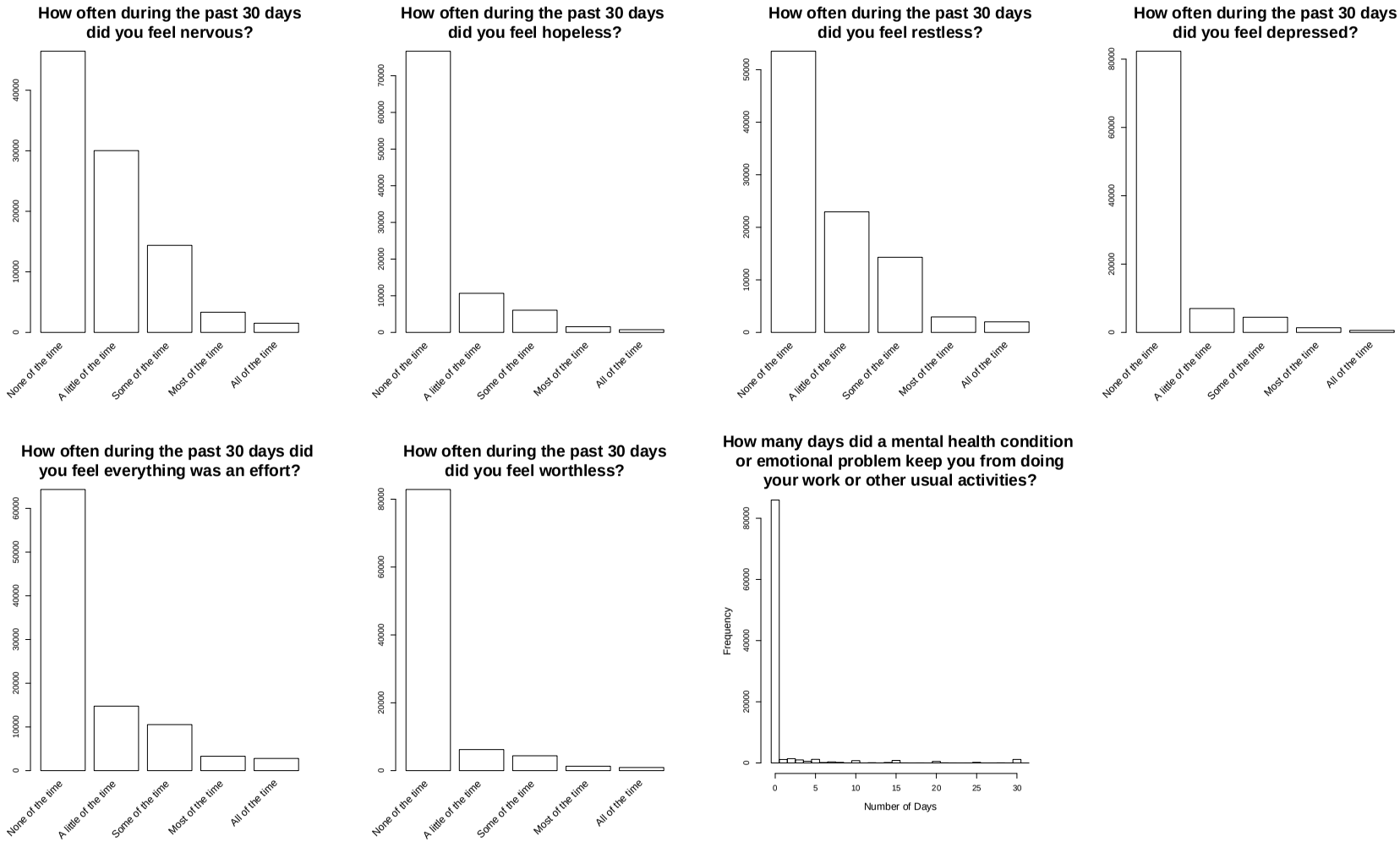


Figure 2. Frequency histograms for the emotional health subscale from the BRFSS (2012) comprising six Likert-type items and one count item.

Chapter 2: Method

1. Primary Aim #1: Addressing Inflation and Heaping in Multivariate Count Data

The motivating data set for Primary Aim #1 comprises responses to the four items from the BRFSS (2014) shown in Figure 1 in the previous chapter; these four items are thought to measure a general emotional health construct, in which higher observed counts indicate worse emotional health. As described in Chapter 1, there are several analytic challenges present in these data. Most notably, there is substantial inflation at 0 days, moderate inflation at 30 days, and heaping at days that are multiples of five. I hypothesized that the inflated proportions of respondents at these values can be accounted for with latent classes, where each latent class is characterized by a different IRT model.

First, I present the most general latent class model – this model applies to any set of item response data. Then, I describe the hypothesized latent classes that are specific to the item response data in Figure 1. Once the latent classes are defined, the class-specific conditional probabilities of a count response u_j can be modeled by specifying IRT models that are appropriate for the specific type of response (e.g., a nominal response model for multinomial responses, a graded response model for ordinal responses, etc.). For simplicity, subscript notation i for individuals is dropped; only subscripts j for item and g for latent class are included.

1.1 A Latent Class Model

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

$$P_j = \sum_{g=1}^{G} \pi_g P_{gj}(U_j = u_j), \tag{17}$$

in which g denotes latent class membership, π_g specifies the probability of belonging to latent class g, and $P_{gj}(U_j = u_j)$ is the conditional probability of observing response u_j from someone in latent class g (Hagenaars & McCuthcheon, 2002; Skrondal & Rabe-Hesketh, 2004). The

proportions of members in each latent class must sum to one, $\sum_{g=1}^{G} \pi_g = 1$. This is the general form of a latent class model: $P_{gj}(U_j = u_j)$ can be any type of probability function (IRT model), and the model can be applied to any set of item response data. In the sections that follow, I propose specific latent classes based on the BRFSS data in Figure 1.

Defining the Latent Classes To model the item response data shown in Figure 1, I propose four mutually exclusive latent classes. One latent class describes some, perhaps many, of the people who respond 0 days to all four items, thus having response vector $\boldsymbol{U} = \mathbf{0} = (0, 0, 0, 0)$. 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. An important distinction is that these people are qualitatively different from those who are not at the floor level of the latent variable but still exhibit a response vector of zeros, such as an individual who has some degree of depression but did not experience any symptoms in the past month. Similarly, a second latent class describes some, perhaps many, of the people who respond 30 days to all four items; these respondents have a response vector $\boldsymbol{U} = \mathbf{30} = (30, 30, 30, 30)$. These two latent classes correspond to response vectors **0** and **30** and will be referred to as the zero class and the maximum class, respectively. The item response models for these two classes are degenerate: People belonging to the zero latent class respond 0 to every item with a probability of 1, and people belonging to the maximum latent class respond 30 to every item with a probability of 1.

In addition to the zero and maximum classes, I propose two graded latent classes that describe the people falling along the continuum of the latent variable. The first of the graded classes, which I refer to as the exact count class, comprises the subset of people whose responses follow a standard count distribution. These are the people who are at some level of the latent variable and respond to the item as intended; that is, they report the exact number of days they experienced each symptom in the last month, regardless of whether those numbers are multiples of five. To model the conditional probability of count response u_j from members of the exact count latent class, a Poisson or negative binomial IRT model can be used; reasons for choosing each type of count model are explicated in later sections. The other graded class, which I refer to as the rounding/selected response class, represents the subset of respondents who select a multiple of five that may be near their true count. For example, someone belonging

to the rounding/selected response class may have a true count of 9 days but an observed count of 10 days. Instead of responding to the item as intended – as an exact count taking on any of 31 non-negative integers – these individuals may treat the item as multiple choice with only seven response categories, $\{0, 5, 10, 15, 20, 25, 30\}$, or they may round their true count to the nearest multiple of five. Because respondents belonging to the rounding/selected response class do not interact with the item as a true count, the Poisson and negative binomial IRT models are not likely to accurately describe the item response function for this class. Instead, a nominal response IRT model, designed for multinomial item response data, may be more appropriate.

Figure 3 shows a diagram of the proposed latent response processes that result in each of the 31 possible observed counts. According to the model, there are three internal response processes that can manifest as a zero count. One possibility is that the respondent is a member of the zero class and thus selects 0 days for every item, $\boldsymbol{U} = \mathbf{0} = (0, 0, 0, 0)$; this option is represented by the direct path from the item to a zero response, without passing through either of the two IRT models. A second possibility is that the person is part of the exact count class and happens to report 0 days; that is, the respondent is at some level of the latent variable but has not experienced any symptoms in the past 30 days. This option is represented by one indirect path from the item to the zero response: The respondent enters the count IRT model, and it is through this IRT model that a response of zero is observed. A third possibility is that the respondent is a member of the rounding/selected response class and reports 0 days; these are people who are inclined to report days that are multiples of five. Instead of passing through the count IRT model, these individuals arrive at zero via the nominal response IRT model. Similar to the three distinct response processes that manifest as a zero count, there are three distinct response processes that result in an observed count of 30 days. These response processes are analogous to those described for the zero case and are shown in Figure 3.

Figure 3 shows that for the subset of people belonging to one of the two graded classes, fewer response processes are possible. Only two distinct paths lead to observed responses that are non-zero and non-30 multiples of five. The response may be from someone in the exact count class and thus represent a count that is a realization of a Poisson or negative binomial random variable. On the other hand, the response may belong to someone in the rounding/selected response class. If the response is from a member of the rounding/selected response class, it is not a realization of

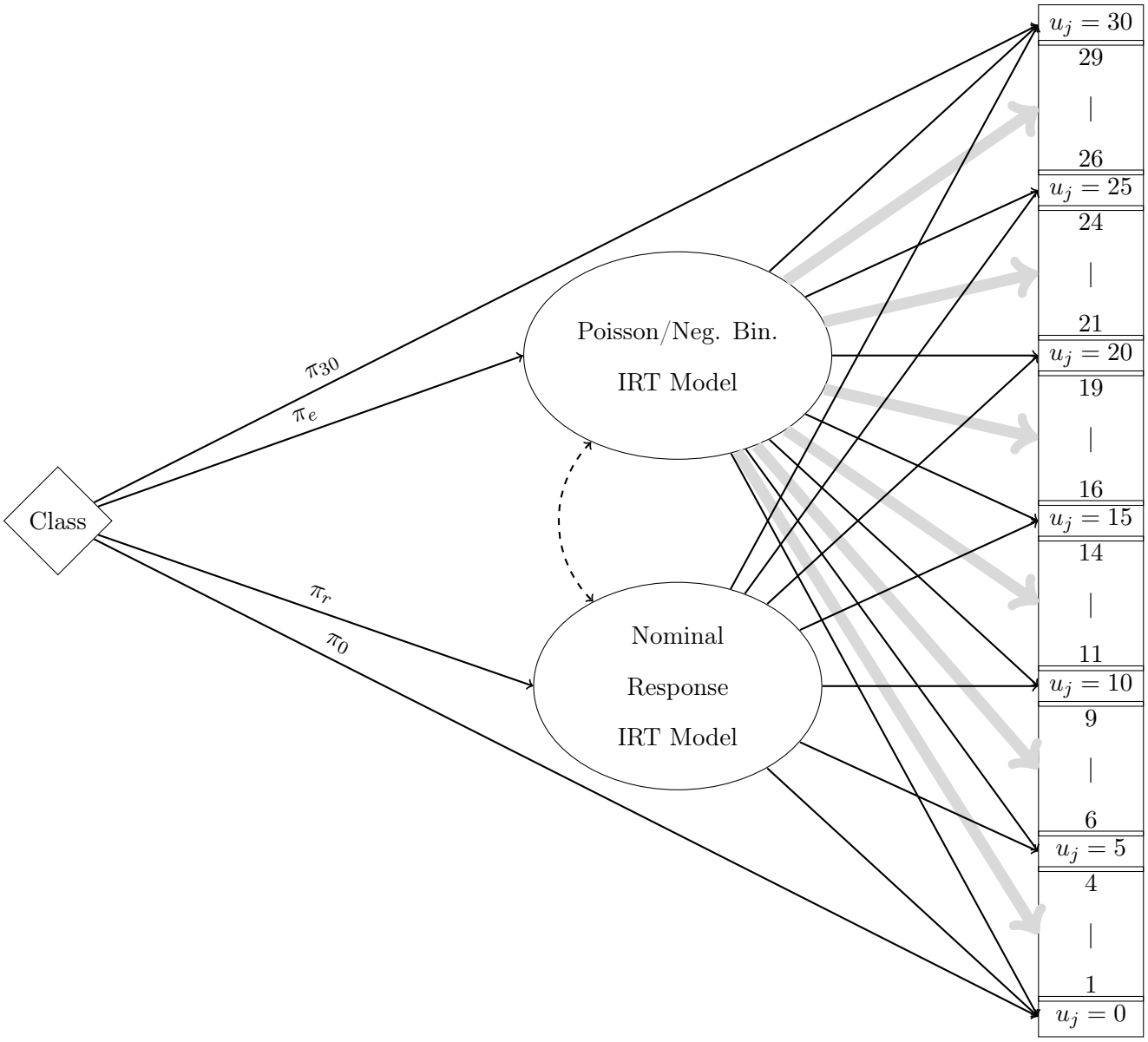


Figure 3. Tree diagram of full latent class IRT model

a count random variable; rather, it is a realization of a multinomial random variable with seven response categories. Both the Poisson/negative binomial and nominal response IRT models may yield the same manifest count; however, it is through different response processes that this multiple-of-five count is reported. The remaining responses that have not been addressed are the exact counts that are not multiples of five. The only path to these exact counts is via the Poisson/negative binomial IRT model; people with responses that are not multiples of five are therefore known to belong to the exact count class. It is worth noting that it is possible to

determine a lower bound estimate of people in the exact count class by simply counting the number of people with response patterns including at least one response that is not a multiple of five.

1.2 Count IRT Models for the Exact Count Class

Typically, latent class item response models use Bernoulli or multinomial conditional response distributions for $P_{jg}(U_j = u_j)$ in Equation 17; in principle, however, the conditional response distribution can be any type of probability function. To model the item responses of people in the exact count class, any IRT model that employs a standard count distribution to describe the conditional item responses can be used. The next two sections describe two different count IRT models: the Poisson IRT model and the negative binomial IRT model.

Poisson IRT Model The conditional probability of response u_j is commonly referred to as the trace line, $T_j(U_j = u_j|\theta)$, because it is the curve that traces the conditional probability of an item response u as a function of the latent variable (Lazarsfeld, 1959; Thissen & Wainer, 2001). To derive the mathematical expression for a Poisson trace line, first consider a random variable U_j that is a count. This count can be modeled as a Poisson distributed random variable,

$$P_j(U_j = u_j) = \frac{\lambda_j^{u_j} \exp(-\lambda_j)}{u_j!}, \quad u_j = 0, 1, 2, ... \tag{18}$$

in which $\lambda_j > 0$ is the expected value of U_j. Within the Poisson regression framework, one can model the expected value of a Poisson random variable using a linear combination of unknown parameters; within the IRT framework, one instead models the expected value of a Poisson random variable with a non-linear combination of parameters that includes the latent variable θ and item parameters a_j and c_j. Using the exponential function to restrict λ_j to non-negative values, the expected value of count item U_j can then be expressed

$$E(U_j|\theta, a_j, c_j) = \lambda_j = \exp(a_j\theta + c_j). \tag{19}$$

The a_j parameter is the item discrimination: The larger the value of a_j, the more discriminating the item is in separating individuals on the latent variable θ. The c_j parameter is the item intercept; it is the expected value of the count item for someone with $\theta = 0$ (i.e., the average level of the latent variable). Wang (2010) expressed the expected count in slope-threshold

form, $a_j(\theta - b_j)$; for computational purposes, I reparameterize the expected count to have slope-intercept form, $a_j\theta + c_j$. Substituting the expression for λ_j into the probability mass function in Equation 18 results in the IRT trace line for a Poisson count item,

$$T_j(U_j = u_j|\theta) = \frac{\exp\{-\exp(a_j\theta + c_j)\} \times \exp\{u_j(a_j\theta + c_j)\}}{u_j!}. \tag{20}$$

The Poisson trace line in Equation 20 is one plausible IRT model for members of the exact count class. Unlike Wang's (2010) model, this Poisson IRT model does not include a zero-inflated mixture component. This is because the proposed latent class IRT model already accounts for zero-inflation by including a zero class for individuals who are at a floor level of the latent variable or to whom the items do not apply.

Negative Binomial IRT Model The negative binominal distribution, which is a generalization of the Poisson distribution, accounts for excess variability in the observed counts by treating the Poisson parameter λ_j in Equation 18 as a random variable. The negative binomial distribution has several different parameterizations; the probability mass function that most clearly shows the negative binomial distribution as a generalization of the Poisson distribution is

$$P_j(U_j = u_j) = \left(\frac{\Gamma(u_j + \delta_j^{-1})}{\Gamma(u_j + 1)\Gamma(\delta_j^{-1})}\right)\left(\frac{\delta_j^{-1}}{\delta_j^{-1} + \lambda_j}\right)^{\delta_j^{-1}}\left(\frac{\lambda_j}{\delta_j^{-1} + \lambda_j}\right)^{u_j}, \tag{21}$$

in which $\lambda_j > 0$ is the expected value of count item U_j, and δ_j is an overdispersion parameter. It is important to note that as δ^{-1} approaches $+\infty$ in Equation 21, or equivalently, as the overdispersion parameter δ approaches 0, the negative binomial distribution converges to the Poisson distribution. Therefore, overdispersion parameter estimates near zero suggest that the Poisson IRT model is sufficient; otherwise, a negative binomial IRT model may be more appropriate. As true of the Poisson IRT model, the expected value λ_j can be modeled as a non-linear function of the latent variable θ and item parameters a_j and c_j, where

$$E(U_j|\theta, a_j, c_j, \delta_j) = \lambda_j = \exp(a_j\theta + c_j). \tag{22}$$

As before, a_j and c_j are the item discrimination and intercept parameters, respectively; they

are interpreted the same way as in the Poisson IRT model, but during parameter estimation, their values are adjusted for overdispersion. Substituting the expression for λ_j in Equation 22 into the probability mass function in Equation 21 yields the trace line for the negative binomial IRT model,

$$T_j(U_j = u_j|\theta) = \left(\frac{\Gamma(u_j + \delta_j^{-1})}{\Gamma(u_j + 1)\Gamma(\delta_j^{-1})}\right)\left(\frac{\delta_j^{-1}}{\delta_j^{-1} + \exp(a_j\theta + c_j)}\right)^{\delta_j^{-1}}\left(\frac{\exp(a_j\theta + c_j)}{\delta_j^{-1} + \exp(a_j\theta + c_j)}\right)^{u_j}. \tag{23}$$

Because I was unsure of the level of dispersion in the BRFSS data, I considered both the Poisson and negative binomial IRT models as possible item response functions for members of the exact count class.

1.3 Nominal Response IRT Model for the Rounding/Selected Response Class

The rounding/selected response class of the latent class IRT model includes individuals who respond only with counts that are multiples of five. Instead of treating the item as having an open-ended count response scale, these individuals treat the item as having a smaller, fixed number of response categories: $\{0, 5, 10, 15, 20, 25, 30\}$. Thus, instead of following a Poisson or negative binomial distribution, conditional responses from members of the rounding/selected response class more plausibly follow a multinomial distribution. To avoid making assumptions about the inherent ordering of the response categories – that is, the assumption that as the level of the latent variable increases, so does the probability of selecting a successively higher count category – the nominal response model (NRM; Bock, 1972, 1997; Thissen, Cai, & Bock, 2010) can be used. The NRM is a multivariate generalization of the logistic regression model – equivalent models from statistics include the polytomous logistic regression, the baseline category logits model, or the generalized logits model. The trace line for the NRM is expressed

$$T_j(U_j = k_j|\theta) = \frac{\exp(a_{jk}\theta + c_{jk})}{\sum_{m=1}^{M}\exp(a_{jm}\theta + c_{jm})}, \tag{24}$$

in which k_j corresponds to the response category for item j. To avoid confusion with the count response u_j that can take any non-negative integer value up to 30, I 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 five. In Equation 24, a_k is the slope parameter and c_k 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. The probability of endorsing response category k is influenced by both the propensity to select that category and the propensities toward selecting the alternatives to that category; thus, the NRM is known as a "divide-by-total" model (Thissen & Steinberg, 1986). An advantage of the NRM over some of the more common IRT models for polytomous data is that the ordering of response categories can be examined empirically after fitting the NRM to the data.

1.4 The Full Latent Class IRT Model

Let I_0, I_e, I_r, and I_{30} be indicator variables denoting membership in the zero class, exact count class, rounding/selected response class, and maximum class, respectively, with probabilities π_0, π_e, π_r, and $\pi_{30} = 1 - \pi_0 - \pi_e - \pi_r$. Assuming these four mutually exclusive latent classes, the general latent class model in Equation 17 can be written

$$\begin{aligned} P_j = {} & \pi_0[P_{0j}(U_j = 0) = 1; P_{0j}(U_j \neq 0) = 0] \\ & + \pi_e T_{ej}(U_j = u_j | \theta_e, I_e = 1; a_j, c_j) \\ & + \pi_r T_{rj}(U_j = k_j | \theta_r, I_r = 1; a_{jk}, c_{jk}) \\ & + \pi_{30}[P_{30j}(U_j = 30) = 1; P_{30j}(U_j \neq 30) = 0] \end{aligned} \tag{25}$$

where $u_j = \{0, 1, ..., 30\}$ 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 $\boldsymbol{U} = \mathbf{0} = (0, 0, 0, 0)$ and $N_{\mathbf{30}}$ be the number of people with response pattern $\boldsymbol{U} = \mathbf{30} = (30, 30, 30, 30)$. Let $N_{\boldsymbol{u}}$ be the number of people with response pattern $\boldsymbol{u}$. For notational simplicity, let $\boldsymbol{e}$ be any response pattern that includes at least one non-zero and non-30 exact count and $\boldsymbol{k}$ be any response pattern that includes only multiples of five, excluding response patterns $\mathbf{0}$ and $\mathbf{30}$. Given response patterns $\boldsymbol{U} = \boldsymbol{u}$, the likelihood of item parameters $\boldsymbol{\alpha} = \{\boldsymbol{a_j}, \boldsymbol{c_j}, \boldsymbol{a_{jk}}, \boldsymbol{c_{jk}}\}$ (or $\boldsymbol{\alpha} = \{\boldsymbol{a_j}, \boldsymbol{c_j}, \boldsymbol{a_{jk}}, \boldsymbol{c_{jk}}, \boldsymbol{\delta_j}\}$ if the negative binomial IRT model is used in place of the Poisson IRT model) for items $j = 1, ..., J$, as well as the latent class proportions π_0, π_e, π_r, and π_{30} can be expressed

$$
\begin{aligned}
L(\boldsymbol{\alpha}, \pi_0, \pi_e, \pi_r, \pi_{30}; \{\boldsymbol{u}\}_1^J) = {} & [\pi_0 + \pi_e T(\boldsymbol{U} = \mathbf{0}|\boldsymbol{\alpha}, \theta, I_e = 1) + \pi_r T(\boldsymbol{U} = \mathbf{0}|\boldsymbol{\alpha}, \theta, I_r = 1)]^{N_{\mathbf{0}}} \\
& \times [\pi_{30} + \pi_e T(\boldsymbol{U} = \mathbf{30}|\boldsymbol{\alpha}, \theta, I_e = 1) + \pi_r T(\boldsymbol{U} = \mathbf{30}|\boldsymbol{\alpha}, \theta, I_r = 1)]^{N_{\mathbf{30}}} \\
& \times \prod_{\boldsymbol{u} \notin \{\mathbf{0}, \mathbf{30}, \boldsymbol{e}\}} [(\pi_r T(\boldsymbol{U} = \boldsymbol{k}|\boldsymbol{\alpha}, \theta, I_r = 1) + \pi_e T(\boldsymbol{U} = \boldsymbol{k}|\boldsymbol{\alpha}, \theta, I_e = 1))^{N_{\boldsymbol{u}}}] \\
& \times \prod_{\boldsymbol{u} \notin \{\mathbf{0}, \mathbf{30}, \boldsymbol{k}\}} [\pi_e T(\boldsymbol{U} = \boldsymbol{u}|\boldsymbol{\alpha}, \theta, I_e = 1)^{N_{\boldsymbol{u}}}].
\end{aligned} \tag{26}
$$

Taking the log of the likelihood yields

$$
\begin{aligned}
\log L(\boldsymbol{\alpha}, \pi_0, \pi_e, \pi_r, \pi_{30}; \{\boldsymbol{u}\}_1^J) = {} & N_{\mathbf{0}} \log[\pi_0 + \pi_e T(\boldsymbol{U} = \mathbf{0}|\boldsymbol{\alpha}, \theta, I_e = 1) + \pi_r T(\boldsymbol{U} = \mathbf{0}|\boldsymbol{\alpha}, \theta, I_r = 1)] \\
& + N_{\mathbf{30}} \log[\pi_{30} + \pi_e T(\boldsymbol{U} = \mathbf{30}|\boldsymbol{\alpha}, \theta, I_e = 1) + \pi_r T(\boldsymbol{U} = \mathbf{30}|\boldsymbol{\alpha}, \theta, I_r = 1)] \\
& + \sum_{\boldsymbol{u} \notin \{\mathbf{0}, \mathbf{30}, \boldsymbol{e}\}} N_{\boldsymbol{u}} \log[\pi_r T(\boldsymbol{U} = \boldsymbol{k}|\boldsymbol{\alpha}, \theta, I_r = 1) + \pi_e T(\boldsymbol{U} = \boldsymbol{k}|\boldsymbol{\alpha}, \theta, I_e = 1)] \\
& + \sum_{\boldsymbol{u} \notin \{\mathbf{0}, \mathbf{30}, \boldsymbol{k}\}} N_{\boldsymbol{u}} \log[\pi_e T(\boldsymbol{U} = \boldsymbol{u}|\boldsymbol{\alpha}, \theta, I_e = 1)].
\end{aligned} \tag{27}
$$

In Equations 26 and 27, $T(\boldsymbol{U} = \boldsymbol{k}|\boldsymbol{\alpha}, \theta, I_r = 1)$ traces the conditional probability of observing response pattern $\boldsymbol{k} = \boldsymbol{u} : u_j \in \{0, 5, 10, 15, 20, 25, 30\}$ for someone in the rounding/selected response class influenced by latent variable θ, $T(\boldsymbol{U} = \boldsymbol{k}|\boldsymbol{\alpha}, \theta, I_e = 1)$ traces the conditional probability of observing response pattern $\boldsymbol{k} = \boldsymbol{u} : u_j \in \{0, 5, 10, 15, 20, 25, 30\}$ for someone in the exact count class influenced by latent variable θ, and $T(\boldsymbol{U} = \boldsymbol{u}|\boldsymbol{\alpha}, \theta, I_e = 1)$ traces the conditional probability of observing response pattern with only exact counts for someone in the exact latent class influenced by latent variable θ.

Note that the individuals with response vector $\boldsymbol{U} = \mathbf{0}$ in the zero class are indistinguishable from individuals with the same response vector who are in the exact count or rounding/selected response classes. Similarly, people with response vector $\boldsymbol{U} = \mathbf{30}$ in the maximum class are indistinguishable from people with the same response vector who are in the exact count or rounding/selected response classes, and people in the rounding/selected response class with response patterns containing only multiples of five are indistinguishable from people in the

exact count class. Thus, the proportion of people in each of the four latent classes must be estimated as part of the model.

Simulation Before fitting the proposed latent class IRT model to the BRFSS data, I carried out a small simulation to determine whether the implementation in R of parameter estimation for the model is able to recover the population parameters when they are known. The simulation served as a model identification and programming check. To mirror the empirical data, I simulated 10,000 open-ended count responses to each of four hypothetical items, where each of the 10,000 observations was assigned to a particular latent class. The proportions of people in each of the zero, exact count, rounding/selected response, and maximum classes were set to 0.3, 0.2, 0.4, and 0.1, respectively. Depending on the latent class, I simulated the data from a different item response model – an IRT model for members of the exact count and rounding/selected response classes, or a degenerate model for members of the zero and maximum classes.

For the 20% of respondents in the exact count class, item response data were generated from a Poisson IRT model with the parameters a_j and c_j in Table 2; simulated counts were capped at 30 to reflect the empirical data that correspond to days of the month. For the 40% of respondents in the rounding/selected response class, item response data were generated from a nominal response IRT model with parameters a_{jk} and c_{jk} in Table 2; accordingly, only counts in $\{0, 5, 10, 15, 20, 25, 30\}$ were possible responses. To account for the remaining 30% and 10% of the population belonging to the zero and maximum classes, I added 3,000 response patterns of $\boldsymbol{U} = (0, 0, 0, 0)$ and 1,000 response patterns of $\boldsymbol{U} = (30, 30, 30, 30)$ to the sample, increasing the total sample size to $N = 10,000$. I then fit the proposed latent class IRT model to the simulated data set. In total, 59 parameters were estimated: 8 parameters for the Poisson IRT model, 48 parameters for the nominal response IRT model, and 3 latent class proportions (the 4 latent class proportions must sum to 1). I used a similar procedure to simulate data from a latent class IRT model that uses a negative binomial IRT model for the exact count class; all IRT parameters and latent class proportions were identical to those used in the Poisson simulation. The only difference was that in using the negative binomial trace line in place of the Poisson trace line, four additional parameters were estimated to account for overdispersion; those parameters are near the bottom of Table 2. I carried out all simulations in R. Sample

code for the simulations can be found in Appendix A.

Empirical Analysis of the BRFSS After fitting the two proposed latent class IRT models to the simulated data to examine parameter recovery and model identification, I then fit the models to the BRFSS data. I used the 2014 version of the BRFSS survey to conduct all analyses within Primary Aim #1. Specifically, I used four count items that form a subscale inquiring about emotional health status, where each item response is reported on a 0-30 day scale. These items relate to pain, depression, anxiety, and energy. For the pain, depression, and anxiety items, an increasing number of days suggests worse health; for energy, an increasing number of days suggests better health. To maintain consistency of scale direction, I reverse coded the energy item such that an increasing number of days is reflective of worse health. Throughout the remainder of this paper, I use the terms "Pain", "Depressed", "Anxious", and (reversed) "Energy" to refer to these items.

The analytic sample for Primary Aim #1 comprised 9,042 individuals who provided responses to the four emotional health count items. The BRFSS is a nationally representative sample; thus, sampling weights should be used if one wishes to draw population-level inferences. Because the goal of this project was to develop a latent class IRT model that could accommodate count data exhibiting inflation and heaping, and not necessarily to draw inferences about general health in the U.S. population, I ignored sampling weights in estimation. Additionally, I wished to make model comparisons using likelihood-based fit criteria; including sampling weights requires pseudo likelihood approaches and thus alternative model comparison techniques.

Estimation Parameter estimation was done via maximum likelihood using **nlm**, R's non-linear optimizer. **nlm** is a general optimizer that directly minimizes a user-specified function using a Newton-type algorithm; to implement maximum likelihood, I used **nlm** to minimize $(-1) \times \log L$, where $\log L$ is expanded in Equation 27. For the latent class model with a Poisson IRT model for the exact count class, a total of 59 parameters were estimated $(\boldsymbol{a}_{jk}, \boldsymbol{c}_{jk}, \boldsymbol{a}_j, \boldsymbol{c}_j, \pi_0, \pi_{30}, \pi_r)$; for the latent class model with a negative binomial IRT model for the exact count class, a total of 63 parameters were estimated $(\boldsymbol{a}_{jk}, \boldsymbol{c}_{jk}, \boldsymbol{a}_j, \boldsymbol{c}_j, \boldsymbol{\delta}_j, \pi_0, \pi_{30}, \pi_r)$. To avoid imposing parameter constraints during estimation, latent class membership proportions were estimated as logits and standard errors were obtained using the delta method. Similarly, for models that included a

Table 2. Simulation parameters for the full latent class IRT model, $N = 10,000$

	Item #1	Item #2	Item #3	Item #4
Nominal Response IRT 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
Count Model IRT Parameters				
a	1.15	1.15	1.30	0.80
c	1.10	0.00	1.20	1.30
δ (Neg. Bin. only)	0.40	0.50	0.48	0.34
Latent Classes	Zero	Exact Count	Rounding	Maximum
Proportion	0.30	0.20	0.40	0.10

negative binomial component, I estimated $\exp(\delta_j)$ to restrict the overdispersion parameter to positive values. Sample R code for model estimation can be found in Appendix A.

Scale Scores After fitting the proposed latent class IRT models to the BRFSS data, I used item parameter estimates from the class-specific IRT models to compute the trace lines for each of the four items. Depending on the latent class, I used two different sets of trace lines to score people: one set of trace lines for the exact count class, the other set of trace lines for the rounding/selected response class. For members of the exact count class, I computed scores using the negative binomial trace lines; for members of the rounding/selected response class, I computed scores using the nominal response trace lines.

For a given response pattern $\boldsymbol{U} = \boldsymbol{u} = (u_1, u_2, u_3, u_4)$, I computed the scale score as the mean of the posterior distribution of θ, where the posterior distribution is 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). These scores are more commonly known as response pattern *expected a posteriori* (EAP) scores and are expressed mathematically as

$$\hat{\theta}_{\text{EAP}} = \frac{\int_{-\infty}^{+\infty} \prod_{j=1}^{4} T_j(u_j)\theta d\theta}{\int_{-\infty}^{+\infty} \prod_{j=1}^{4} T_j(u_j) d\theta} \tag{28}$$

where $T_j(u_j)$ is the trace line for item j. In practice, the mean of the posterior density is computed by approximating the integral over a range of quadrature points q,

$$\hat{\theta}_{\text{EAP}} \approx \frac{\sum_1^q \prod_{j=1}^{4} T_{jq}(u_j)\theta_q d\theta}{\sum_1^q \prod_{j=1}^{4} T_{jq}(u_j) d\theta}. \tag{29}$$

Standard errors of scale scores are computed as the standard deviation of the posterior distribution of θ,

$$SD(\hat{\theta}_{\text{EAP}}) \approx \sqrt{\frac{\sum_1^q \prod_{i=1}^{4} T_j(u_j)(\theta_q - \hat{\theta}_{\text{EAP}})^2 d\theta}{\sum_i^q \prod_{i=1}^{4} T_j(u_j) d\theta}} \tag{30}$$

For individuals known to be members of the exact count class (i.e., their response patterns included at least one count that was not a multiple of five), I computed a single scale score using the estimated count model trace lines for $T_j(u_j)$ in Equation 28. 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 five), I computed two plausible scores. The first score assumes the person is a member of the exact count class and was computed using the count model trace lines; the second score assumes the person is a member of the rounding/selected response class and was computed using the NRM trace lines. It is important to note that some people with $\boldsymbol{U} = \mathbf{0} = (0, 0, 0, 0)$ or $\boldsymbol{U} = \mathbf{30} = (30, 30, 30, 30)$ response patterns may be members of the zero class or maximum class, respectively. In theory, members of these two latent classes should not be scored because it is likely that it is a different latent variable that influences their item responses – that is, it may not be the same general health latent variable that describes members of the two graded classes. To account for the proportion of individuals with all-0 or all-30 response patterns who belong to one of the graded classes, I computed scale scores for 726 of the people with $\boldsymbol{U} = \mathbf{0} = (0, 0, 0, 0)$ and 45 of the people with $\boldsymbol{U} = \mathbf{30} = (30, 30, 30, 30)$ as though these response patterns belonged to members of the exact count or rounding/selected response class. The number of all-0 and all-30 response patterns to score was determined from the latent class proportions that were estimated from fitting the latent class IRT model to the BRFSS data. Because anyone with a response pattern including only multiples of five was scored twice – once according to the set of count trace lines, and once according to the set of NRM trace lines – I computed the correlation between the two sets of scores to examine the practical implications of choosing one scoring method over the other.

2. Primary Aim #2: Are Complex Models Really Needed?

The second primary goal of this research was to evaluate whether such complex modeling techniques are really needed to analyze the four count items from the BRFSS, or whether constraining to zero different parts of the log likelihood in Equation 27 yields a more parsimonious model that describes the data just as well. I tested whether the rounding/selected response latent class is needed by comparing the fit of the model based on the log likelihood in Equation 27 to the fit of the model based on a log likelihood that fixes the probability of being a member of the rounding/selected response class to 0:

$$\begin{aligned}\log L(\boldsymbol{\alpha}, \pi_0, \pi_e, \pi_{30}; \{\boldsymbol{u}\}_1^J) = N_{\mathbf{0}} \log[\pi_0 + \pi_e T(\boldsymbol{U} = \mathbf{0} | \boldsymbol{\alpha}, \theta, I_e = 1)] \\ + N_{\mathbf{30}} \log[\pi_{30} + \pi_e T(\boldsymbol{U} = \mathbf{30} | \boldsymbol{\alpha}, \theta, I_e = 1)] \\ + \sum_{\boldsymbol{u} \notin \{\mathbf{0}, \mathbf{30}\}} N_{\boldsymbol{u}} \log[\pi_e T(\boldsymbol{U} = \boldsymbol{u} | \boldsymbol{\alpha}, \theta, I_e = 1)].\end{aligned} \tag{31}$$

I also tested whether the maximum class is needed by fixing the probability of being a member of the maximum class to zero in the model log likelihood,

$$\begin{aligned}\log L(\boldsymbol{\alpha}, \pi_0, \pi_e, \pi_r; \{\boldsymbol{u}\}_1^J) = N_{\mathbf{0}} \log[\pi_0 + \pi_e T(\boldsymbol{U} = \mathbf{0} | \boldsymbol{\alpha}, \theta, I_e = 1) + \pi_r T(\boldsymbol{U} = \mathbf{0} | \boldsymbol{\alpha}, \theta, I_r = 1)] \\ + \sum_{\boldsymbol{u} \notin \{\mathbf{0}, \boldsymbol{e}\}} N_{\boldsymbol{u}} \log[\pi_r T(\boldsymbol{U} = \boldsymbol{k} | \boldsymbol{\alpha}, \theta, I_r = 1) + \pi_e T(\boldsymbol{U} = \boldsymbol{k} | \boldsymbol{\alpha}, \theta, I_e = 1)] \\ + \sum_{\boldsymbol{u} \notin \{\mathbf{0}, \boldsymbol{k}\}} N_{\boldsymbol{u}} \log[\pi_e T(\boldsymbol{U} = \boldsymbol{u} | \boldsymbol{\alpha}, \theta, I_e = 1)]\end{aligned} \tag{32}$$

and whether the zero class is needed by fixing the probability of being a member of the zero class to zero,

$$\begin{aligned}\log L(\boldsymbol{\alpha}, \pi_e, \pi_r, \pi_{30}; \{\boldsymbol{u}\}_1^J) = N_{\mathbf{30}} \log[\pi_{30} + \pi_e T(\boldsymbol{U} = \mathbf{30} | \boldsymbol{\alpha}, \theta, I_e = 1) + \pi_r T(\boldsymbol{U} = \mathbf{30} | \boldsymbol{\alpha}, \theta, I_r = 1)] \\ + \sum_{\boldsymbol{u} \notin \{\mathbf{30}, \boldsymbol{e}\}} N_{\boldsymbol{u}} \log[\pi_r T(\boldsymbol{U} = \boldsymbol{k} | \boldsymbol{\alpha}, \theta, I_r = 1) + \pi_e T(\boldsymbol{U} = \boldsymbol{k} | \boldsymbol{\alpha}, \theta, I_e = 1)] \\ + \sum_{\boldsymbol{u} \notin \{\mathbf{30}, \boldsymbol{k}\}} N_{\boldsymbol{u}} \log[\pi_e T(\boldsymbol{U} = \boldsymbol{u} | \boldsymbol{\alpha}, \theta, I_e = 1)].\end{aligned} \tag{33}$$

I used the AIC and BIC, which are likelihood based fit criteria that penalize for model complexity, to evaluate whether excluding latent classes substantially worsens model fit. In total, I made three model comparisons: the model with all four latent classes vs. a three-class model without the rounding/selected response class, the model with all four latent classes vs. a three-class model without the maximum class, and the model with all four latent classes vs.

a three-class model without the zero class. All latent class IRT models were estimated via maximum likelihood using R's **nlm** optimizer as described in the previous section, where the log likelihood that was maximized depended on the latent classes that were included in the model. I computed the AIC and BIC based on the maximized value of the log likelihood, $\log(\hat{L})$:

$$\text{AIC} = 2k - 2 \cdot \log(\hat{L}) \tag{34}$$

and

$$\text{BIC} = -2 \cdot \log(\hat{L}) + k \cdot \log(n) \tag{35}$$

where k is the number of estimated model parameters and n is the sample size ($n = 9,042$).

3. Secondary Aim: What Value Do Count Items Add to Scales?

The Secondary Aim of this research was motivated by an additional set of items from the BRFSS, shown in Figure 2 at the end of Chapter 1. Like the set of items used to address Primary Aim #1, these items also measure an emotional health construct. The first six items are Likert-type; the last item is a count. The purpose of the Secondary Aim was twofold: I was interested in whether a modified version of the latent class IRT model described in Primary Aim #1 could be fit to these data to accommodate both the Likert-type items and the count item, and if so, the degree to which the single count item contributes to measurement precision. If its contribution is trivial, one may be able to justify omitting the count item, obviating the need for more complex count IRT models. A scale comprising only Likert-type items poses far fewer analytic challenges than a scale that includes a count item, especially when the count item exhibits inflation and heaping.

To investigate whether a latent class IRT model could be used to describe the item responses from this scale of mixed item types, I proposed a similar model to the one explained earlier in this chapter. The general form of the log likelihood for the latent class IRT model that includes six Likert-type items and one count item is identical to that used in Primary Aim #1:

$$
\begin{aligned}
\log L(\boldsymbol{\alpha}, \pi_0, \pi_e, \pi_r, \pi_{30}; \{\boldsymbol{u}\}_1^J) = {} & N_{\mathbf{0}} \log[\pi_0 + \pi_e T(\boldsymbol{U} = \mathbf{0} | \boldsymbol{\alpha}, \theta, I_e = 1) + \pi_r T(\boldsymbol{U} = \mathbf{0} | \boldsymbol{\alpha}, \theta, I_r = 1)] \\
& + N_{\mathbf{30}} \log[\pi_m + \pi_e T(\boldsymbol{U} = \mathbf{30} | \boldsymbol{\alpha}, \theta, I_e = 1) + \pi_r T(\boldsymbol{U} = \mathbf{30} | \boldsymbol{\alpha}, \theta, I_r = 1)] \\
& + \sum_{\boldsymbol{u} \notin \{\mathbf{0}, \mathbf{30}, \boldsymbol{e}\}} N_{\boldsymbol{u}} \log[\pi_r T(\boldsymbol{U} = \boldsymbol{k} | \boldsymbol{\alpha}, \theta, I_r = 1) + \pi_e T(\boldsymbol{U} = \boldsymbol{k} | \boldsymbol{\alpha}, \theta, I_e = 1)] \\
& + \sum_{\boldsymbol{u} \notin \{\mathbf{0}, \mathbf{30}, \boldsymbol{k}\}} N_{\boldsymbol{u}} \log[\pi_e T(\boldsymbol{U} = \boldsymbol{u} | \boldsymbol{\alpha}, \theta, I_e = 1)].
\end{aligned}
\tag{36}
$$

While the general form of this log likelihood is the same as in Equation 27, some modifications are required to account for the additional items that have a different type of response format. First, the response vectors $\boldsymbol{U} = \boldsymbol{u}$ are longer due to an increased number of items: Instead of having four items, this scale has seven items, yielding response vectors $\boldsymbol{U} = \boldsymbol{u} = \{u_j, u_j, u_j, u_j, u_j, u_j, u_j\}$. Second, the item responses comprising $\boldsymbol{u} \notin \{\mathbf{0}, \mathbf{30}, \boldsymbol{e}\}$ and $\boldsymbol{u} \notin \{\mathbf{0}, \mathbf{30}, \boldsymbol{k}\}$ must now reflect the Likert-type response format. The first six items in Figure 2 have five possible response categories $U_j = z$, $z \in \{0, 1, 2, 3, 4\}$. The zero class is represented by response pattern $\boldsymbol{U} = \mathbf{0} = (0, 0, 0, 0, 0, 0, 0)$, corresponding to endorsement of the "None of the time" response category for all six Likert items and a count of 0 on the last item. The maximum class, represented by response pattern $\boldsymbol{U} = \mathbf{30} = (4, 4, 4, 4, 4, 4, 30)$, corresponds to endorsement of the "All of the time" response category for the six Likert items and a count of 30 on the last item. Because responses to Likert-type items are not influenced by digit preference, it is only the last item in the response pattern, the count item, that determines whether a response pattern $\boldsymbol{u}$ is in $\boldsymbol{e}$ or $\boldsymbol{k}$. If a non-**0** and non-**30** response pattern ends in a count that is not a multiple of five, it belongs to $\boldsymbol{e}$; if the response pattern ends in a multiple of five, it belongs to $\boldsymbol{k}$.

Graded Response IRT Model For the exact and rounding/selected response classes, the conditional probability of response $U_j = z$ for each of the first six Likert-type items can be viewed as a trace line for the graded response model (GRM; Samejima, 1969). For ordered responses $U_j = z$, $z \in \{0, 1, 2, 3, 4\}$, the conditional probability of endorsing response category z is expressed

$$T(U_j = z|\theta) = \frac{1}{1 + \exp[-a_j(\theta - b_{jz})]} - \frac{1}{1 + \exp[-a_j(\theta - b_{j(z+1)})]}, \tag{37}$$

in which a_j is item discrimination and b_{jz} is the threshold for category z. The value of b_{jz} is the value of θ at which the respondent has a 50% probability of endorsing category z or higher. For members of the exact count and rounding/selected response classes, responses to the first six items in Figure 2 in Chapter 1 can be modeled using the GRM. For computational purposes, I reparameterized the model into slope-intercept form, with $\text{logit}[T(U_j \geq z|\theta)] = -a_j\theta + c_{jz}$.

Count Response IRT Model As was true of the model described in Primary Aim #1, the specific IRT model used for the count item depends on latent class membership. For members of the exact count class, the conditional probability of count response $U_j = u_j$ can be modeled using a Poisson (Equation 20) or negative binomial (Equation 23) IRT model; for members of the rounding/selected response class, it can be modeled using the NRM (Equation 24). Thus, the set of trace lines comprising the joint probability of the seven item responses depends on whether the individual is a member of the exact count class or the rounding/selected response class.

Simulation Before fitting the proposed latent class IRT models to the BRFSS data, I carried out a small simulation to verify model identification and determine whether the implementation in R of parameter estimation for the model is able to recover known population parameters. I set the proportions of people belonging to the zero, exact count, rounding/selected response, and maximum classes to 0.25, 0.50, 0.24, and 0.01, respectively. To examine whether the count responses exhibit overdispersion, I tested two different latent class IRT models: one that uses a Poisson IRT model for the exact count class, and one that uses a negative binomial IRT model for the exact count class. Simulation parameters for each model are in Table 3. Sample R code for the Secondary Aim simulations can be found in Appendix B.

Empirical Analysis of the BRFSS After fitting the proposed latent class IRT models to the simulated data, I fit both models to the mixed item-type scale from the BRFSS. Nearly 100,000 individuals responded to the seven items comprising this scale; to make the analyses of the Secondary Aim more comparable with those of the Primary Aims, I took a simple random sample of 10,000 to use as an analytic sample. After fitting the latent class IRT model to the

Table 3. Simulation parameters for the full latent class IRT model, $N = 10,000$

	Item #	Item #2	Item #3	Item #4	Item #5	Item #6	Item #7
Graded Response							
IRT Parameters							
a	1.75	2.65	1.50	2.30	2.20	3.20	–
c_1	1.00	0.50	1.50	-1.00	1.50	2.00	–
c_2	0.25	-0.50	0.20	-2.00	0.50	0.75	–
c_3	-2.00	-0.80	-1.20	-3.00	-0.40	0.10	–
c_4	-4.00	-1.5	-3.40	-4.00	-2.00	-0.75	–
Nominal Response							
IRT Paramaters							
a_1	–	–	–	–	–	–	0.00
a_2	–	–	–	–	–	–	1.00
a_3	–	–	–	–	–	–	1.00
a_4	–	–	–	–	–	–	1.50
a_5	–	–	–	–	–	–	1.50
a_6	–	–	–	–	–	–	2.00
a_7	–	–	–	–	–	–	1.25
c_1	–	–	–	–	–	–	0.00
c_2	–	–	–	–	–	–	-0.30
c_3	–	–	–	–	–	–	-0.75
c_4	–	–	–	–	–	–	-0.50
c_5	–	–	–	–	–	–	-0.25
c_6	–	–	–	–	–	–	-0.80
c_7	–	–	–	–	–	–	-1.00
Count Model							
IRT Parameters							
a	–	–	–	–	–	–	1.15
c	–	–	–	–	–	–	1.10
δ							
(Neg. Bin. only)	–	–	–	–	–	–	0.40

Latent Classes	Zero	Exact Count	Rounding	Maximum
Proportion	0.25	0.50	0.24	0.01

empirical data, I compared model fit statistics to examine whether the Poisson IRT model or negative binomial IRT model is better supported by the data.

Estimation Parameter estimation for all models was carried out using R's **nlm** optimizer, described earlier in this chapter. When the Poisson IRT model was used to describe the exact count class, a total of 47 parameters were estimated: 30 GRM IRT parameters, 12 NRM IRT parameters, 2 Poisson IRT model parameters, and 3 latent class parameters. When the negative binomial IRT model was used for the exact count class, a total of 48 parameters were estimated: 30 GRM IRT parameters, 12 NRM IRT parameters, 3 negative binomial IRT model parameters, and 3 latent class parameters.

What Value is Added by the Count Item? The other goal of the Secondary Aim was to examine the contribution of the count item to the mixed item-type scale that is shown in Figure 2: What does the single count item contribute to measurement precision above and beyond the six Likert-type items? One method of evaluating the degree to which an item contributes to the measurement precision of a scale is to compare the posterior standard deviations of scale scores with and without that item. The posterior standard deviation quantifies the precision with which a latent variable θ is measured across its range (Thissen & Wainer, 2001; Ostini & Nering, 2006). Because the level of precision depends on the level of the latent variable, the posterior standard deviation varies with θ and is often depicted graphically, with $\hat{\theta}_{\text{EAP}}$ along the x-axis and $SD(\hat{\theta}_{\text{EAP}})$ along the y-axis. To assess the informative value of the count item, I computed the posterior standard deviation twice for the same individual – once for the scale score that includes the count item, and once for the scale score that excludes the count item. Plots of the posterior standard deviations as a function of $\hat{\theta}_{\text{EAP}}$ can then be compared to assess the improvement in measurement precision that is attributable to the count item.

I computed IRT scale scores ($\hat{\theta}_{\text{EAP}}$: Equation 29) and posterior standard deviations ($SD(\hat{\theta}_{\text{EAP}})$: Equation 30) for all response patterns observed in the sample, after removing the proportions of $\boldsymbol{U} = \mathbf{0} = (0, 0, 0, 0, 0, 0, 0)$ and $\boldsymbol{U} = (4, 4, 4, 4, 4, 4, 30)$ response patterns that were estimated to belong to members of the zero and maximum classes, respectively. As true of the model described in Primary Aim #1, individuals with a multiple-of-five response for the count item could belong to either the exact count or rounding/selected response class; thus, I computed

two possible scale scores and posterior standard deviations for these response patterns – one resulting from the GRM and Poisson trace lines, the other resulting from the GRM and NRM trace lines. I then plotted the posterior standard deviations as a function of the scale scores to depict how measurement precision varies across levels of the latent variable that the items measure. To assess the specific contribution of the count item, I compared the measurement precision for scale scores computed with and without the count item.

Chapter 3: Results

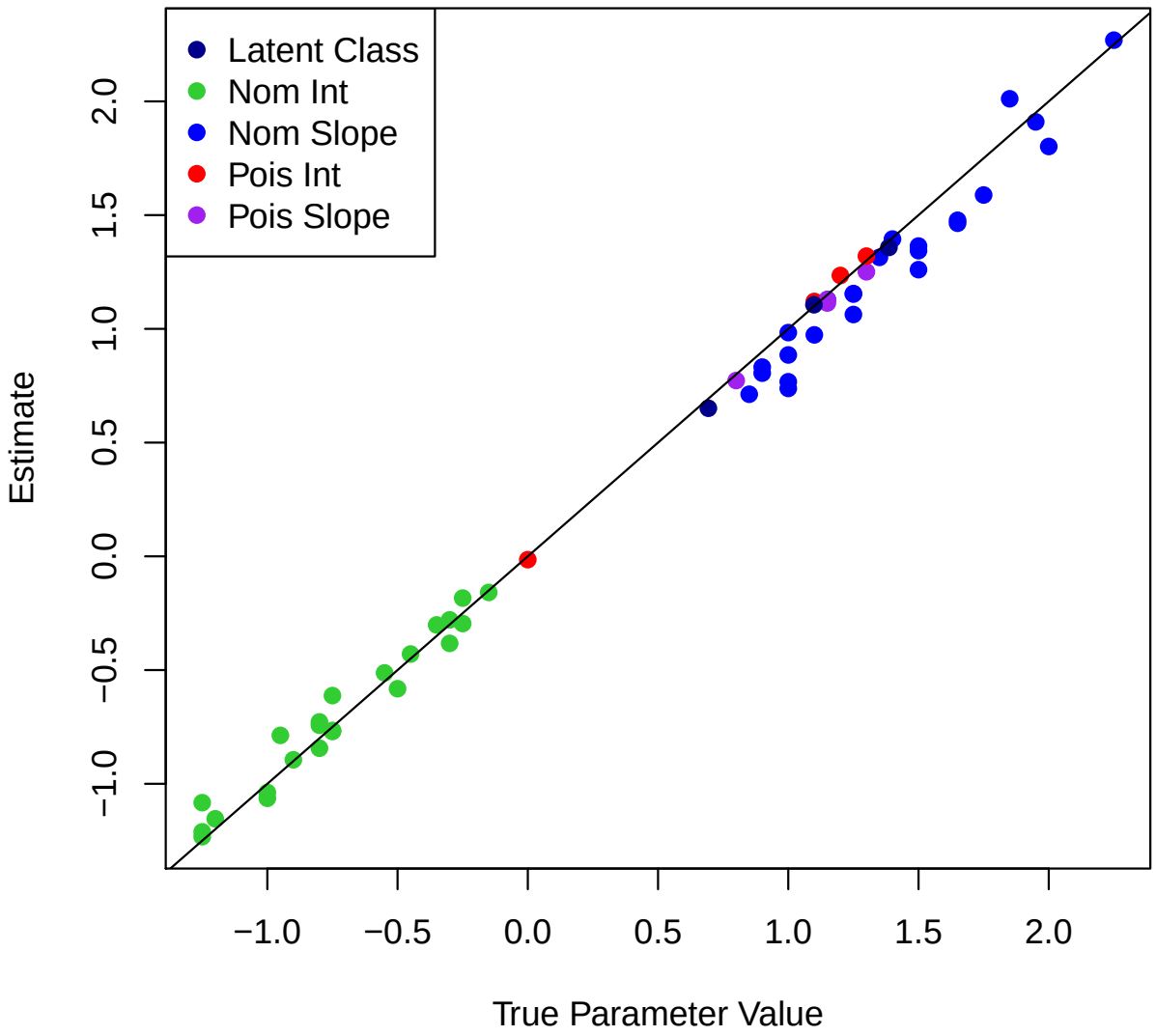
1. Primary Aim #1: Addressing Inflation and Heaping in Multivariate Count Data

1.1 Simulation

The Latent Class IRT Model with a Poisson IRT Model for the Exact Count Class When fit to the data simulated from the parameters in Table 2 in Chapter 2, the model converged after 266 iterations, requiring approximately 12.5 hours on a desktop computer with a quad-core 2.4GHz Intel Core processor and 4GB of RAM. The parameter estimates are plotted against the data-generating parameters in Figure 4. 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. Figure 4 shows reasonable parameter recovery, with all points hovering around the identity line. Deviations are greatest for the slope parameters of the NRM. This is unsurprising, given that the NRM is the most highly parameterized component of the latent class IRT model and typically requires larger sample sizes to obtain accurate estimates. The remaining fluctuation around the identity line is likely due to sampling error. The results of the simulation provide evidence that the full latent class IRT model with a Poisson component for the exact count class is identified and that parameter estimation as implemented in R is successful.

The Latent Class IRT Model with a Negative Binomial IRT Model for the Exact Count Class In practice, count data tend to exhibit more variability than accounted for by a Poisson distribution; in such cases, the negative binomial distribution is often a more realistic description of the data, because it allows for overdispersion – that is, the variance of the distribution can be greater than its mean. For this reason, I tested an alternative latent class IRT model that uses a negative binomial model to describe the item responses of members from the exact count class. This model is identical to the Poisson IRT model, with the exception that the negative binomial IRT model includes an additional parameter for each item to characterize overdispersion. When fit to the data generated from the parameters in Table 2, the model

Figure 4. Latent class IRT model with a Poisson IRT model for exact count class and a nominal response IRT model for rounding/selected response class: Estimated parameters vs. data-generating parameters, for data simulated using the parameters in Table 2.



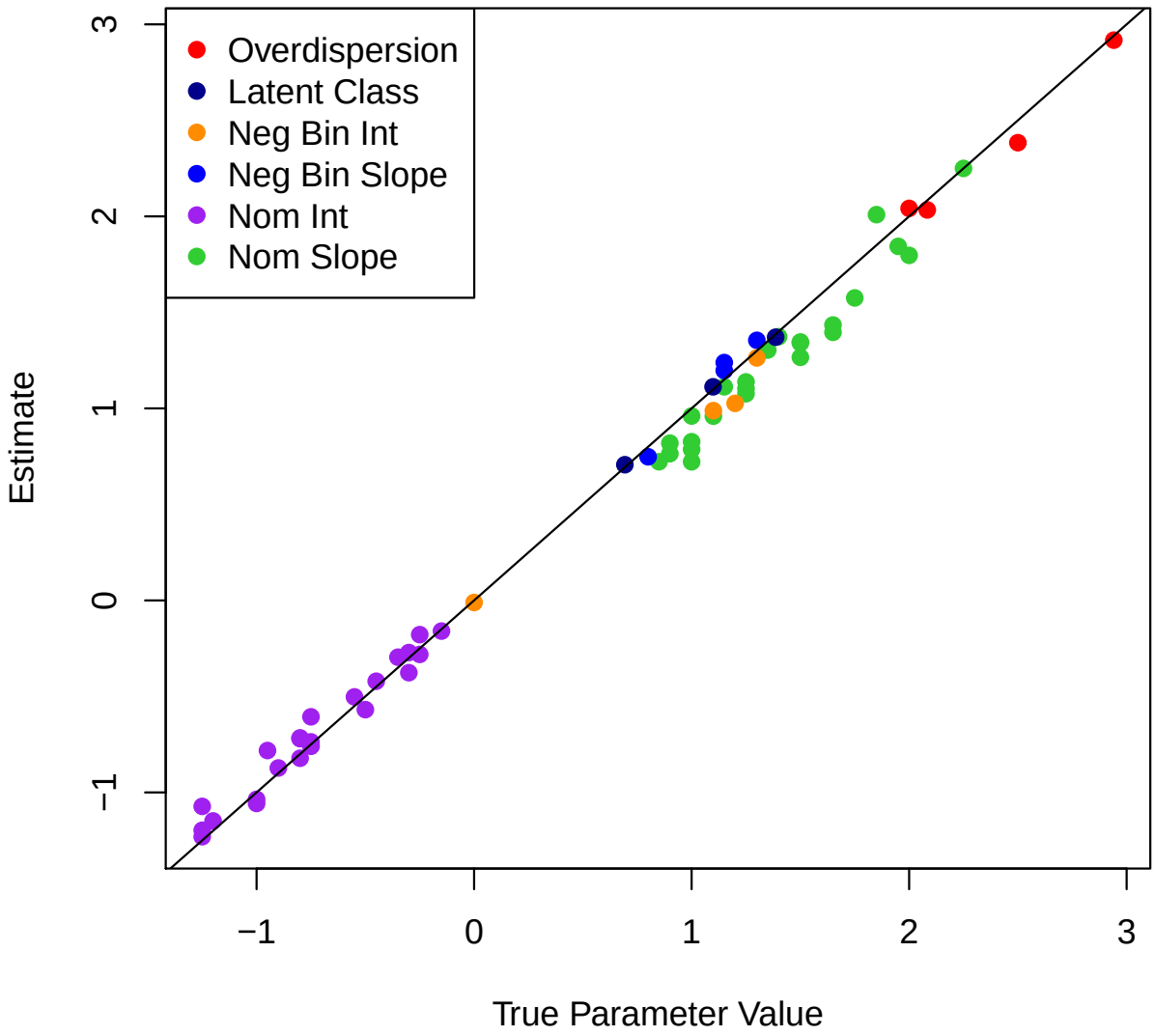
converged after 285 iterations, requiring approximately 13 hours on a desktop computer with a quad-core 2.4GHz Intel Core processor and 4GB of RAM. The results of the simulation test of software implementation of the estimation procedure are shown in Figure 5. Like its Poisson counterpart, Figure 5 shows that the R program recovers all parameters reasonably well, including the four overdispersion parameters that are unique to the negative binomial IRT model. As was true of the latent class model with the Poisson component, the largest deviations between the model estimates and data-generating parameters correspond to the slopes for the nominal response component of the model.

1.2 Empirical Analysis of BRFSS Data

Latent Class IRT Model with Poisson IRT Model for Exact Count Class The results of both simulations provide evidence that the latent class IRT model with a Poisson component for the exact count class, as well as the latent class IRT model with a negative binomial component for the exact count class, are identified, and that the R program produces reasonably accurate estimates of model parameters. After testing the model with simulated data, I fit each of the latent class IRT models to the motivating data set: the Pain, Depressed, Anxious, and (reversed) Energy items from the BRFSS. First, I tried fitting the most parsimonious model, with a Poisson component for the exact count class. After only eight iterations, however, the R program terminated because the gradient of the log likelihood approached infinity, suggesting a model misspecification. To identify the problem, I examined the empirical response distributions of the four items and found that the Poisson distribution overpredicts the frequency of counts at the lower end of the scale, and conversely, underpredicts the frequency of counts at the higher end of the scale. Such a discrepancy between observed and predicted counts suggests overdispersion in the data. Therefore, an alternative model that could account for the excess variability in the empirical data was needed.

The Latent Class IRT Model with a Negative Binomial IRT Model for the Exact Count Class After concluding that the four count items from the BRFSS exhibit greater variability than what can be explained with a Poisson distribution, I respecified the model by replacing the Poisson IRT model with a negative binomial IRT model to describe the response process for members of the exact count class. I then fit this modified latent class IRT model to the BRFSS data. Unlike the latent class model with a Poisson IRT component, the parameters

Figure 5. Latent class IRT model with a negative binomial IRT model for exact count class and a nominal response IRT model for rounding/selected response class: Estimated parameters vs. data-generating parameters, for data simulated using the parameters in Table 2.



of the alternative model with a negative binomial component for the exact count class were estimated successfully. The model converged after 472 iterations and approximately 16.5 hours.

Estimates of the proportion of people belonging to each of the four latent classes can be found at the bottom of Table 4. The proportion of respondents estimated to belong to the zero class, exact count class, rounding/selected response class, and maximum class are 0.16, 0.52, 0.31, and 0.01, respectively, suggesting that 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 five. Only about half of respondents utilized the full range of the open-ended count scale in selecting non-multiple-of-five counts. While 24% of the sample endorsed zero for every item, $\boldsymbol{U} = \mathbf{0} = (0, 0, 0, 0)$, only 16% of respondents were estimated to belong to the zero class. This decomposition of all-0 response patterns indicates that someone endorsing zero for all four items has approximately 67% probability of belonging to the zero class and 33% probability of belonging to the exact count or rounding/selected response class. Approximately 68% of the 138 people who endorsed 30 for every item, $\boldsymbol{U} = \mathbf{30} = (30, 30, 30, 30)$, 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 that the scale likely measures: 'Poor Emotional Health'.

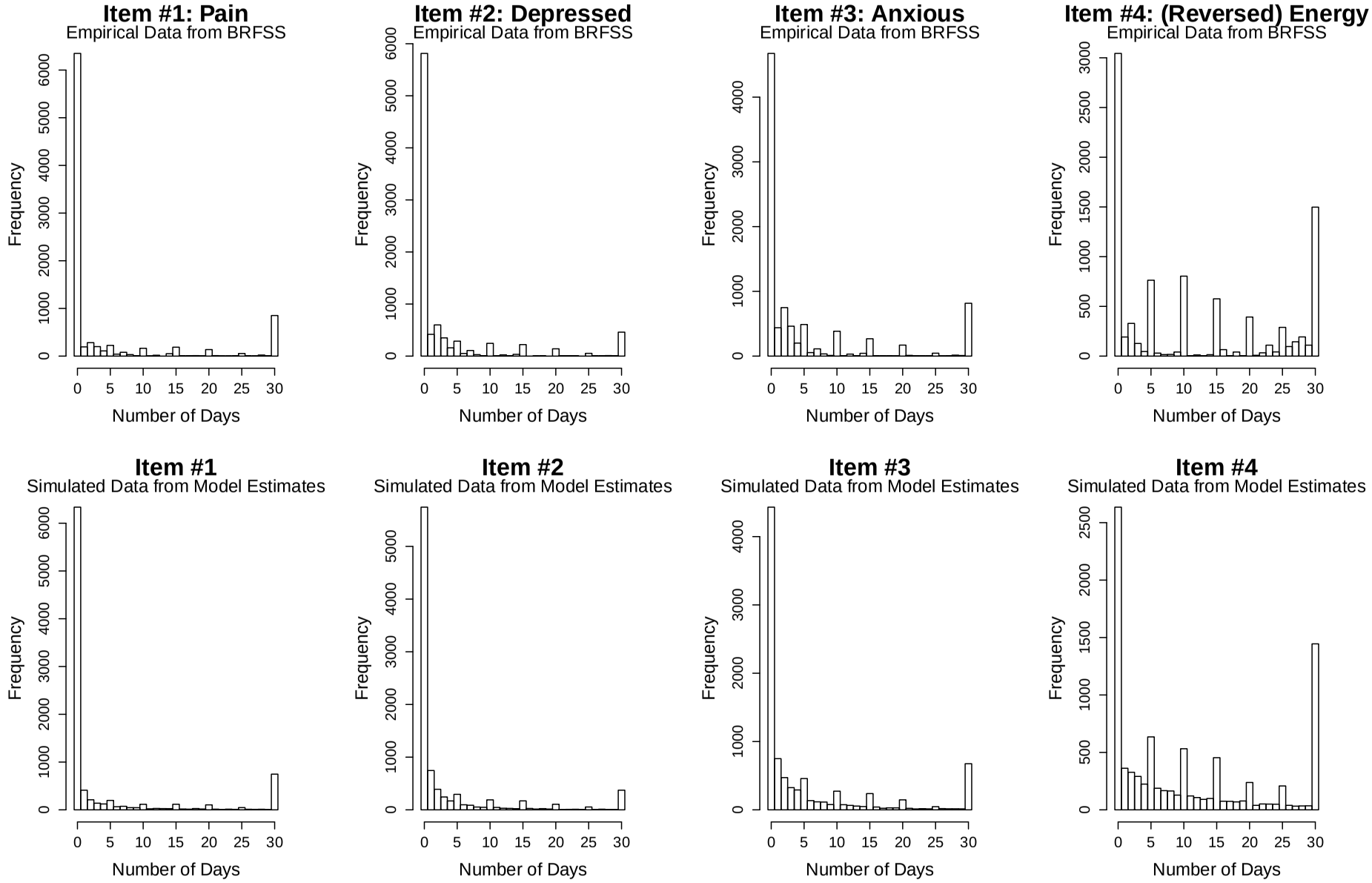
IRT parameter estimates for the four BRFSS count items can also be found in Table 4. To examine how closely the IRT parameters estimated from the model could reproduce the empirical response distributions, I simulated 9,042 responses to each of four items from the parameter estimates in Table 4. The upper panel of Figure 6 shows the empirical response distributions for the BRFSS data; the lower panel shows the data that were simulated from the parameter estimates. While comparison of the upper and lower panels does not allow a thorough analysis of model fit at the response pattern level, comparing the empirical and simulated response distributions can help inform whether the specific IRT models are appropriate for the data – specifically, whether negative binomial and nominal response IRT models accurately describe the shape of the response distributions. The empirical response distributions for Pain, Anxious, and Depressed – shown in the first three columns of Figure 6 – are reproduced fairly well, suggesting that the negative binomial and nominal response IRT models are appropriate IRT model choices for these three items; however, the empirical response distribution for (reversed) Energy,

which is shown in the rightmost column of Figure 6, is not nearly as well reproduced via simulation. Specifically, the negative binomial IRT model overpredicts the frequency of respondents reporting counts between 20 and 30 (reversed responses between 0 and 10) and underpredicts the frequency of respondents reporting counts between 0 and 10 (reversed responses between 20 and 30). This discrepancy between the observed and predicted counts occurs because the negative binomial distribution cannot account for the increasing number of people reporting counts toward the upper limit of the 0-30 count range. As the latent class IRT model is only intended to be an approximation of the true count distribution, however, I proceed with interpretation of the latent class IRT model that includes the negative binomial component for all four items. Recommendations for alternative model specifications that may better accommodate the Energy item are described in Chapter 4.

Table 4. Parameter estimates from the negative binomial latent class IRT model fit to the BRFSS data , $N = 10,000$

	# Days Pain	# Days Depressd	# Days Anxious	(30-) # Days Energy
Nominal Response IRT 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 (0.011)	-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 IRT 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)

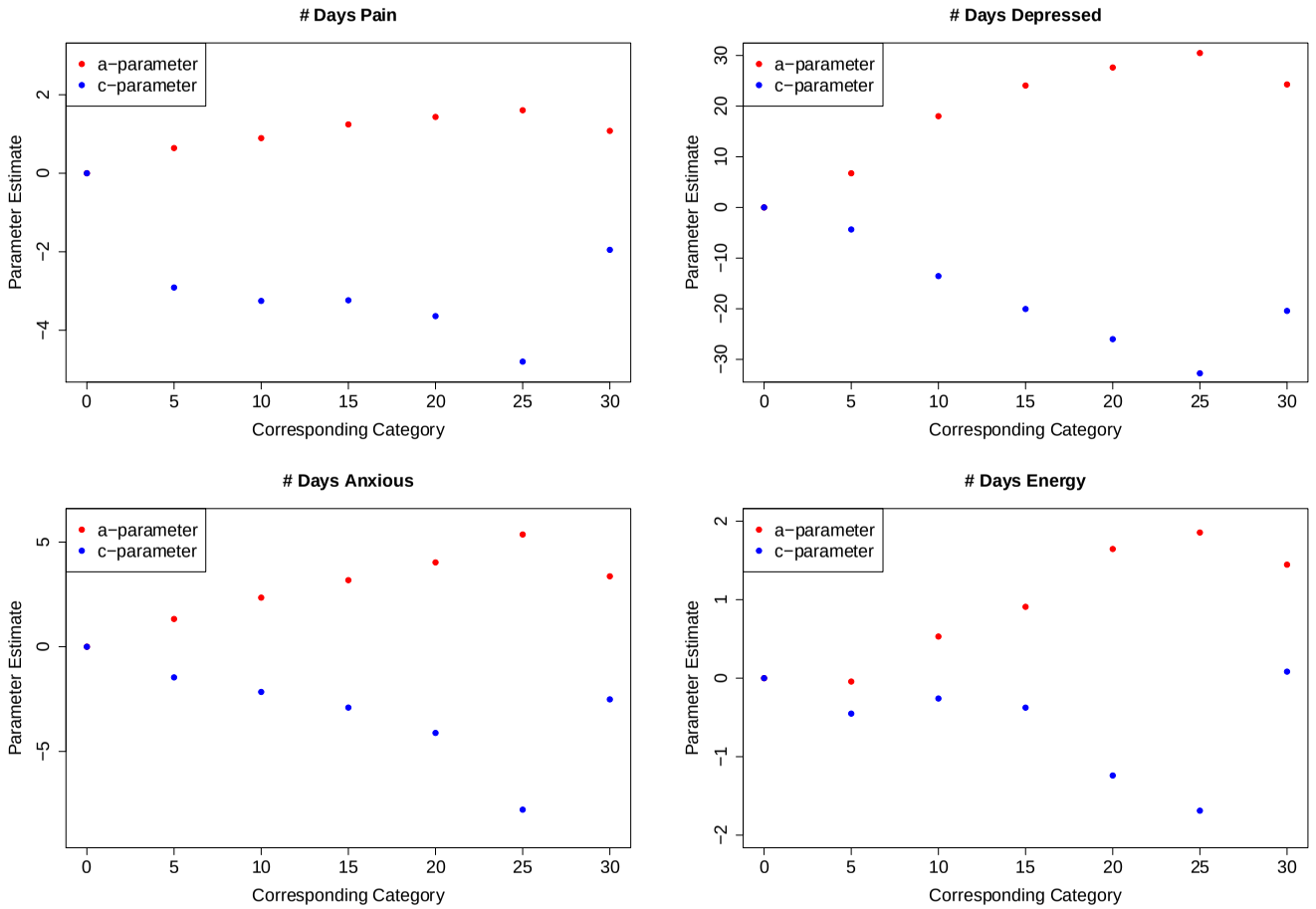
Figure 6. Empirical response distributions vs. response distributions simulated from the estimated model parameters in Table 4.



The Rounding/Selected Response Class Several features of the item parameters in Table 4 are noteworthy. Because the NRM was used to describe the response process for members of the rounding/selected response class, the ordering of response categories could be examined empirically. For this latent class, the NRM item parameters show a relatively linear trend: As the number of days category increases, the a-parameters tend to increase and the c-parameters tend to decrease nearly linearly, with the exception of the a- and c-parameters corresponding to the 30-day category (recall that a_1 and c_1 were fixed to 0 for all items). Figure 7 more clearly shows the relationship between the NRM item parameters and response categories; within each item, the a-parameters increase and the c-parameters decrease from the 5-day to 25-day response categories. For the 30-day response category, the a- and c-parameters reverse direction. While the magnitudes of the parameter estimates vary depending on the item, this pattern is consistent across all four items.

Within the rounding/selected response class, Anxious and Depressed are most discriminating based on their slope parameters, suggesting that these two 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 Figure 8, which depicts the NRM trace lines for the rounding/selected response class. The trace lines for Anxious and Depressed tend to function more similarly to each other than either of the other items, likely because they both measure aspects of mental health. The Poor Emotional Health latent variable is really defined by these two items, with Pain and Energy acting as ancillary items. Anxious and Depressed are also the two items that exhibit the least overdispersion.

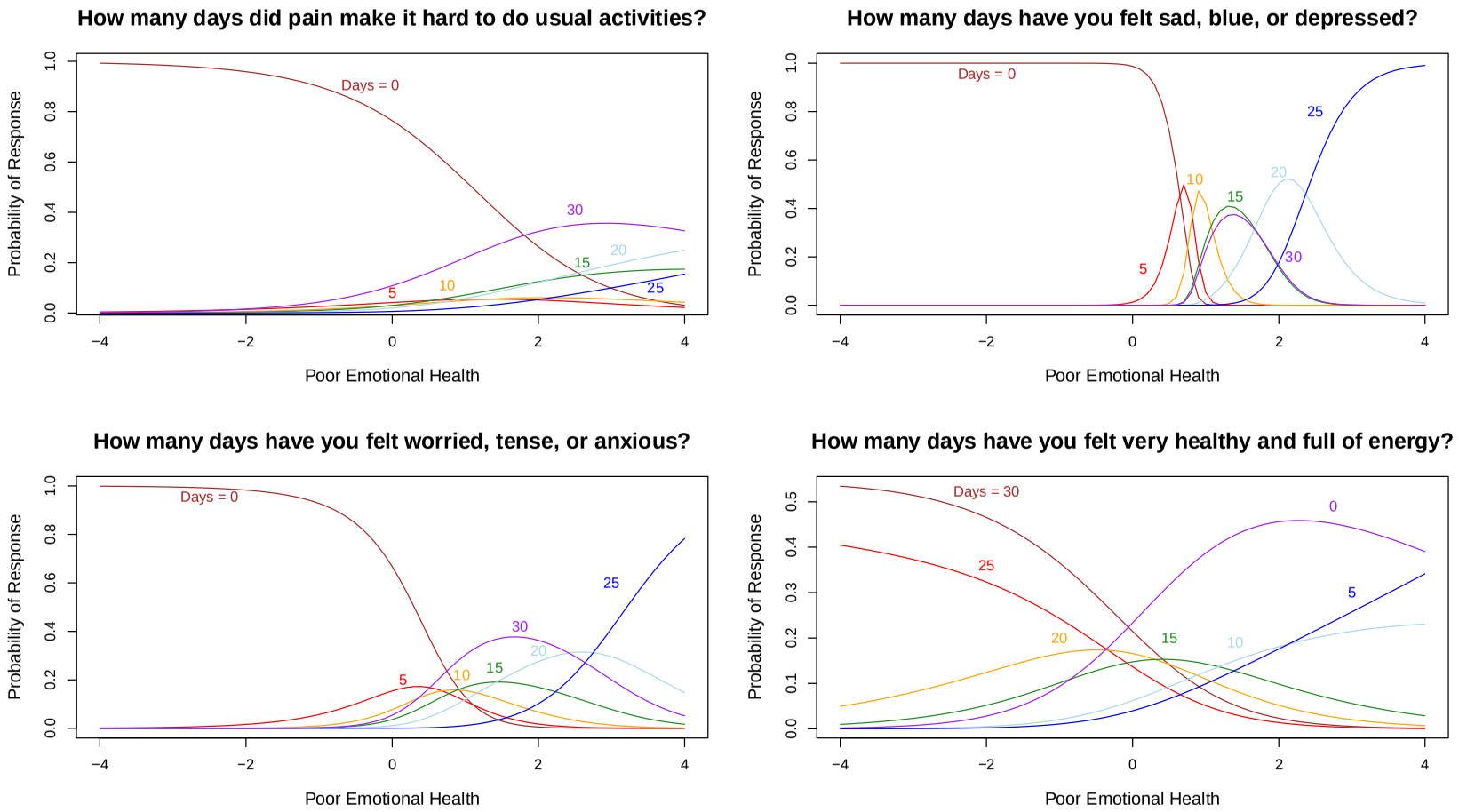
Figure 7. NRM item parameter estimates as a function of response category within the rounding/selected response class.



The trace lines in Figure 8 also reveal several other interesting characteristics of the items. First, the 0-days and 30-days response categories tend to be associated with the greatest probabilities of endorsement, regardless of someone's level of the latent variable: Across the entire continuum of Poor Emotional Health, 0-days or 30-days is almost always the most likely response category. This phenomenon is especially 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 8 shows that the response categories tend to be in increasing order for 5-days, 10-days, 15-days, 20-days, and 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 days (or, a lesser number of days for the Energy item). The increasing order does not hold for the 0-days and 30-days response options, however; this is most clearly seen in the trace lines for Depressed and Anxious, where for people at very high levels of Poor Emotional Health (i.e., approximately 2.5 or more standard deviations above average), 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 – in other words, these items are not very informative for relatively healthy people. This is seen in Figure 8, where for these three items, it is not until $\theta \geq 0$ that the trace lines begin to cross. The relatively flat trace lines for the Energy item suggest that this item is not very strongly related to the latent variable, as people tend to endorse a smaller number of days for this item regardless of their level of Poor Emotional Health. For example, even for someone at low levels of θ, 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, shown in Figure 8 at the location where the 0-days and 30-days trace lines cross. Compared to the other three items, the endorsement of lower counts is common.

Figure 8. NRM trace lines for the rounding/selected response class.



The Exact Count Class The negative binomial item parameter estimates for the exact count class are also shown in Table 4. As was seen in the trace lines for the rounding/selected response class, the exact count class parameters in Table 4 indicate that Anxious and Depressed are the most discriminating items, with a-parameters that are considerably larger than those for Pain and (reversed) Energy. Within the exact count latent class, the item discrimination for a count item can be interpreted as the log expected change in the number of days associated with a one standard deviation increase in Poor Emotional Health; this value can then be exponentiated to be placed on a more interpretable scale. For a one standard deviation 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; again, this value can be exponentiated for ease of interpretation. 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). All expected values and slopes have been adjusted for overdispersion, with Pain and Energy exhibiting the greatest variability in responses.

In fitting a count IRT model to item response data, there are multiple ways to graphically depict characteristics of the items (e.g., 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 Figure 9. Within a particular item, each curve corresponds to one 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 best discriminates among individuals. One can choose a level of the Poor Emotional Health latent variable and find the endorsement probability that corresponds to each of the 31 counts. For the Pain item, for example, someone with $\theta = 0$ has roughly a 0.60 probability of endorsing 0 days, a 0.1 probability of endorsing 1 day, a 0.05 probability of endorsing 2 days, and a near-zero probability of endorsing any of the counts greater than 3 days.

Figure 9. Negative binomial model trace lines for the exact count class.

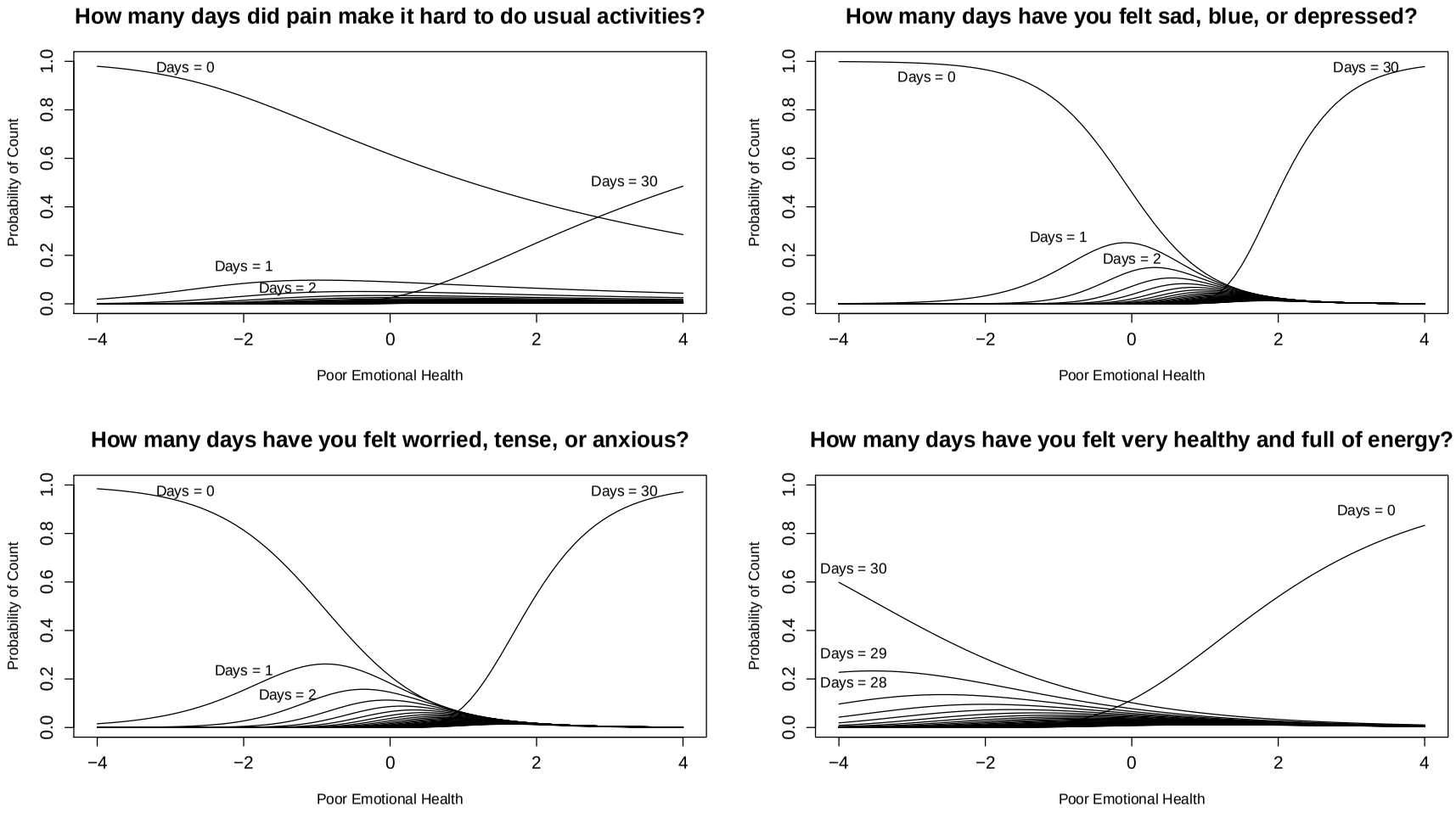


Figure 9 shows that the Anxious and Depressed items are considerably more discriminating than the Pain and Energy items. Overall, both Pain and Energy do a poor job of separating individuals on the Poor Emotional Health latent variable. Like the rounding/selected response class, the 0-days and 30-days 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 9 also shows that lower counts are more easily endorsed for the Energy item, even for individuals at low levels of Poor Emotional Health. For example, someone at low levels of Poor Emotional Health has only a modeled 40-60% probability of reporting 30 days for the Energy item (or equivalently, 0 days for the reverse coded Energy item); for the other three 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 three 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 two items appear less relevant.

The negative binomial trace lines also show a weak relationship between the Pain item and the Poor Emotional Health latent variable. Only counts of 0-days and 30-days have reasonably high probabilities of endorsement. The trace lines corresponding to the remaining 29 counts rise little above a probability of 0. Even for someone who is more than two standard deviations above average on Poor Emotional Health, the count associated with the highest probability is 0-days for Pain. One must be very high on Poor Emotional Health for 0 to not be the most likely response.

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. Because the negative binomial IRT model assumes that the probability of endorsing increasingly higher counts is a monotonically increasing function of the latent variable, one can plot the expected count λ_j as a function of Poor Emotional Health. Expected count plots are shown in Figure 10, with each color representing a different item – in this case, symptom. Figure 10 shows the same item characteristics as the trace lines in Figure 9, only now the y-axis represents expected counts – or, number of days – rather than probability. The low discriminating power of the Energy item

compared to the Depressed and Anxious items is shown with its flatter slope. As was seen in the negative binomial trace lines in Figure 9, the relatively high expected counts for (reversed) Energy – even for people with low levels of Poor Emotional Health – is shown in Figure 10. For someone who is two standard deviations 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.

Figure 10. Expected counts as a function of the latent variable for members of the exact count class.

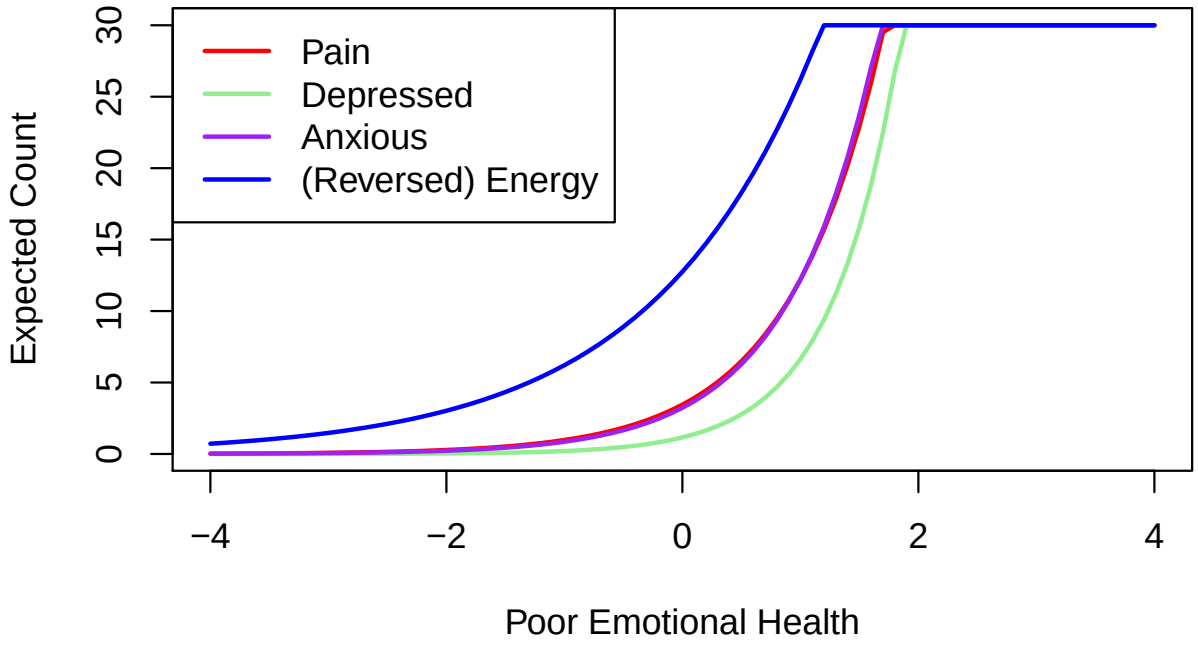


Figure 10 also reveals a rapid increase in the expected number of Pain, Depressed, and Anxious days for someone between one and two standard deviations above average on Poor Emotional Health. The sharp increase is especially noticeable for Depressed, where someone falling two standard deviations above average on Poor Emotional Health ($\theta = 2$) is expected to report feeling depressed for approximately 25 more days per month than someone falling only one standard deviation above average on Poor Emotional Health ($\theta = 1$). Once someone's

level of Poor Emotional Health approaches two standard deviations above average ($\theta = 2$), the expected number of days is 30 for all symptoms.

Scale Scores Because clinicians are often interested in using scale scores in subsequent statistical analyses, an additional goal was to compute IRT scale scores for individuals based on their responses to the four items on the BRFSS. In this case, scale scores depend not only on response patterns but latent class membership as well. One of the complexities involved in scoring people according to a latent class IRT model 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. Depending on an individual's response pattern, however, it may be possible to identify class membership; if a person responds with at least one count that is not a multiple of five, that person must belong to the exact count class. For all other response patterns, however, there exist multiple possible class memberships, and consequently, more than one plausible scale score.

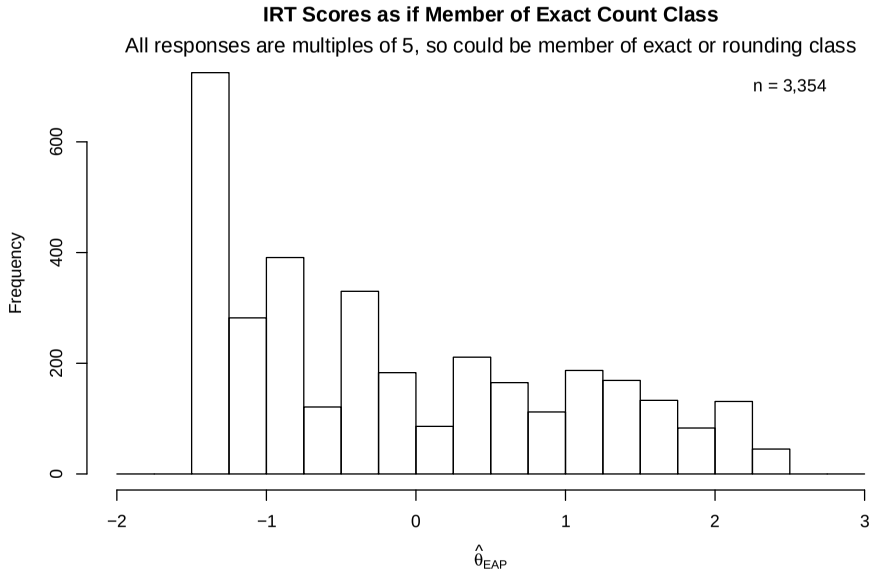
It is important to note that at the population level, only members of the exact count and rounding/selected response classes fall along the latent variable continuum – this means that only 83% of the population, or 7,498 of the 9,042 people in the sample, should receive scores on Poor Emotional Health. To account for the 16% and 1% of the population belonging to the zero and maximum classes, respectively, I removed 1,451 all-0 response patterns and 93 all-30 response patterns from the sample, 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, I discuss scores only at the population level and not at the individual level.

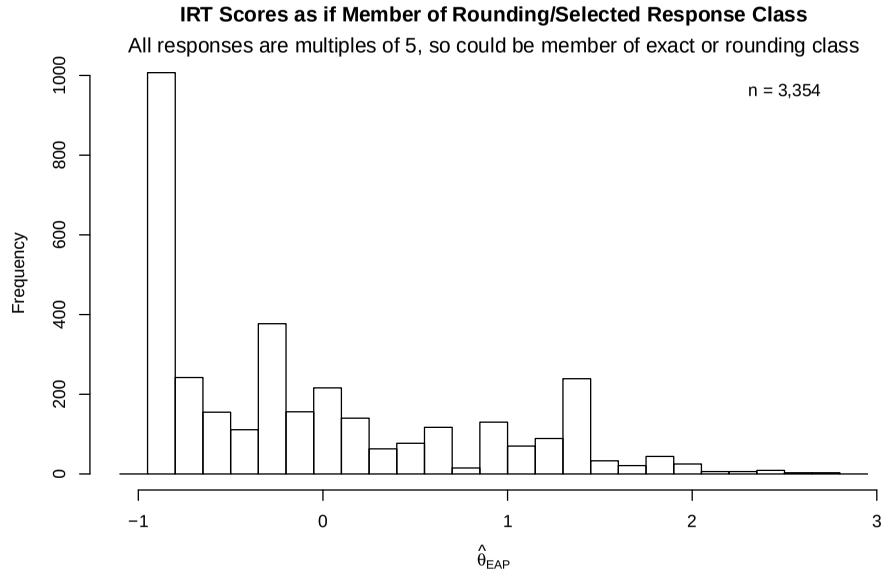
After removing 16% of the all-0 response patterns and 1% of the all-30 response patterns, I computed scores for the 7,498 people who remained in the BRFSS sample. For the 55% of the scoring sample with response patterns including at least one non-multiple-of-five count (e.g., $\boldsymbol{U} = [0, 1, 0, 2]$, $\boldsymbol{U} = [5, 0, 4, 0]$), scoring was rather straightforward; because these individuals must belong to the exact count class, I used the set of negative binomial trace lines to compute their scale scores and posterior standard deviations according to Equations 29 and 30 in Chapter 2.

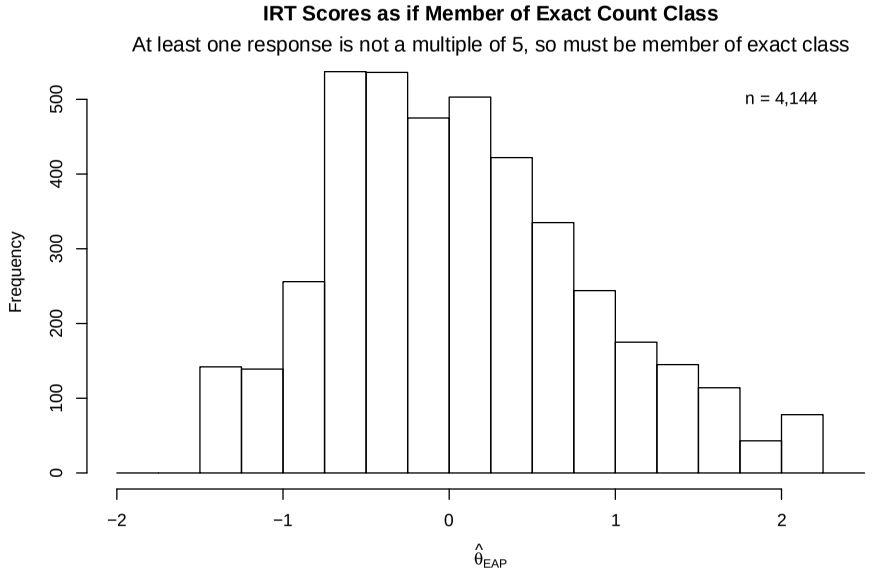
A histogram of the scale scores for people who must belong to the exact count class is shown in the bottom left panel of Figure 11. These scale scores, interpreted on a z-score metric, are roughly normally distributed with a mean of 0.07 and variance of 0.63, although there is a slight positive skew that is representative of the generally healthy population. Scale scores range from $\hat{\theta}_{EAP} = -1.34$, which corresponds to response pattern $\boldsymbol{U} = (0, 0, 0, 1)$, to $\hat{\theta}_{EAP} = 2.24$, which corresponds to response pattern $\boldsymbol{U} = (30, 30, 30, 29)$. Note that because these scores represent only those individuals who must be members of the exact count class, the histogram does not include response patterns that comprise only multiples of five, including $\boldsymbol{U} = \mathbf{0} = (0, 0, 0, 0)$ and $\boldsymbol{U} = \mathbf{30} = (30, 30, 30, 30)$.

The remaining 45% of the scoring sample exhibited response patterns that included only multiples of five; because this type of response pattern can manifest from either an exact count or rounding/selected response process, two different scores are plausible for individuals with this type of response pattern. I computed one score as if the person belonged to the exact count class by taking the product of the negative biniomial trace lines to obtain the posterior distribution. I computed the second score as if the person were a member of the rounding/selected response class by taking the product of the NRM trace lines to obtain the posterior distribution. Score histograms corresponding to each of the two plausible scoring methods are shown in the upper panel of Figure 11.

Figure 11. IRT scale scores (response pattern EAPs) for members of the exact count and rounding/selected response classes.







Unlike the histogram of scores for individuals who must be members of the exact count class, which is shown in the lower left corner of Figure 11, the score histograms of individuals who could belong to either the exact count or rounding/selected response class, shown in the upper panel of Figure 11, exhibit peakedness; the largest peak corresponds to the response pattern $\boldsymbol{U} = \mathbf{0} = (0, 0, 0, 0)$ that is commonly observed among respondents. Importantly, even though the 16% of people estimated to belong to the zero class was removed from the scoring sample, there are still many people with all-0 response patterns who fall along the Poor Emotional Health latent variable continuum; thus, some of the all-0 response patterns have scores associated with Poor Emotional Health, and these scores are represented by the highest peak in these two histograms. For a member of the exact count class, the scale score associated with a response pattern of $\boldsymbol{U} = \mathbf{0} = (0, 0, 0, 0)$ is $\hat{\theta}_{EAP} = -1.40$; for a member of the rounding/selected response class, the scale score associated with this same all-0 response pattern is $\hat{\theta}_{EAP} = -0.91$. For a member of the exact count class, the scale score associated with a response pattern of $\boldsymbol{U} = \mathbf{30} = (30, 30, 30, 30)$ is $\hat{\theta}_{EAP} = 2.38$; a member of the rounding/selected response class with the same all-30 response pattern has an estimated scale score of $\hat{\theta}_{EAP} = 1.41$. The discrepancy between the associated scale scores is due to the non-linear orderling of response categories within the NRM. Consequently, response patterns $\boldsymbol{U} = \mathbf{0} = (0, 0, 0, 0)$ and $\boldsymbol{U} = \mathbf{30} = (30, 30, 30, 30)$ are not representative of the most extreme levels of Poor Emotional Health. For the rounding/selected response class, it is actually a response pattern of $\boldsymbol{U} = (0, 0, 0, 5)$ that is associated with the lowest scale score, $\hat{\theta}_{EAP} = -0.93$, and a response pattern of $\boldsymbol{U} = (30, 25, 25, 25)$ that is associated with the highest scale score, $\hat{\theta}_{EAP} = 2.75$.

One question that may be of interest to clinicians is the extent to which it matters what type of score is computed for people who could belong to either the exact count or rounding/selected response class: How related are the two types of scores? The correlation between the two sets of scores that are shown in the upper panel of Figure 11 is 0.95 with a standard error of 0.01; Figure 12 shows this relationship in scatterplot form. This scatterplot reveals a slightly funnel-shaped relationship between the two types of scale scores, in which scores are not as highly correlated at the extreme positive end of Poor Emotional Health. Table 5 further highlights this trend. The scale score estimates that correspond to response patterns with higher counts, such as $\boldsymbol{U} = (25, 25, 25, 25)$ and $\boldsymbol{U} = (30, 30, 30, 30)$, show larger discrepancies between the

exact count and rounding/selected response scoring methods than the scale scores representing response patterns with lower counts, such as $\boldsymbol{U} = (5, 5, 5, 5)$ and $\boldsymbol{U} = (10, 10, 10, 10)$. This is because according to the negative binomial IRT model for the exact count class, expected counts are a monotonically increasing function of Poor Emotional Health, whereas according to the nominal response IRT model for the rounding/selected response class, expected counts do not always increase with higher levels of Poor Emotional Health. A side effect of this non-monotonic 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 one standard deviation above average.

Table 5. Expected scale scores and posterior standard deviations for different response patterns: The exact count class vs. the rounding/selected response class.

	Exact Count	Selected Response
Response Pattern $\boldsymbol{U}$	$\hat{\theta}_{\text{EAP}}$ (SE)	$\hat{\theta}_{\text{EAP}}$ (SE)
(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)

Posterior standard deviations (i.e., the standard errors of the scale scores) are plotted in Figure 13. 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. As tends to be true of IRT scores, the posterior standard deviations are larger at the extreme ends of Poor Emotional Health

Figure 12. Scatterplot of scale scores computed for the same persons from negative bimomial model trace lines (x-axis) vs. NRM trace lines (y-axis).

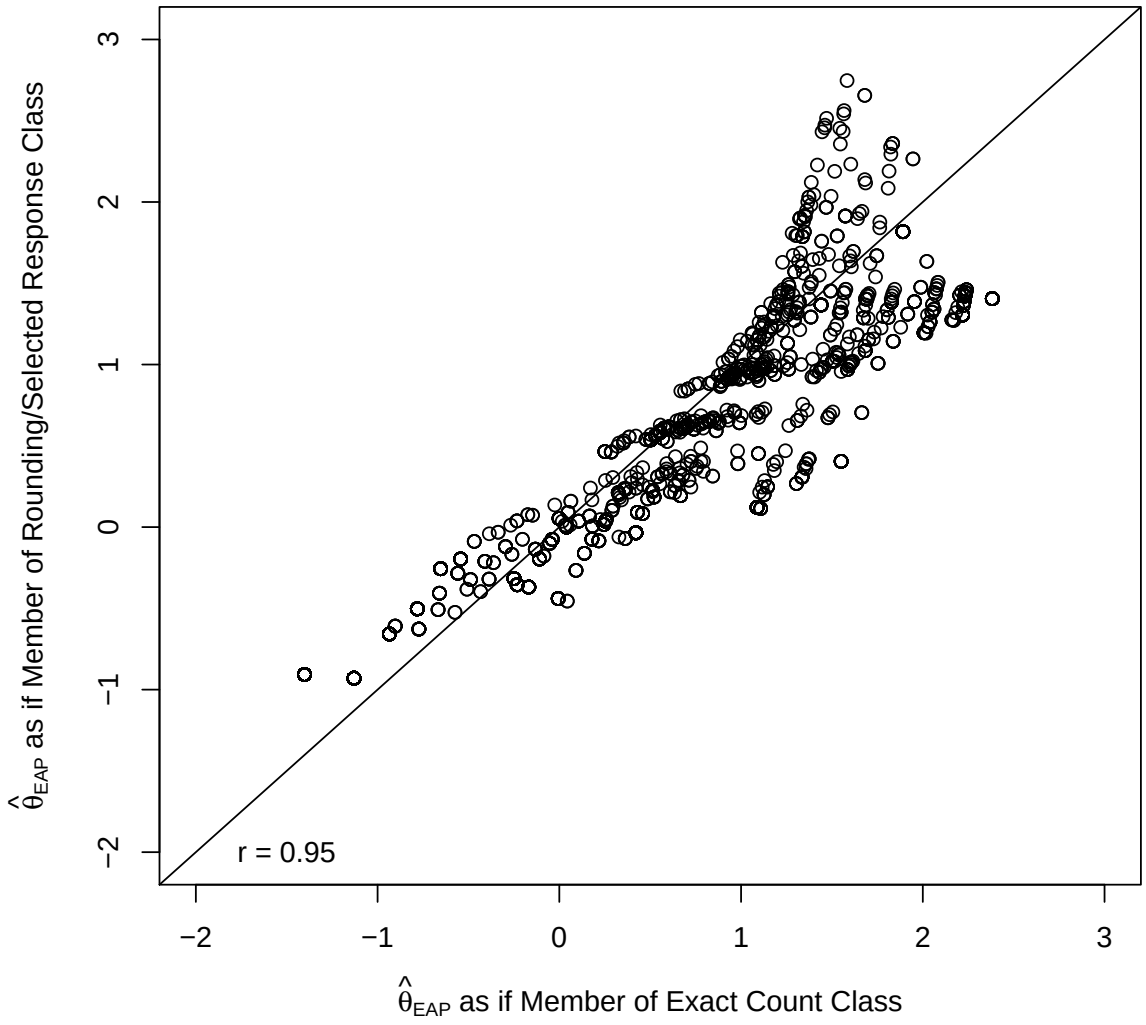
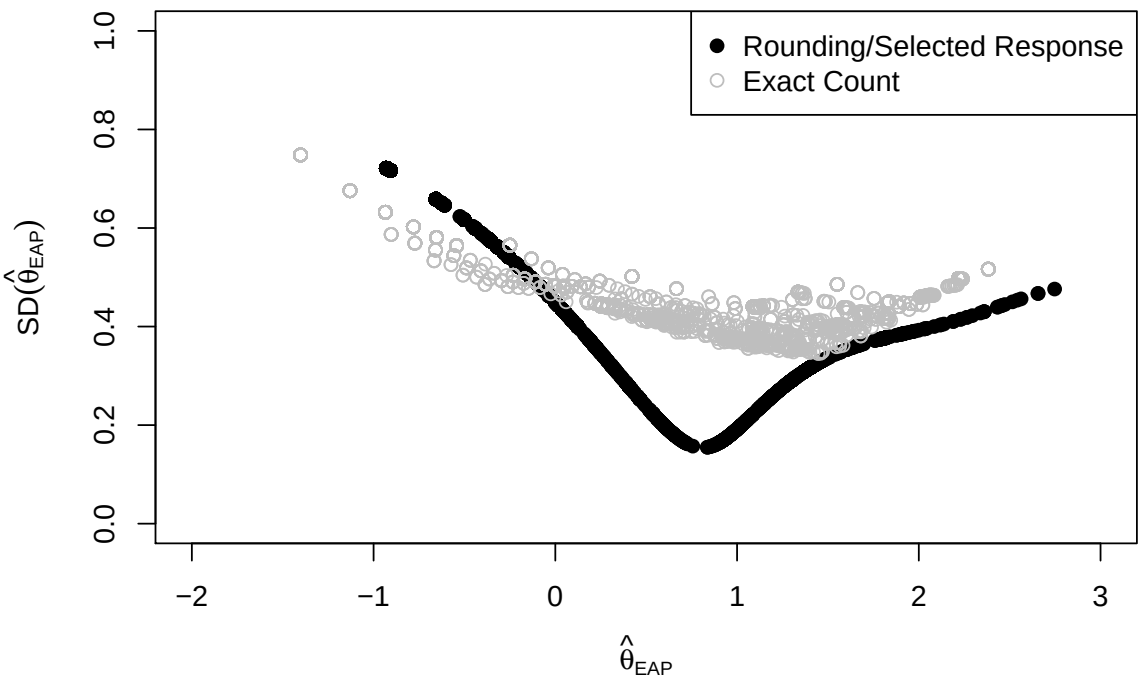


Figure 13. Posterior standard deviations (SEs) as function of scale scores.



and smaller near values of Poor Emotional Health where the items are most discriminating. For the exact count class, posterior standard deviations are the smallest for Poor Emotional Health scores that are approximately 1.5 standard deviations above average; for the rounding/selected response class, posterior standard deviations are the smallest for Poor Emotional Health scores that are just under one standard deviation above average.

It is worth noting that the posterior standard deviations 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. This indicates 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; however, when counts are higher, Poor Emotional Health is measured with greater precision in

the rounding/selected response class.

2. Primary Aim #2: Are Complex Models Really Needed?

The results thus far suggest that a latent class model with a negative binomial IRT model for an exact count class, a nominal response IRT model for a rounding/selected response class, and degenerate IRT models for zero and maximum classes can help account for zero-inflation, maximum inflation, and heaping in multivariate count data. However, a model with four latent classes and two IRT models is complex, involves estimation of a large number of parameters, and requires several hours of computing time. Therefore, an additional goal of this reseach was to ascertain whether such complex modeling techniques are needed, or whether more parsimonious models without all four latent classes are sufficient.

To address this question, I proposed three alternative models: a three-class model that excludes the zero class, a three-class model that excludes the rounding/selected response class, and a three-class model that excludes the maximum class. After fitting each of these alternative models to the BRFSS data, I used the value of the model log likelihood to compute the AIC and BIC fit statistics, as shown in Equations 34 and 35 in Chapter 2. Table 6 contains the AIC and BIC values corresponding to each of the three competing latent class IRT models, with the smallest values, representative of the best-fitting model, shown in bold. Both the AIC and BIC support the four class model, providing evidence that all four classes are needed to describe these item response data. Because a model without the rounding/selected response class does not require estimation of 48 nominal response parameters, it is a much more parsimonious model than either of the models that include a rounding/selected response component; however, model fit substantially worsens, suggesting that a count IRT model alone does not sufficiently capture the response process people use in answering these four count items. Further, even though the proportion of people estimated to belong to the maximum class is small (1%), the AIC and BIC both suggest that this class is needed: At the cost of only one additional parameter, the AIC and BIC values drop considerably.

Overall, these results indicate that, while more parsimonious, IRT models that do not take into account heaping and inflation omit important characteristics of the response mechanisms that people likely use in responding to these four count items. A count IRT model alone cannot sufficiently describe the way in which respondents interact with retrospectively reported count

Table 6. AIC and BIC values for competing latent class IRT models; the best-fitting values are shown in bold.

	All 4 Classes	No Zero Class	No Maximum Class	No Rounding Class
# Parameters	63	62	62	14
AIC	**117955.6**	124055.8	124125.7	132420.4
BIC	**118403.5**	124496.6	124566.2	132519.9

items on the BRFSS; some type of mixture component that can also account for a different type of response style is necessary – in this case, a NRM for people who round their answers to multiples of five or treat the items as multiple choice questions.

3. Secondary Aim: What Value Do Count Items Add to Scales?

The Secondary Aim of this research was to examine the contribution of a single count item to a scale from the BRFSS comprising six Likert-type items. The items on this scale are shown in Figure 2 in Chapter 1. To address this question, I proposed a modified latent class IRT model, in which the six Likert items were fit with a graded response IRT model and the count item was fit with either a count or nominal response IRT model, depending on the latent class. For members of the exact count class, a Poisson or negative binomial IRT model was used; for members of the rounding/selected response class, a nominal response IRT model was used. As was true of the model used to analyze the BRFSS scale comprising four count items, the proportions of respondents in the zero, maximum, exact count, and rounding/selected response classes were estimated as part of the model. Unlike the previous model, however, it was only a single count item that determined the proportion of individuals in each of the two graded classes.

Defining the Zero and Maximum Classes For a scale that includes only count items, the zero and maximum classes are easily defined, with a response pattern of all zeros characterizing members of the zero class and a response pattern of all 30s characterizing members of the maximum class. Thus, defining the latent classes for the models used to address the Primary Aims was straightforward – responses of 0 and 30 map onto the minimum and maximum number of days, respectively. The Secondary Aim of this research used a scale with mixed item types,

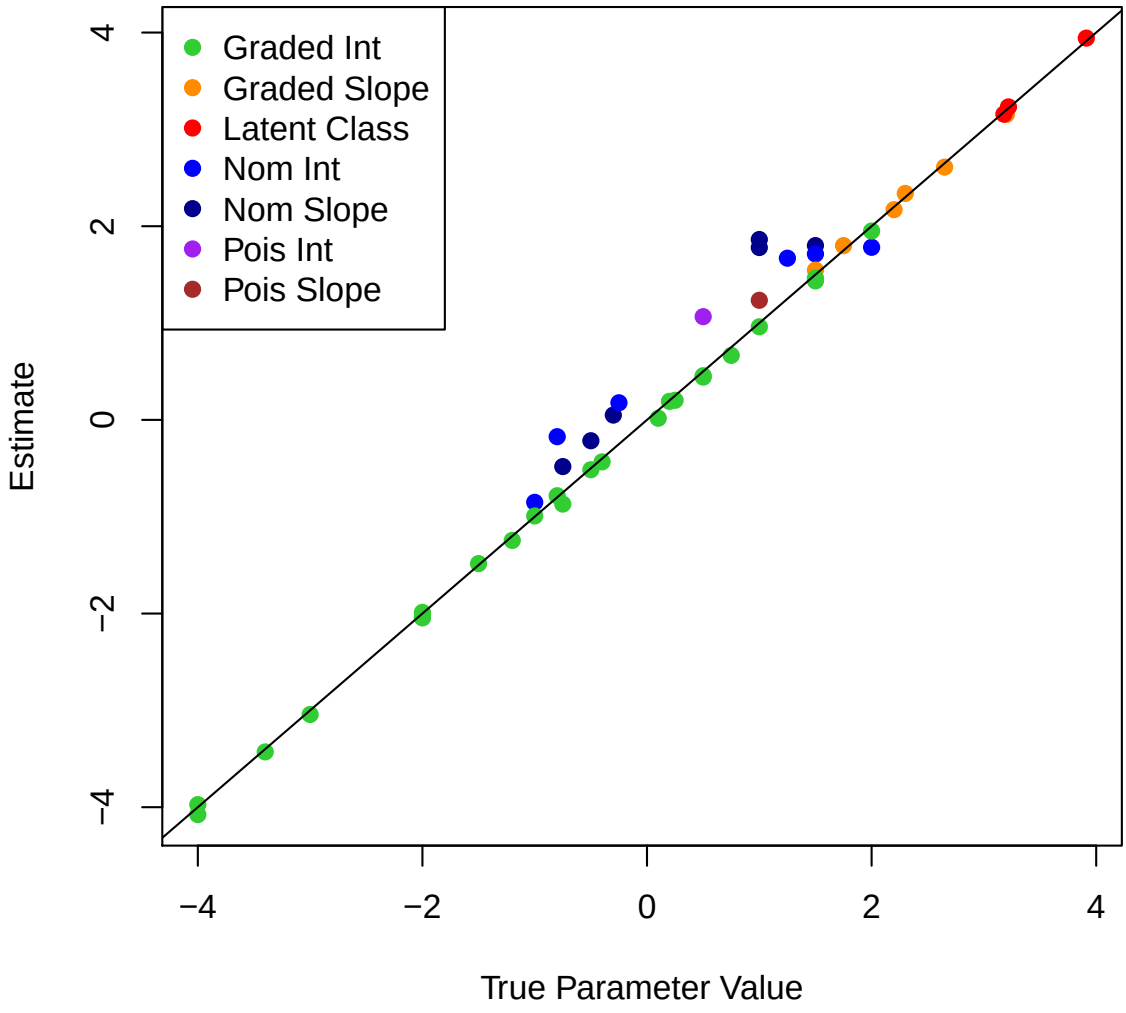
however, changing the response patterns comprising the zero and maximum classes. The zero class was still defined as a response of zero to all seven items: $\boldsymbol{U} = \mathbf{0} = (0, 0, 0, 0, 0, 0, 0)$; however, due to their Likert-type response format, the first six zeros represent a response of "None of the time" on a 0-4 Likert scale, not 0 days on a 0-30 count scale. Likewise, the maximum class was still defined as the most severe response for each item, but for the six Likert items, the most severe response correspond to endorsement of "All of the time" on a 0-4 Likert scale instead of 30 days on a 0-30 count scale: $\boldsymbol{U} = (4, 4, 4, 4, 4, 4, 30)$. As a result, the terms "zero class" and "maximum class" are used slightly more liberally within the Secondary Aim.

3.1 Simulation

The Latent Class Model with a Poisson IRT Model for the Exact Count Class To test the implementation of parameter estimation in the R program, I first fit the latent class IRT model with mixed item types to data that were simulated from known population parameters, shown in Table 3 in Chapter 2. When the latent class IRT model with a Poisson component for the exact count class was fit to data that were simulated from the parameters in Table 3, the model converged after 112 iterations, requiring approximately 5 hours on a desktop computer with a quad-core 2.4GHz Intel Core processor and 4GB of RAM. Parameter estimates are plotted against the data-generating parameters in Figure 14. This figure shows that the R program recovers the data-generating parameters fairly well, with most points hovering around the identity line. As was true of the previous model fittings, the parameter estimates showing the greatest deviation from the identity line correspond to the nominal response IRT model. Importantly, the results of the simulation also suggest that this particular latent class IRT model with mixed item types is identified; even though there is only a single count item on the scale, the proportions of people in each of the latent classes are recovered.

The Latent Class Model with a Negative Binomial IRT Model for the Exact Count Class I also carried out a small simulation to examine whether implementation of parameter estimation in the R program could recover the parameters when data were simulated from a latent class model with a negative binomial component for the exact count class. When this alternative latent class IRT model was fit to data that were simulated from the parameters in Table 3, the model converged after 111 iterations and approximately 5.5 hours on a desktop computer with a quad-core 2.4GHz Intel Core processor and 4GB of RAM. Unlike the previ-

Figure 14. Latent class IRT model with a Poisson IRT model for the exact count class and a nominal response IRT model for the rounding/selected response class: Estimated parameters vs. data-generating parameters, for data simulated using the parameters in Table 3.



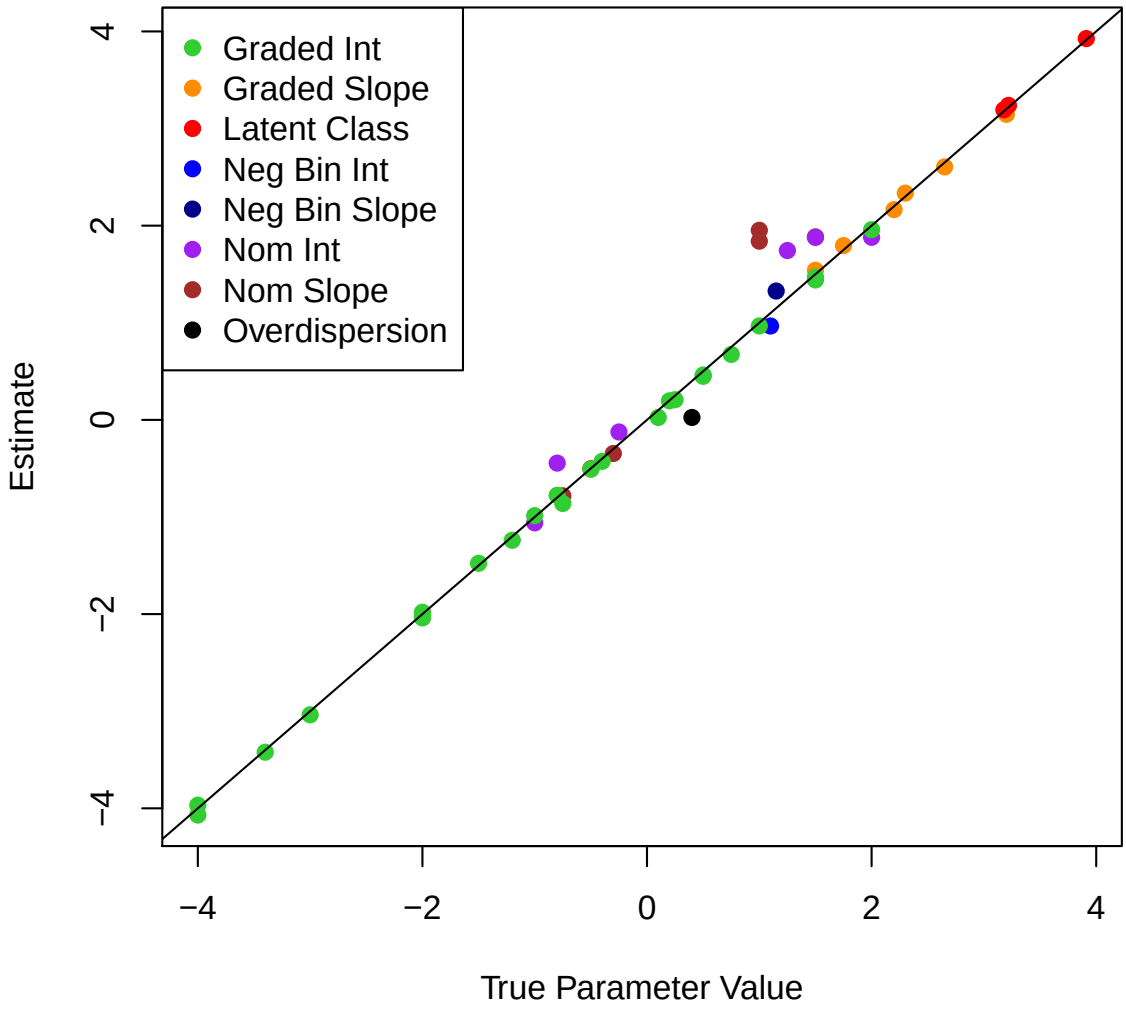
ous model, which uses a Poisson IRT model for the exact count class, this model includes an additional parameter to account for overdispersion in responses to the count item. Figure 15 shows the parameter estimates plotted against the data-generating parameters. Like its Poisson counterpart, this figure indicates that the R program recovers the data-generating parameters reasonably well, including the overdispersion parameter from the negative binomial IRT model.

3.2 Empirical Analysis of the BRFSS Data

The Latent Class IRT Model with a Poisson IRT Model for the Exact Count Class

The results of both simulations suggest that the latent class IRT model with a Poisson component for the exact count class, as well as the latent class IRT model with a negative binomial component for the exact count class, are identified, and that implementation of parameter estimation in the R program produces reasonably accurate estimates of both the IRT and latent class parameters. After testing the model with simulated data, I fit each of the latent class IRT models to the motivating data set for the Secondary Aim: the mixed item-type scale from the BRFSS comprising six Likert-type items and one count item, shown in Figure 2 at the end of Chapter 1. I first fit the most parsimonious model that omits the overdispersion parameter; this model assumes that, conditional on the latent variable, responses to the count item follow a Poisson distribution. The model converged after 228 iterations and approximately 6 hours on a desktop computer with a quad-core 2.4GHz Intel Core processor and 4GB of RAM; however, after refitting the model with different starting values that yielded larger log likelihoods, I concluded that the solution was a local maximum. To identify the global maximum, I used different sets of starting values that were determined based on a systematic trial-and-error procedure that involved refitting the model several times: For each refitting, I fixed the latent class proportions to different plausible values, where across all models, the proportion of people in the zero and maximum classes were fixed to 0.17 and 0, respectively. With each model fitting, I recorded the log likelihood and used the IRT parameter estimates and fixed latent class proportions associated with the best log likelihood as starting values in estimating the original model, where latent class proportions were treated as unknown. The model with revised starting values converged after 10 iterations and 50 minutes on a desktop computer with a quad-core 2.4GHz Intel Core processor and 4GB of RAM. Note that the small number of iterations and relatively short computing time are artifacts of using starting values that are close to the estimated model

Figure 15. Latent class IRT model with a negative binomial IRT model for the exact count class and a nominal response IRT model for the rounding/selected response class: Estimated parameters vs. data-generating parameters, for data simulated using the parameters in Table 3.



parameters; using naive starting values would require many more iterations before convergence. Parameter estimates for this model can be found in Table 7. While one cannot be certain that the estimates in Table 7 represent a global maximum, there is strong evidence that they are the most likely solution.

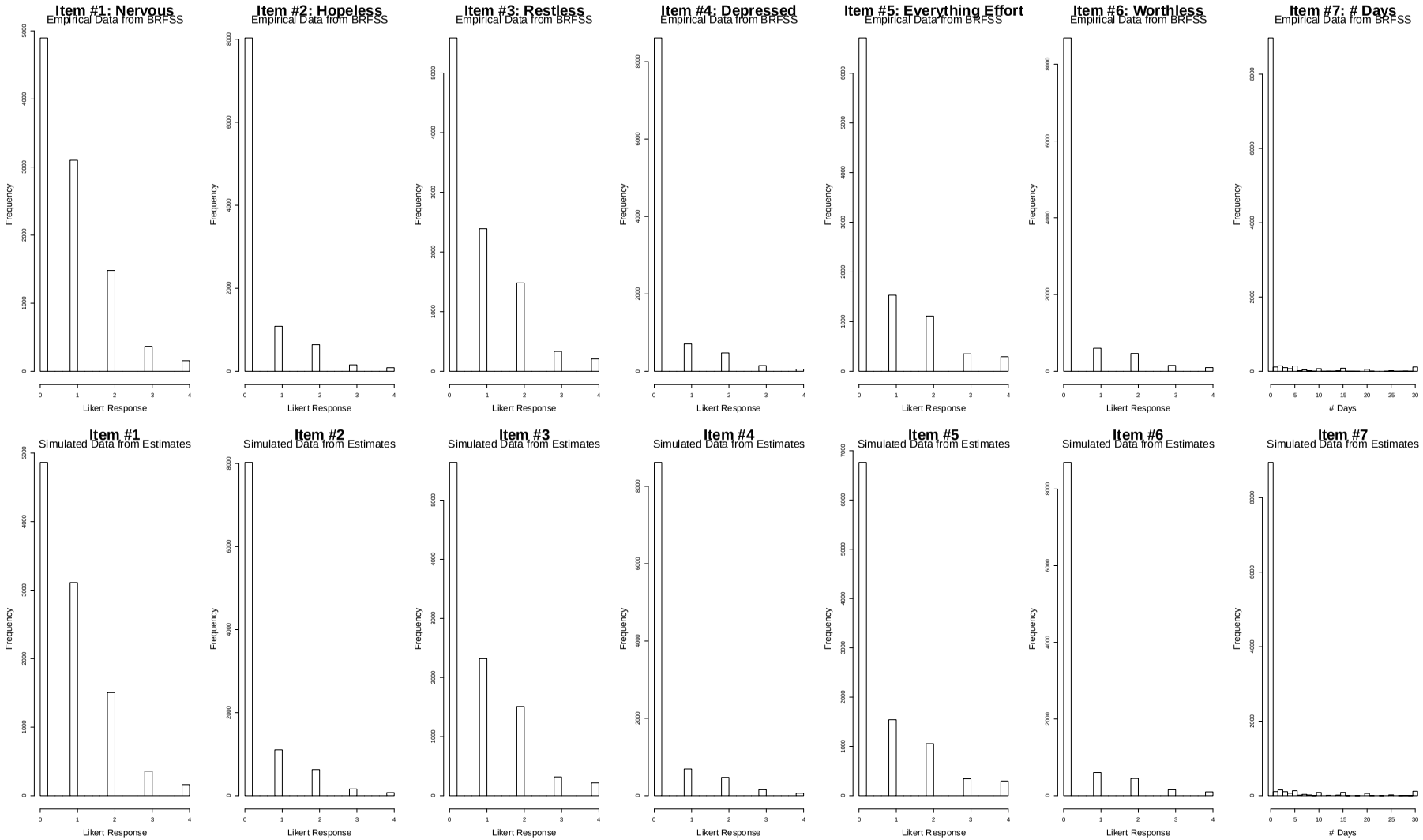
Unlike the results from the analysis of the four count items, in which the majority of individuals were estimated to belong to the exact count class with a smaller proportion estimated to belong to the rounding/selected response class, the reverse is true of the results from the analysis of the scale with mixed item types: Only a small fraction of the sample was estimated to belong to the exact count class (7%), and a much larger proportion was estimated to belong to the rounding/selected response class (75%). Possible explanations for the division of latent class proportions are presented in Chapter 4. Similar to the results found in the Primary Aim analyses, 17% of the sample was found to belong to the zero class; less than 1% was estimated to belong to the maximum class, likely because only 13 of the 10,000 people in the sample had response patterns of $\boldsymbol{U} = (4, 4, 4, 4, 4, 4, 30)$.

To examine how closely the IRT and latent class parameters estimated from the model could reproduce the empirical response distributions, I simulated 10,000 responses to each of the seven items from the parameter estimates in Table 7. The upper panel of Figure 16 displays the empirical response distributions for the BRFSS data; the lower panel shows the data that were simulated from the parameter estimates. As explained in the results section for Primary Aim #1, comparison of the upper and lower panels does not allow a thorough comparison of model fit at the response pattern level; however, the similarity of the empirical and simulated response distributions provides support for the idea that the IRT models are correctly specified. Figure 16 shows that the empirical response distributions for all seven items are reproduced fairly well by the parameters estimated from the model. Table 8 contains the observed and expected frequencies of responses to the count item "How many days did a mental health condition or emotional problem keep you from doing your work or other usual activities?" While this table provides only a limited picture of model fit – it does not take into consideration people's responses to any of the six Likert items – it does suggest that the mixture of Poisson and nominal response IRT models is able to reproduce the observed count distribution very well.

Table 7. Parameter estimates for the latent class IRT model with a Poisson IRT model for the exact count class fit to BRFSS data, $N = 10,000$.

	Item #1	Item #2	Item #3	Item #4	Item #5	Item #6	Item #7
Graded Response IRT Parameters							
a	1.45 (0.03)	3.21 (0.04)	1.34 (0.03)	3.16 (0.04)	1.68 (0.03)	3.02 (0.03)	–
c_1	0.72 (0.03)	-2.50 (0.04)	0.17 (0.02)	-3.43 (0.04)	-0.64 (0.02)	-3.45 (0.04)	–
c_2	-1.45 (0.03)	-4.59 (0.05)	-1.44 (0.02)	-5.08 (0.05)	-2.00 (0.03)	-4.87 (0.05)	–
c_3	-3.54 (0.04)	-7.40 (0.09)	-3.43 (0.04)	-7.52 (0.09)	-3.60 (0.04)	-6.96 (0.08)	–
c_4	-5.06 (0.08)	-9.60 (0.15)	-4.57 (0.06)	-9.75 (0.16)	-4.65 (0.05)	-8.63 (0.12)	–
Nominal Response IRT Paramaters							
a_1	–	–	–	–	–	–	0.00 (–)
a_2	–	–	–	–	–	–	1.28 (0.13)
a_3	–	–	–	–	–	–	1.33 (0.12)
a_4	–	–	–	–	–	–	1.38 (0.10)
a_5	–	–	–	–	–	–	1.25 (0.13)
a_6	–	–	–	–	–	–	0.81 (0.34)
a_7	–	–	–	–	–	–	1.42 (0.09)
c_1	–	–	–	–	–	–	0.00 (–)
c_2	–	–	–	–	–	–	-5.08 (0.14)
c_3	–	–	–	–	–	–	-5.24 (0.13)
c_4	–	–	–	–	–	–	-5.18 (0.11)
c_5	–	–	–	–	–	–	-5.42 (0.13)
c_6	–	–	–	–	–	–	-6.09 (0.22)
c_7	–	–	–	–	–	–	-5.04 (0.10)
Count Model IRT Parameters							
a	–	–	–	–	–	–	0.89 (0.04)
c	–	–	–	–	–	–	0.89 (0.05)
Latent Classes	Zero	Exact Count	Rounding	Maximum			
Proportion	0.17 (0.04)	0.07 (0.04)	0.75 (0.03)	0.00 (0.28)			

Figure 16. Empirical response distributions vs. response distributions simulated from estimated parameters from the latent class IRT model with a Poisson component.



The Latent Class IRT Model with a Negative Binomial IRT Model for the Exact Count Class To test whether an overdispersion parameter is needed to describe the empirical response distribution of the count item, I also fit a latent class IRT model with a negative binomial component for the exact count class. I used as starting values the parameter estimates from the latent class IRT model with the Poisson component for the exact count class. The negative binomial model converged after 60 iterations and approximately 2 hours on a desktop computer with a quad-core 2.4GHz Intel Core processor and 4GB of RAM. Again, note that close starting values significantly decreases the number of iterations – and thus estimation time – necessary for model convergence. Parameter estimates can be found in Table 9.

Comparison of the Latent Class IRT Models with Poisson and Negative Binomial IRT Models for the Exact Count Class In comparing the parameter estimates from the Poisson and negative binomial latent class IRT models, shown in Tables 7 and 9, respectively, some features of the results are noteworthy. First, the IRT parameter estimates are nearly identical between the two models. This suggests that including the overdispersion parameter does not change the a- and c-parameters that are estimated from the count IRT model: $a \approx 0.89$ and $c \approx 0.89$ for both the Poisson and negative IRT models. Second, the overdispersion parameter estimate for the negative binomial IRT model is very small ($\delta = 0.007 \approx 0$). The near-zero value of the overdispersion parameter indicates that the negative binomial IRT model essentially functions as a Poisson IRT model when fit to these data, and that the Poisson IRT model is likely sufficient. Table 10 provides further evidence that the Poisson IRT model is sufficient for this particular count item, with both the AIC and BIC indicating that the overdispersion parameter does not improve model fit. Therefore, I proceed with interpretation of the latent class IRT model that includes a Poisson component for the exact count class, ignoring overdispersion.

4. The Latent Class Model with a Poisson IRT Model for the Exact Count Class

Trace lines for the seven items comprising the mixed item-type scale can be found in Figures 17 and 18. Figure 17 shows the GRM trace lines for the six Likert-type items; these trace lines apply to both the exact count and rounding/selected response classes, because count response style is not relevant for the Likert-type items. Figure 18 shows the Poisson trace lines for the exact count class (upper panel) and NRM trace lines for the rounding/selected response class

Table 8. Frequencies of observed vs. expected responses to "How many days did a mental health condition or emotional problem keep you from doing your work or other usual activities?" according to latent class IRT model with a Poisson component for the exact count class and a nominal component for the rounding/selected response class.

# Days Condition	Observed Frequency	Expected Frequency
0	8945	8918
1	110	158
2	155	122
3	108	95
4	64	56
5	133	125
6	14	22
7	35	31
8	19	17
9	2	20
10	87	102
11	0	9
12	10	3
13	0	3
14	16	3
15	89	95
16	1	4
17	0	1
18	1	3
19	0	4
20	60	52
21	3	1
22	0	2
23	1	0
24	0	0
25	21	15
26	0	1
27	2	0
28	4	0
29	1	0
30	119	136

Table 9. Parameter estimates for the latent class IRT model with a negative binomial IRT model for the exact count class fit to BRFSS data, $N = 10,000$.

	Item #1	Item #2	Item #3	Item #4	Item #5	Item #6	Item #7
Graded Response IRT Parameters							
a	1.45 (0.03)	3.21 (0.04)	1.34 (0.03)	3.16 (0.04)	1.68 (0.03)	3.02 (0.03)	–
c_1	0.72 (0.03)	-2.50 (0.04)	0.17 (0.02)	-3.43 (0.04)	-0.64 (0.02)	-3.45 (0.04)	–
c_2	-1.45 (0.03)	-4.59 (0.05)	-1.44 (0.02)	-5.08 (0.05)	-2.00 (0.03)	-4.87 (0.05)	–
c_3	-3.54 (0.04)	-7.40 (0.09)	-3.43 (0.04)	-7.52 (0.09)	-3.60 (0.04)	-6.96 (0.08)	–
c_4	-5.06 (0.08)	-9.60 (0.15)	-4.57 (0.06)	-9.75 (0.16)	-4.65 (0.05)	-8.63 (0.12)	–
Nominal Response IRT Paramaters							
a_1	–	–	–	–	–	–	0.00 (–)
a_2	–	–	–	–	–	–	1.29 (0.13)
a_3	–	–	–	–	–	–	1.32 (0.12)
a_4	–	–	–	–	–	–	1.38 (0.10)
a_5	–	–	–	–	–	–	1.25 (0.13)
a_6	–	–	–	–	–	–	0.81 (0.34)
a_7	–	–	–	–	–	–	1.42 (0.09)
c_1	–	–	–	–	–	–	0.00 (–)
c_2	–	–	–	–	–	–	-5.08 (0.14)
c_3	–	–	–	–	–	–	-5.24 (0.13)
c_4	–	–	–	–	–	–	-5.18 (0.11)
c_5	–	–	–	–	–	–	-5.42 (0.13)
c_6	–	–	–	–	–	–	-6.09 (0.22)
c_7	–	–	–	–	–	–	-5.04 (0.10)
Count Model IRT Parameters							
a	–	–	–	–	–	–	0.88 (0.04)
c	–	–	–	–	–	–	0.89 (0.05)
δ	–	–	–	–	–	–	0.01 (0.07)
Latent Classes	Zero	Exact Count	Rounding	Maximum			
Proportion	0.17 (0.04)	0.07 (0.04)	0.75 (0.03)	0.00 (0.28)			

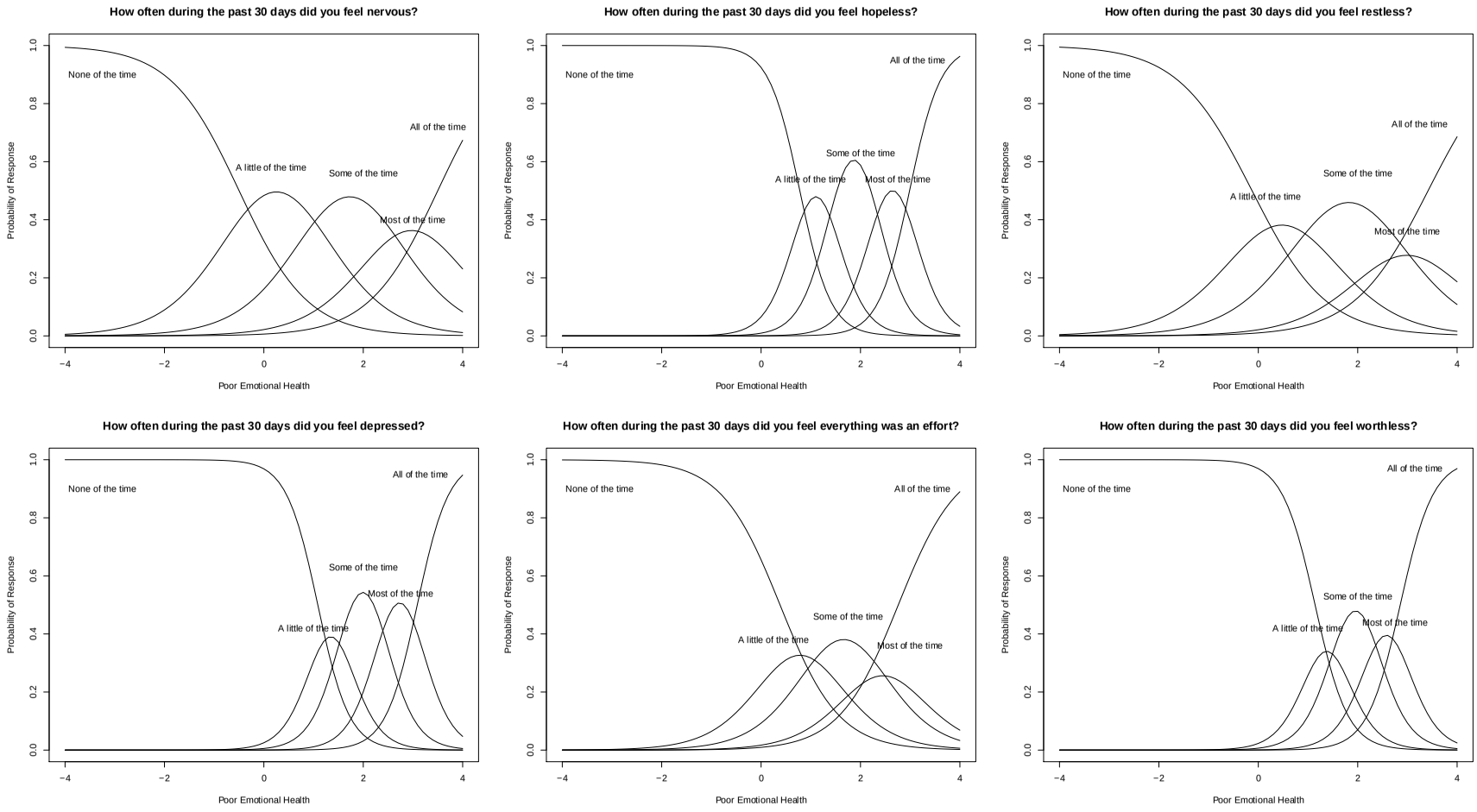
Table 10. Model comparison of the Poisson vs. the negative binomial latent class IRT models.

	# Parameters	AIC	BIC
Poisson	47	**98071.70**	**98410.59**
Negative Binomial	48	98073.70	98419.80

(lower panel) for the count item.

The GRM trace lines in Figure 17 indicate that the Hopeless, Depressed, and Worthless items are the most discriminating; in particular, these items are able to separate individuals who are at above average levels of Poor Emotional Health ($\theta \geq 0$). Like the NRM trace lines described in Primary Aim #1, the GRM trace lines trace the probability of each response category across different levels of the Poor Emotional Health latent variable. For example, for someone with a Poor Emotional Health level of $\theta = 2$, the "Some of the time" response option consistently has the highest probability of endorsement across all six Likert-type items. While less discriminating, the Nervous, Restless, and Effort items also tend to do a better job of separating individuals who fall at higher levels of Poor Emotional Health ($\theta \geq 0$); people at below average levels of Poor Emotional Health tend to have near-one probability of endorsing the "None of the time" response option.

Figure 17. GRM trace lines for the six Likert items on the BRFSS mixed item-type scale.



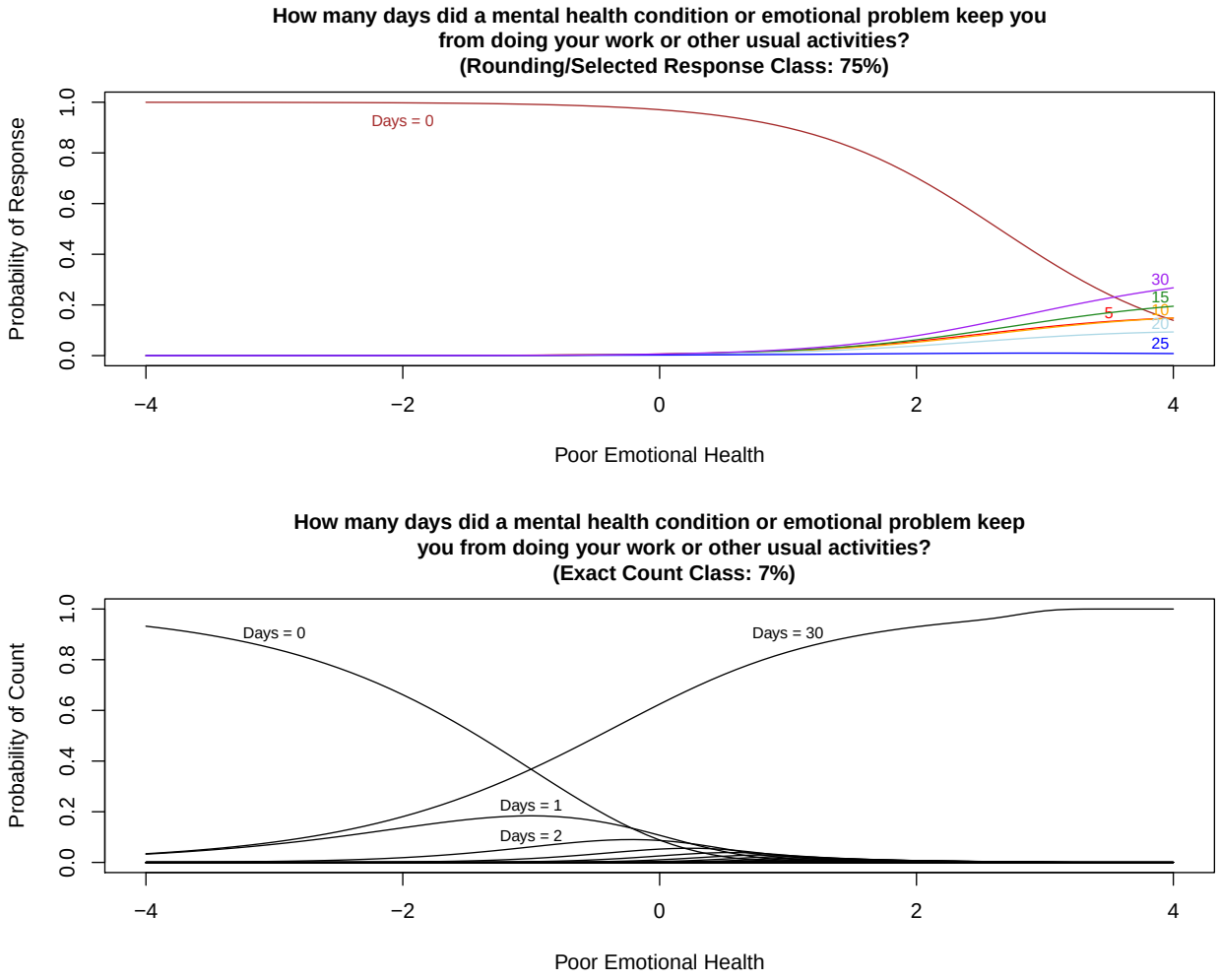
The relatively flat curves in both sets of trace lines in Figure 18 reveal the overall low discriminating power of the count item on the mixed item-type scale. Part of the reason for the low discrimination parameter in both classes is because 90% of the sample reported 0 days for this item. Within the rounding/selected response class (the upper panel of Figure 18), 0 days has the highest probability of endorsement across nearly all levels of the Poor Emotional Health latent variable. It is not until θ is more than three standard deviations above average, which represents only a very small fraction of the sample, that any of the trace lines cross, and when they do cross, 30 becomes the most likely response category. Within the exact count class, 0 days and 30 days are also the most likely responses across all levels of the latent variable; however, days ranging from one to five also have non-zero probabilities of endorsement for people who are low on Poor Emotional Health. These trace lines suggest that the Poisson IRT model best describes people who endorse low counts (i.e., fewer than 6) and are thus low on the Poor Emotional Health; with the exception of 30, there is essentially zero probability of endorsing any count greater than 5. People who endorse higher counts that are multiples of five are much more likely to belong the rounding/selected response class, either rounding their answer to the nearest multiple of five or treating the item as a multiple choice question. This finding is supported by the observed counts in Table 8, which shows that "exact counts" (i.e., those that are not multiples of five) are typically only observed at the low end of the open-ended count scale. Less than 1% of the sample reported a non-multiple-of-five count greater than 15.

In interpreting the Poisson and NRM trace lines in Figure 18, it is important to keep in mind that only 7% of the sample was estimated to belong to the exact count class, whereas 75% of the sample was estimated to belong to the rounding/selected response class. Thus, the NRM trace lines are much more representative of this sample than the Poisson trace lines. Further, with only one count item determining the composition of these two graded classes, the exact count and rounding/selected response latent class estimates are likely rather unstable. For these reasons, caution should be exercised in drawing too strong conclusions from the Poisson trace lines, as these estimates are based on a very small proportion of the sample.

4.1 The Contribution of the Count Item to Measurement Precision

An additional goal of this research was to determine the value of including count items on scales; specifically, how much does the count item on this particular mixed item-type scale con-

Figure 18. Poisson (upper) and NRM (lower) trace lines for the count item on the BRFSS mixed item-type scale.

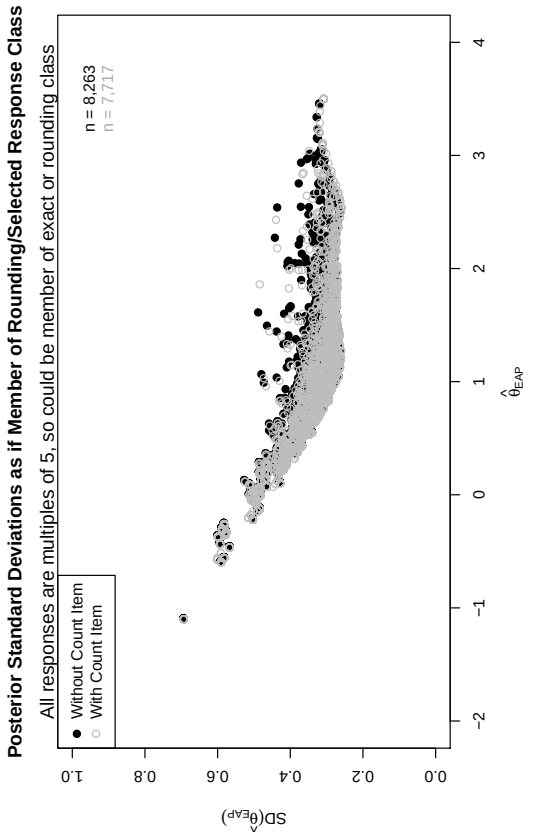


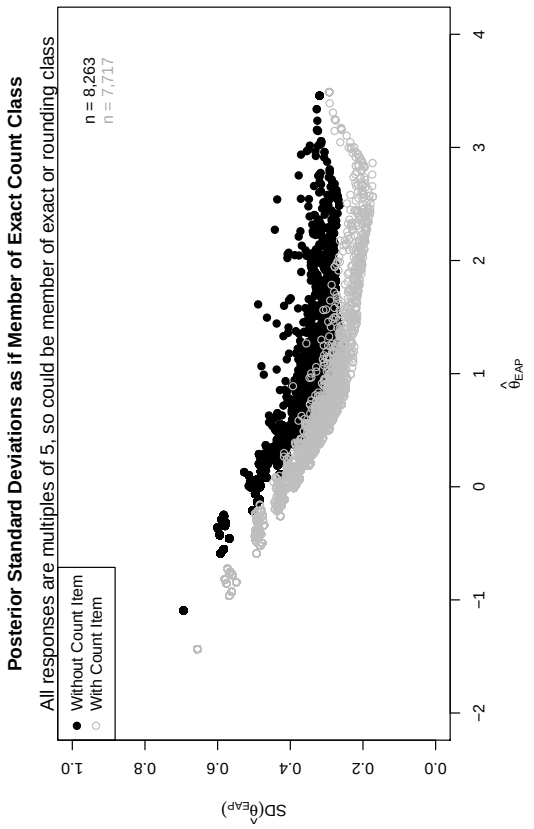
tribute to the precision of measurement? To address this question, I computed the posterior standard deviations of the scale scores twice. First, I estimated scale scores and posterior standard deviations for respondents when only the six Likert-type items were included on the scale. I computed these values for 8,263 individuals in the sample, which excludes the (approximately) 17% of the sample estimated to belong to the zero class; because less than 1% of the sample was estimated to belong to the maximum class, I did not exclude any of the $\boldsymbol{U} = (4, 4, 4, 4, 4, 4, 30)$ response patterns from the scoring sample. To visualize the relationship between level of Poor Emotional Health and scale score precision, I plotted each scale score relative to its posterior standard deviation. Then, I estimated scale scores and posterior standard deviations for the same respondents, only this time, I included the count item as a seventh item on the scale. Because including the count item introduces two new latent classes that reflect response style – the exact count class and the rounding/selected response class – I subdivided the scoring samples based on plausible latent class membership. After removing 17% of the sample that represents the zero class, I divided people into two types of response patterns: those with a multiple-of-five count response, and those with a non-multiple-of-five count response. The 7,717 individuals with a multiple-of-five count response could belong to either the exact count or rounding/selected response class; thus, I computed both types of scores for these individuals, using either the Poisson or nominal response IRT model for the count item. The remaining 546 individuals with a non-multiple-of-five count response could only belong to the exact count class; these individuals received only one score that used the Poisson IRT model for the count item.

Figure 19 shows the posterior standard deviations for each type of scoring method (represented by location of the plot) and scale (represented by dot color). The top row of Figure 19 represents the majority of people in the sample – the 75% belonging to the rounding/selected response class. The bottom row represents the 7% of the sample that belongs to the exact count class. The black dots are identical in all three plots: These are scores that are based only on the six Likert-type items for all 8,263 individuals in the scoring sample (i.e., the proportion of people who are not members of the zero class). Because the exact count and rounding/selected response classes do not exist when the scale only includes the six Likert-type items, these scores do not depend on latent class membership. The grey dots represent scores that are based on

the six Likert-type items in addition to the count item. Because introducing the count item also introduces two latent classes for response style, the number of grey dots in each plots varies depending on the latent class. The 7,717 grey dots in the plot in the top left corner of Figure 19 are scores that are computed from the full scale when the count item is scored according to a Poisson IRT model; these scores are for individuals with a multiple-of-five count response. The 7,717 grey dots in the plot in the top right corner of Figure 19 are scores that are computed from the full scale when the count item is scored according to a nominal response IRT model; these scores are also for individuals with a multiple-of-five count response. In actuality, the majority of these 7,717 people belong to the rounding/selected response class, with only a very small fraction belonging to the exact count class. The grey dots in the plot in the bottom left corner of Figure 19 are scores that are computed from the full scale when the count item is scored according to a Poisson IRT model; these scores correspond to count responses that are not multiples of five, representing the remaining 546 individuals in the scoring sample who must belong to the exact count class.

Figure 19. Precision of measurement as a function of scale scores that are computed with and without the count item.





Each of the plots in Figure 19 shows the contribution of the count item to the measurement precision of the scale across all levels of the Poor Emotional Health latent variable; however, the improvement in measurement precision depends on latent class membership. Most notably, the improvement in measurement precision is very small when the count item is scored as if the person is a member of the rounding/selected response class: The grey dots nearly map onto the black dots in the plot shown in the upper right corner of Figure 19. The very slight decrease in the posterior standard deviations is what would be expected by including an additional item on a scale, regardless of the type of item, and it is important to point out that this minimal contribution of the count item to overall measurement precision describes 75% of the sample. Thus, for most people in the sample, the increase in measurement precision is trivial.

When the count item is scored according to a Poisson IRT model – and therefore, its true count properties are preserved – the count item has a considerably larger contribution to the measurement precision of the scale: The left column of Figure 19 shows that across nearly all levels of the Poor Emotional Health latent variable, the grey dots (the full scale) represent smaller posterior standard deviations than the black dots (the only-Likert-items scale). This suggests that when someone answers the open-ended count item according to a strict Poisson process, the item does substantially improve measurement precision; however, this improvement in precision should be interpreted with caution. It is only 7% of the population – 700 of the combined 8,263 people that are shown in the left-hand plots of Figure 19 – that are members of the exact count class; thus, one would expect a substantial improvement in measurement precision for a smaller number of people than what is shown in these figures. This may be considered a relatively small gain for the computational burden of including the count item on the scale.

Chapter 4: Discussion & Conclusions

The primary goal of this research was to develop a latent class IRT model that could be used to account for zero inflation, maximum inflation, and heaping in multivariate open-ended count item response data. Results of the empirical analyses suggest that a latent class IRT model that uses a Poisson or 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 mixture IRT modeling approach is useful in analyzing data from scales with multiple open-ended count items – in particular, the four count items that comprise a subscale on the BRFSS. In addition to IRT parameter estimates, the model also provides estimates of the proportion of people in each of the four proposed latent classes, which may represent distinct subpopulations. The results also indicate that it is necessary to include all four latent classes to reflect different types of response styles and subpopulations. While an IRT model that assumes a standard count distribution for item responses may be able to describe individuals who respond to items according to a strict count process, it cannot account for heaping at preferred digits and inflation at zero and the maximum. The results are a bit more tenuous and offer less compelling support for constructing scales that include only a single count item, even when a mixture IRT model may be able to accommodate heaping and inflation in the count item responses. It is likely that the latent class estimates are unstable when they are based on only a single count item. While the latent class estimates may be unstable, the results also suggest that including a single count item on a scale can substantially improve measurement precision for individuals who respond to the item according to a count process; however, for people who treat the count item as a multiple choice question, the increase in measurement precision is very small. This chapter offers a more thorough analysis and interpretation of these results.

1. Discussion of the Empirical Results of the Primary Aims

The results reveal a peculiarity in the way some people respond to open-ended count items on questionnaires, and this peculiarity likely has implications for scale development. When responding to open-ended count items, a sizable proportion of individuals do not appear to treat the item as an open-ended count with 31 response options, $U_j = \{0, 1, 2, ..., 30\}$, but instead treats it as a selected response item with seven possible response categories, $U_k = \{0, 5, 10, 15, 20, 25, 30\}$. Or, these individuals round their count responses to the nearest multiple of five. Further, within the rounding/selected response latent class, the seven response categories do not appear to be in 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 for the Depressed and Anxious items, which are the two items most strongly related to the latent variable that is measured by the scale. 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.

Related to this shortcut that some respondents seem to use in choosing 30 days over 25 days, an explanation for the unexpected ordering of response categories is that there may be an additional latent class of individuals in the population who treat open-ended count items as binary: Instead of treating these items as a multiple choice items with seven possible response categories, they dichotomize their responses into one of two categories. If they fall somewhere on the lower end of the open-ended count scale, they choose 0 days, and if they fall somewhere on the higher end of the scale, they choose 30 days. This explanation is consistent with the large number of people selecting 0 or 30 days who are not members of the zero or maximum classes.

This explanation is also consistent with the trace lines in Figure 8 in Chapter 3, where for all four items, the response categories that are typically associated with the greatest probabilities of endorsement across the entire Poor Emotional Health continuum are 0 days and 30 days.

The empirical results also support the idea 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. There are multiple reasons someone may belong to one of these two classes. One reason is that respondents may be at some floor or ceiling level of the latent variable. 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 Poor Emotional Health should be used instead. Or, perhaps it is a different latent variable altogether that describes these individuals. In either case, further assessment of people who may belong to the zero or maximum classes is a logical next step.

A second 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; therefore, the items do not apply to these respondents. While the analog to emotional health symptoms is not quite as intuitive, it is possible that some respondents are so low (or perhaps absent) on psychopathology that they view the items as irrelevant. These individuals, who have response patterns $\boldsymbol{U} = \mathbf{0} = (0, 0, 0, 0)$ and belong to the zero class, may be viewed similarly to substance use abstainers. Regardless of the reason, in theory, the majority of people who endorse 0 days or 30 days for every item should not be treated and scored as if they fall along the same Poor Emotional Health continuum as the rest of the population. Ignoring these extreme classes leads

to a biased representation of the Poor Emotional Health latent variable in the population.

While it is possible to use the estimated latent class proportions to determine the number of individuals in the sample who should not be scored along the latent variable continuum (i.e., the number of zero and maximum class members), it is not possible to identify from the IRT model the specific individuals who belong to these classes. If the goal of the research is to examine scores at the population level – for example, to show the distribution of scale scores in the population – this may not be a disadvantage; one can simply remove the proportions of people with all-0 and all-30 response patterns that represent the zero and maximum classes, respectively, and score everyone else according to one of the two graded components. However, if the goal of the research is to assign scores to individuals for clinical assessment, this is a major limitation of using a latent class IRT model. One cannot definitively assign an all-0 (or all-30) response pattern to the zero (or maximum) class, because it is also possible that someone with such a response pattern belongs to one of the two graded classes. Finkelman et al. (2011) recommend scoring individuals with all-0 and all-30 response patterns as if they belong to the graded class; however, the model in Finkelman et al. (2011) includes only one graded class, so scoring is more straightforward. When there is more than one graded class, as is true of these models, it is similarly not possible to separate members of the two graded classes for scoring purposes. If the response pattern includes at least one non-multiple-of-five count, the respondent necessarily belongs to the exact count class and only one type of scoring method is applicable. However, for all other response patterns, which comprise a substantial proportion of the sample, it is not possible to assign respondents to latent classes. If the goal of the research is to score people, one option is to assign people with multiple-of-five response patterns two different scores: one score as if the respondent were a member of the exact count class, and the other score as if the respondent were a member of the rounding/selected response class. A second option is to simply choose one scoring method. Depending on where a person falls on the latent variable, however, the scoring method may be highly consequential: The two sets of scores are essentially uncorrelated at extreme levels of Poor Emotional Health. For these reasons, scoring remains a complex issue in latent class IRT modeling.

2. Discussion of the Empirical Results of the Secondary Aim

The Secondary Aim of this research had two main goals. One goal was to develop a latent class IRT model, similar to those used for the analyses in the Primary Aims, for scales comprising mixed item types – specifically, a scale with six Likert-type items and a single count item that exhibits inflation and heaping. The other goal was to evaluate the contribution of the single count item to the scale's precision of measurement: Does including the count item reduce the posterior standard deviations associated with scale scores?

The first goal was met with moderate success: It is possible to estimate a latent class IRT model using a scale that comprises mixed item types. Even though the scale used includes just one count item, the proportions of individuals belonging to the exact count and rounding/selected response classes were estimated successfully. Unlike the results from the analysis of the scale comprising only count items, which suggest that the majority of respondents treat count items as open-ended scales and utilize the full range of possible responses, analysis of the scale with mixed item types suggests that a large majority of respondents – 75% – treat the count item as a multiple choice question (or round their answers to the nearest multiple of five). The reversal of latent class estimates may be due to differences in the structure and item ordering of the two scales. The scale used to address the Primary Aims comprises only count items; however, the scale used to address the Secondary Aim comprises six Likert-type items and one count item, in which respondents answer the six Likert-type questions before responding to the count item. It is likely that the response options of the Likert-type items, which are categorized into five increasing units of time along a continuum, primed respondents to categorize the open-ended count item into similar categories, explaining the very low proportion of people estimated to belong to the exact count class. In completing the questionnaire composed of only count items, it is less likely that respondents were conditioned to categorize their responses, possibly explaining the much larger proportion of people who used the full range of the count scale.

While priming with Likert-type items may explain the division of latent class proportions, it should be emphasized that with only one count item included on the scale, the latent class estimates are likely unstable. Due to an apparently multimodal and ridged log likelihood surface, there were several stationary points that the **nlm** optimizer treated as maxima. Multiple sets of starting values were required to arrive at the solution that is believed to be the global maximum,

and depending on the starting values used, the division of proportions between the exact count and rounding/selected response classes varied widely. Regardless of whether the parameter estimates presented here are the correct solution, the latent classes are probably not well-defined. Additionally, the zero and maximum latent classes carry less meaning on scales of mixed item types. While it is still possible to define the zero and maximum classes as comprising the most extreme responses, endorsing a Likert response of "None of the time" is qualitatively different from choosing a count response of 0 days. Likewise, selecting a Likert response of "All of time" does not necessarily suggest the same symptom severity as reporting a count response of 30 days. For all of these reasons, researchers should proceed more cautiously in interpreting latent classes on scales with heterogeneous response formats, especially in the presence of only a single count item.

An additional goal of this research was to evaluate the inclusion of count items on scales with mixed item types. The results suggest that including a count item can substantially improve the precision of measurement, with some posterior standard deviations dropping by as much as 50%; however, this statement requires some important qualification. First, measurement precision is only substantially improved for people belonging to the exact count class; this class comprises a mere 7% of the population. For the remaining 75% of the scoring sample – those who either round their answers or treat the item as multiple choice – the contribution of the count item is trivial, and it may not be worth the additional effort to include the count item on the scale. However, if the researcher believes that members of the exact count class report counts according to a strict count process, such as that dictated by the Poisson IRT model, the count item can be very informative. Second, as mentioned in the earlier discussion of the limitations of using scale scores from latent class IRT models, it is impossible to determine which people with a multiple-of-five count response belong to the exact count class; thus, the large reduction in posterior standard deviations cannot be of practical benefit in assigning scores to individuals in the sample. A clinician may decide to use the exact count scoring method only for the 546 individuals who must belong to the exact count class because their count responses are non-multiples-of-five; however, this is not entirely accurate, as it excludes 1.5% of the population with multiple-of-five responses who also belong to the exact count class, and it creates an asymmetrical scoring approach that may make it difficult to compare individuals in

the sample.

3. Recommendations

The models developed as part of this research are computationally complex. Not only does parameter estimation require several hours of computing time, but to my knowledge, user-friendly software that can implement these types of latent class IRT models is limited, or perhaps non-existent. 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, such as writing one's own EM algorithm. Researchers can obviate such complex modeling techniques by not including items that elicit a retrospective count response on their scales and questionnaires. There are several alternative methods of framing the question that can greatly simplify item-level analyses.

Perhaps the most obvious strategy that can be used to avoid eliciting open-ended count responses is to bin the response options before administering the questionnaire. For example, instead of presenting respondents with an open-ended count scale, $u_j = \{0, 1, 2, ..., 30\}$, researchers could provide a fixed-category response format – for example, $u_j = \{0, 1-5, 6-10, ...\}$. This type of modification serves at least two purposes. First, it eliminates the issue of individual differences in count response style; that is, the exact count and rounding/selected response classes are no longer needed. Second, it may reduce recall error. Framing the question with binned counts is less taxing on the respondent's memory than asking for an exact count that is cumulative over a period of time, which is highly prone to recall error. Binning count responses does not eliminate the need for zero and maximum classes, however; I recommend that researchers still include stand alone zero and maximum response options to help identify individuals who may belong to either of these classes, even if all other counts are binned.

If a researcher is truly interested in the frequency of a specific thought or behavior, such as the number of days someone feels depressed in a month, an alternative and likely more reliable approach could be to use a daily diary response format instead of retrospective counts. At the end of each day, the respondent could indicate whether he or she felt depressed on that particular day; then, the researcher could tally the counts at the end of the month to get a more accurate total frequency instead of asking the researcher to recall the total count

retrospectively. One clear 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 standard deviations 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.

If a researcher wishes to include retrospective self-report open-ended count items on scales, it is advisable to include more than one count item. A single count item may not provide enough information about response style to produce stable estimates of latent class membership, as described in the empirical discussion of the results. Including multiple count items on a scale is likely to result in improved latent class definition.

4. Limitations

4.1 Absolute Model Fit

One of the major limitations of using open-ended count IRT models is that it is difficult (or perhaps impossible) to assess absolute model fit, due to the extremely large number of possible response patterns. Goodness of fit is typically examined by comparing observed and expected values; in the case of IRT, the observed value is the number of people in the data with a given response pattern, and the expected value is the IRT model-predicted number of people with that particular response pattern. The less discrepancy between the observed and expected response pattern frequencies, the better the model fit. When the number of possible response patterns is limited, it is feasible to compare the observed and expected frequencies for each response pattern and calculate a measure of overall model fit. In this application, however, it is likely not possible. Four count items, each with 31 possible open-ended count responses, yield nearly one million possible response patterns; as the number of count items increases, this number becomes even more unmanageable, producing multi-way contingency tables with extreme sparseness. Such extreme sparseness occurs because as the number of response categories for each item increases, the number of possible response patterns increases, and the sum of the probabilities across all

response patterns must sum to 1 (Bartholomew & Tzamourani, 1999; Maydeu-Olivares, 2013). Therefore, even with a small number of items, many of the cells in the contingency table are expected to have frequencies of zero, creating challenges in the development of goodness of fit statistics. Because the absolute fit of IRT models that are fit to open-ended count items is so difficult to assess, I was limited to measures of relative model fit that can be computed from the model log likelihood. While statistics such as the AIC and BIC are useful for model selection, one is unable to draw conclusions about the degree to which the model fits the data – and thus, the validity of the inferences drawn from the model – without tests of overall goodness of fit.

4.2 Bounded vs. Unbounded Counts

The Poisson and negative binomial distributions used in defining the count IRT models assume that the observed counts are unbounded. To account for the fact that the open-ended counts could not exceed 30, the trace line corresponding to the 30-days count was calculated as one minus the sum of the trace lines for the preceding counts. For most of the count items, the Poisson and negative binomial IRT models served as reasonable approximations for the empirical response distributions – very high counts were rarely observed in the data, and most of the 30s were manifestations of a rounding or selected response process, not a count process. However, it is likely that an alternative count IRT model – specifically, one that can accommodate bounded counts – is more appropriate for the Energy item. As discussed in Chapter 3, the empirical response distribution for the (reversed) Energy item is not well-approximated by an IRT model that assumes unbounded counts: The negative binomial IRT model overpredicts the frequency of respondents reporting counts between 20 and 30 (reversed responses between 0 and 10) and underpredicts the frequency of respondents reporting counts between 0 and 10 (reversed responses between 20 and 30). This discrepancy between the observed and predicted counts occurs because the negative binomial distribution, as well as other standard count distributions, cannot account for the increasing number of people reporting counts toward the upper limit of the 0-30 count range. 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 model could include beta-binomial IRT models for the exact count class to better accommodate bounded count responses.

4.3 Generalizability

It is important to emphasize that this dissertation was not a large parameter recovery simulation study. The purpose of using simulated data was to evaluate the implementation of parameter estimation in R and to test whether the models are identified, as all proposed models include dozens of parameters. For each model that I tested, I simulated a single data set with a very large sample size such that parameter estimates would be close to the true values. While a small simulation study for each proposed model verified that the R programs were written correctly and that the models are identified, such a limited simulation does not permit broad generalizability to multiple conditions. Large scale simulation studies typically involve many more replications and variability in data generating conditions. The simulations presented here were not designed to permit generalizability of the proposed models to multiple conditions. Due to the limited scope of the simulations, these models may not be appropriate with all types of count data generated under different conditions (e.g., bounded vs. unbounded counts, retrospectively- vs. concurrently-reported counts, etc.). However, the results of the empirical studies suggest that the proposed methods and models have usefulness in the analysis of questionnaires with retrospectively-reported count responses, and it is likely that similar techniques may be applicable to other types of data. A larger simulation study to determine the conditions under which these latent class IRT models are appropriate is a next step in this line of research.

5. Future Directions

While the results of these studies address the research questions posed in the Primary and Secondary Aims, several questions remain to be addressed in future work. As noted in the discussion of the empirical results of Primary Aim #1, there may be an additional class of individuals who dichotomize 0-30 open-ended count responses into two discrete categories: 0 for counts in the lower half of the scale, and 30 for counts in the upper half of the scale. This hypothesis could be empirically tested in a follow-up study that experimentally manipulates response formats, in which each participant responds to the same item twice: once as an open-ended count and once as a binned count. For example, the item "During the past 30 days, for about how many days have you felt sad, blue, or depressed?" could be posed with response options $\{0, 1, 2, 3, ..., 30\}$ in one condition and $\{0, 1-5, 6-10, 11-15..., 30\}$ in the other condition. It would be informative to compare the same individual's response for each type of count

response format, paying particular attention to whether people are more likely to select 0 or 30 as responses when the question is framed as an open-ended count than when the question is framed as a binned count.

The current research suggests that latent class IRT models may be useful in analyzing item response data from questionnaires that include count items. While I only fit the proposed models to data from the BRFSS, it is likely that these methods may have utility with other questionnaires. For example, the World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0; World Health Organization, 2010) is a 36-item questionnaire that measures six domains of mental and physical functioning. The 36-item measure, as well as its 12-item short form, includes only Likert-type items, but there are three supplementary items at the end of the questionnaire that elicit counts. In scoring the WHODAS 2.0, these three count items are typically ignored; the online scoring instructions provided by the WHO exclude these three items. Because the current research suggests that including count items on scales can substantially improve measurement precision, future research could focus on adapting the models presented in this dissertation to the WHODAS 2.0, with the goal of obtaining more informative scale scores.

Lastly, future research could investigate new methods of assessing overall model fit for IRT models with count items. While there is a well-known literature on model fit for the generalized linear model with a single count outcome (e.g., Pearson's X^2 statistic and the likelihood ratio G^2 statistic – refer to Agresti (2002) or Bishop, Fienberg, and Holland (1975) for descriptions of fit statistics for GLMs with count outcomes), as well as a substantial literature on model fit for IRT models with binary, ordinal, and nominal item responses (e.g., the M_2 statistic – refer to Maydeu-Olivares (2013) for a review of limited information methods commonly used in IRT), to my knowledge there is no existing method of assessing overall model fit for IRT models with multivariate count outcomes. It is possible that the methods of assessing model fit described in Maydeu-Olivares (2013) can be generalized to multivariate count data; however, as was mentioned in the discussion of model limitations, developing a measure of overall model fit will require circumventing the major challenge of sparseness in the contingency tables – pooling cells, or binning counts, would almost certainly be necessary.

6. 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 models and methods were devised by integrating 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 models developed in this dissertation are 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 I believe that they show promise of wider applicability in the field of psychological measurement.

Appendix A. R Code for Simulations in Primary Aim #1

```
##############################################################################
#### Simulation for Primary Aim #1, Poisson Component for Exact Count Class ####
##############################################################################

############################
#### 4-Class IRT Model ####
############################

set.seed(6014526)
N <- 10000
pnom <- 0.40
ppois <- 0.20
pmin <- 0.30
pmax <- 1 - (pnom + ppois + pmin)
nnom <- pnom*N
npois <- ppois*N
nmin <- pmin*N
nmax <- pmax*N
pi_0 <- pmin
pi_e <- ppois
pi_r <- pnom
pi_30 <- pmax
nitems <- 4
theta <- as.data.frame(rnorm(npois,0,1))

################################
#### Poisson item parameters ####
################################

a1 <- 1.15
a2 <- 1.15
a3 <- 1.30
a4 <- 0.80
c1 <- 1.10
c2 <- 0.00
c3 <- 1.20
c4 <- 1.30

ap <- c(a1, a2, a3, a4)
ap <- as.data.frame(ap)
cp <- c(c1, c2, c3, c4)
cp <- as.data.frame(cp)

#####################################################################
#### Compute Poisson parameters based on above item parameters ####
#####################################################################
```

```
J <- nrow(ap)
lambda <- matrix(-1, npois, J)
for (i in 1:npois){
  for (j in 1:J){
    lambda[i,j] <- exp(ap[j,] %*% theta[i,] + cp[j,])
  }
}

########################################
#### Simulate count data from lambdas ####
########################################

response_p <- matrix(-1, nrow=npois, ncol=J)
for (i in 1:npois){
  for (j in 1:J){
    response_p[i,j] = rpois(1, lambda[i,j])
    if(response_p[i,j] > 30) response_p[i,j] <- 30
  }
}

response_p <- as.data.frame(response_p)
colnames(response_p) <- c("x1", "x2", "x3", "x4")

############################
#### NRM item parameters ####
############################

a11 <- 0
a12 <- 1.00
a13 <- 1.00
a14 <- 1.50
a15 <- 1.50
a16 <- 2.00
a17 <- 1.25

c11 <- 0
c12 <- -0.30
c13 <- -0.75
c14 <- -0.50
c15 <- -0.25
c16 <- -0.80
c17 <- -1.00

a21 <- 0
a22 <- 0.90
a23 <- 1.25
a24 <- 1.50
```

```
a25 <- 1.00
a26 <- 1.40
a27 <- 1.35

c21 <- 0
c22 <- -0.35
c23 <- -0.75
c24 <- -1.25
c25 <- -0.55
c26 <- -0.80
c27 <- -1.20

a31 <- 0
a32 <- 0.85
a33 <- 1.10
a34 <- 1.15
a35 <- 1.85
a36 <- 2.25
a37 <- 1.75

c31 <- 0
c32 <- -0.15
c33 <- -0.25
c34 <- -0.45
c35 <- -0.75
c36 <- -0.90
c37 <- -0.95

a41 <- 0
a42 <- 1.95
a43 <- 1.65
a44 <- 1.00
a45 <- 1.65
a46 <- 0.90
a47 <- 1.25

c41 <- 0
c42 <- -0.75
c43 <- -0.30
c44 <- -1.25
c45 <- -0.80
c46 <- -1.00
c47 <- -1.25

a1 <- cbind(a11, a12, a13, a14, a15, a16, a17)
a2 <- cbind(a21, a22, a23, a24, a25, a26, a27)
a3 <- cbind(a31, a32, a33, a34, a35, a36, a37)
a4 <- cbind(a41, a42, a43, a44, a45, a46, a47)
```

```
a <- rbind(a1,a2,a3,a4)

c1 <- cbind(c11, c12, c13, c14, c15, c16, c17)
c2 <- cbind(c21, c22, c23, c24, c25, c26, c27)
c3 <- cbind(c31, c32, c33, c34, c35, c36, c37)
c4 <- cbind(c41, c42, c43, c44, c45, c46, c47)
c <- rbind(c1,c2,c3,c4)

########################################
#### Generate 4 nominal response items ####
########################################

ncat <- length(c1)
p <- matrix(c(rep(0)), nitems, ncat)
resp <- matrix(c(rep(0)), ncat, nitems)
resp1 <- matrix(c(rep(0)), nnom, nitems)

for (i in 1:nnom){
  set.seed(i)
  theta <- rnorm(1,0,1)
  for (j in 1:nitems){
    for (k in 1:ncat){
      p[j,k] <- exp(a[j,k]*theta + c[j,k])/
        (exp(a[j,1]*theta + c[j,1]) + exp(a[j,2]*theta + c[j,2]) +
           exp(a[j,3]*theta + c[j,3]) + exp(a[j,4]*theta + c[j,4]) +
           exp(a[j,5]*theta + c[j,5]) + exp(a[j,6]*theta + c[j,6]) +
           exp(a[j,7]*theta + c[j,7]))
    }
    resp[,j] <- rmultinom(1,1,p[j,])
    resp[,j] <- ifelse(resp[1,j] == 1, 0,
                       ifelse(resp[2,j] == 1, 5,
                              ifelse(resp[3,j] == 1, 10,
                                     ifelse(resp[4,j] == 1, 15,
                                            ifelse(resp[5,j] == 1, 20,
                                                   ifelse(resp[6,j] == 1, 25,
                                                   ifelse(resp[7,j] == 1, 30, NA)))))))
    resp1[i,] <- resp[1,]
  }
}

response_n <- as.data.frame(resp1)
colnames(response_n) <- c("x1", "x2", "x3", "x4")

################################
#### Generate All-0 People ####
################################

response_0 <- matrix(c(0), nmin, nitems)
```

```
response_0 <- as.data.frame(response_0)
colnames(response_0) <- c("x1", "x2", "x3", "x4")

################################
#### Generate All-30 People ####
################################

response_30 <- matrix(c(30), nmax, nitems)
response_30 <- as.data.frame(response_30)
colnames(response_30) <- c("x1", "x2", "x3", "x4")

########################################
#### Combine Poisson, NRM, 0, 30 data ####
########################################

response <- rbind(response_p, response_n, response_0, response_30)

# 1 if the person must be in the exact class, 0 otherwise

response$e <- ifelse((response$x1 != 0 & response$x1 != 5 & response$x1 != 10 &
                        response$x1 != 15 & response$x1 != 20 &
 response$x1 != 25 & response$x1 != 30) |
                       (response$x2 != 0 & response$x2 != 5 & response$x2 != 10 &
                          response$x2 != 15 & response$x2 != 20 &
 response$x2 != 25 & response$x2 != 30) |
                       (response$x3 != 0 & response$x3 != 5 & response$x3 != 10 &
                          response$x3 != 15 & response$x3 != 20 &
 response$x3 != 25 & response$x3 != 30) |
                       (response$x4 != 0 & response$x4 != 5 & response$x4 != 10 &
                          response$x4 != 15 & response$x4 != 20 &
 response$x4 != 25 & response$x4 != 30),
                     1, 0)

# All Zero Response Pattern

response$all0 <- ifelse(response$x1 == 0 & response$x2 == 0 & response$x3 == 0 &
                          response$x4 == 0, 1, 0)

# 1 if person is NOT in max class, 0 if they could be

response$all30 <- ifelse(response$x1 == 30 & response$x2 == 30 & response$x3 == 30 &
                           response$x4 == 30, 1, 0)

#######################################################
#### Restructure data based on response patterns ####
#######################################################

library(plyr)
```

```
patterns <- ddply(response, ~x1+x2+x3+x4+e+all0+all30, summarize, n=length(x4))
patterns <- as.data.frame(patterns)
colnames(patterns) <- c("x1", "x2", "x3", "x4", "e", "all0", "all30", "r")
x1 <- patterns$x1
x2 <- patterns$x2
x3 <- patterns$x3
x4 <- patterns$x4
x <- cbind(x1,x2,x3,x4)
r <- patterns$r
e <- patterns$e
all0 <- patterns$all0
all30 <- patterns$all30
r_030 <- patterns$r

all0[1] <- r[1]
all30[length(r)] <- r[length(r)]
r_030[1] <- 0
r_030[length(r)] <- 0

n <- sum(r)
data <- data.frame(I(x), r, e, all0, all30, r_030)
theta <- seq(-4,4,0.1)

########################################
#### Normal population distribution ####
########################################

Gaussian.pts <-function(mu,sigma,theta) {
  curve <- exp(-0.5*((theta - mu)/sigma)^2)
  curve <- curve/sum(curve)
}

########################################
#### Function for Poisson trace line ####
########################################

counts <- seq(0,max(x),by=1)
factlist <- factorial(counts)

trace.line.pts.pois <- function(a,c,theta) {
  mat <- matrix(0, nrow=length(a), ncol=length(theta))
  n1 <- length(counts)
  itemtrace <- lapply(seq_len(n1), function(X) mat)
  for (i in 1:length(a)) {
    sumtraces <- rep(0, length(theta))
    for (y in 1:(length(counts)-1)){
      itemtrace[[y]][i,] <- (exp(-exp(a[i]*theta + c[i]))*
                               exp((y-1)*(a[i]*theta + c[i])))/factlist[y]
```

```
      sumtraces <- sumtraces + itemtrace[[y]][i,]
    }
    itemtrace[[length(counts)]][i,] <- 1 - sumtraces
  }
  return(itemtrace)
}

#####################################
#### Function for NRM trace line ####
#####################################

num <- list()
mat <- matrix(0, nrow=nitems, ncol=length(theta))
num <- lapply(seq_len(ncat), function(X) mat)
denom <- matrix(0, nrow=nitems, ncol=length(theta))
itemtrace <- num

trace.line.pts.nrm <- function(a,c,theta) {
  for (j in 1:nitems){
    for (k in 1:ncat){
      num[[k]][j, ] <- exp(a[j,k]*theta + c[j,k])
    }
    denom[j,] <- num[[1]][j,] + num[[2]][j,] + num[[3]][j,] + num[[4]][j,] +
    num[[5]][j,] + num[[6]][j,] + num[[7]][j,]
    itemtrace[[1]][j, ] <- num[[1]][j, ]/denom[j,]
    itemtrace[[2]][j, ] <- num[[2]][j, ]/denom[j,]
    itemtrace[[3]][j, ] <- num[[3]][j, ]/denom[j,]
    itemtrace[[4]][j, ] <- num[[4]][j, ]/denom[j,]
    itemtrace[[5]][j, ] <- num[[5]][j, ]/denom[j,]
    itemtrace[[6]][j, ] <- num[[6]][j, ]/denom[j,]
    itemtrace[[7]][j, ] <- num[[7]][j, ]/denom[j,]
  }
  return(itemtrace)
}

#####################################
#### Function for the likelihood ####
#####################################

ll.poisnrm.ip <- function(p,testdata,theta) {
  nParmsPerNomItem <- 2*(ncat-1)
  nParmsPerPoisItem <- 2
  nParmsPi <- 3
  a <- matrix(rep(0,nitems*(ncat+1)), nitems, ncat+1)
  c <- matrix(rep(0,nitems*(ncat+1)), nitems, ncat+1)
  pi_r <- c(0)
  for (j in 1:nitems) {
    a[j,1] <- 0
```

```
  c[j,1] <- 0
  for (k in 2:ncat){
    a[j,k] <- p[(j-1)*nParmsPerNomItem + (k-1)]
    c[j,k] <- p[(j-1)*nParmsPerNomItem + (ncat-1) + (k-1)]
  }
  a[j,ncat+1] <- p[nitems*nParmsPerNomItem + (2*j-1)]
  c[j,ncat+1] <- p[nitems*nParmsPerNomItem + (2*j-1) + 1]
  zpi_0 <- p[nitems*nParmsPerNomItem + nitems*nParmsPerPoisItem + 1]
  zpi_e <- p[nitems*nParmsPerNomItem + nitems*nParmsPerPoisItem + 2]
  zpi_r <- p[nitems*nParmsPerNomItem + nitems*nParmsPerPoisItem + 3]
  pi_0 <- exp(zpi_0)/(1 + exp(zpi_0) + exp(zpi_e) + exp(zpi_r))
  pi_e <- exp(zpi_e)/(1 + exp(zpi_0) + exp(zpi_e) + exp(zpi_r))
  pi_r <- exp(zpi_r)/(1 + exp(zpi_0) + exp(zpi_e) + exp(zpi_r))
}

itemtrace_nrm <- trace.line.pts.nrm(a[,1:ncat],c[,1:ncat],theta)
itemtrace_pois <- trace.line.pts.pois(a[,ncat+1],c[,ncat+1],theta)
expected_nrm <- rep(0,length(testdata$r))
expected_pois <- rep(0, length(testdata$r))
for (i in 1:length(testdata$r)) {
  if (testdata$e[i]) {
    posterior_nrm <- 0
  } else {
    posterior_nrm <- Gaussian.pts(0,1,theta)
    for (item in 1:ncol(testdata$x)) {
      x <- I(testdata$x[i,item])
      if (x == 0)
        posterior_nrm <- posterior_nrm*itemtrace_nrm[[1]][item,]
      if (x == 5)
        posterior_nrm <- posterior_nrm*itemtrace_nrm[[2]][item,]
      if (x == 10)
        posterior_nrm <- posterior_nrm*itemtrace_nrm[[3]][item,]
      if (x == 15)
        posterior_nrm <- posterior_nrm*itemtrace_nrm[[4]][item,]
      if (x == 20)
        posterior_nrm <- posterior_nrm*itemtrace_nrm[[5]][item,]
      if (x == 25)
        posterior_nrm <- posterior_nrm*itemtrace_nrm[[6]][item,]
      if (x == 30)
        posterior_nrm <- posterior_nrm*itemtrace_nrm[[7]][item,]
    }
    expected_nrm[i] <- sum(posterior_nrm)
  }
}
for (i in 1:length(testdata$r)) {
  posterior_pois <- Gaussian.pts(0,1,theta)
  for (item in 1:ncol(testdata$x)) {
    x <- I(testdata$x[i,item])
```

```
      posterior_pois <- posterior_pois*itemtrace_pois[[x+1]][item,]
    }
    expected_pois[[i]] <- sum(posterior_pois)
  }
  l <- (-1)*(sum(testdata$all0*log(pi_0 + (1-pi_0-pi_e-pi_r)*expected_pois +
   pi_r*expected_nrm)) + sum(testdata$all30*log((1-pi_0-pi_e-pi_r) +
   pi_e*expected_pois + pi_r*expected_nrm)) +
   sum(testdata$r_030*(log(pi_r*expected_nrm + pi_e*expected_pois))))
}

acoefs_nom <- matrix(c(rep(2)), nitems, ncat-1)
ccoefs_nom <- matrix(c(rep(-2)), nitems, ncat-1)
a_pois <- matrix(c(rep(1)), nitems, 1)
c_pois <- matrix(c(rep(0)), nitems, 1)
acoefs <- cbind(acoefs_nom, a_pois)
ccoefs <- cbind(ccoefs_nom, c_pois)
zpi_0 <- 2)
zpi_e <- 2
zpi_r <- 2
nParmsPerNomItem <- 2*(ncat-1)
nParmsPerPoisItem <- 2
nParmsPi <- 3
p <- rep(0, 2*(nitems*(ncat-1+1))+nParmsPi)
for (j in 1:nitems) {
  for (k in 1:(ncat-1)){
    p[(j-1)*nParmsPerNomItem + k] <- acoefs[j,k]
    p[(j-1)*nParmsPerNomItem + (ncat-1) + k] <- ccoefs[j,k]
  }
  p[nitems*nParmsPerNomItem + (2*j-1)] <- acoefs[j,ncat]
  p[nitems*nParmsPerNomItem + (2*j-1) + 1] <- ccoefs[j,ncat]
  p[nitems*nParmsPerNomItem + nitems*nParmsPerPoisItem + 1] <- zpi_0
  p[nitems*nParmsPerNomItem + nitems*nParmsPerPoisItem + 2] <- zpi_e
  p[nitems*nParmsPerNomItem + nitems*nParmsPerPoisItem + 3] <- zpi_r
}

system.time(result <- nlm(f=ll.poisnrm.ip,p=p,hessian=TRUE,
                          testdata=data,theta=theta,print.level=2,iterlim=1000))
```

Appendix B. R Code for Simulations in Secondary Aim

```
##############################################################################
#### Simulation for Secondary Aim, Neg Bin Component for Exact Count Class ####
##############################################################################

set.seed(6014526)
N <- 10000
pnom <- 0.24
pnegbin <- 0.50
pmin <- 0.25
pmax <- 1 - (pnom + pnegbin + pmin)
nnom <- pnom*N
nnegbin <- pnegbin*N
nmin <- pmin*N
nmax <- pmax*N
pi_0 <- pmin
pi_e <- pnegbin
pi_r <- pnom
pi_30 <- pmax
nitems <- 7
theta <- as.data.frame(rnorm(nnegbin,0,1))

#################################
#### Neg Bin item parameters ####
#################################

a1 <- 1.15
c1 <- 1.10
dispnb <- 2.5

anb <- as.data.frame(a1)
cnb <- as.data.frame(c1)

#####################################################################
#### Compute Poisson parameters based on above item parameters ####
#####################################################################

J <- nrow(anb)
lambda <- matrix(-1, nnegbin, J)
for (i in 1:nnegbin){
  for (j in 1:J){
    lambda[i,j] <- exp(anb[j,] %*% theta[i,] + cnb[j,])
  }
}
```

```
#####################################################
#### Simulate count data from simulated lambdas ####
#####################################################

response_p <- matrix(-1, nrow=nnegbin, ncol=J)
for (i in 1:nnegbin){
  for (j in 1:J){
    response_p[i,j] = rnbinom(1, size=dispnb[j], mu=lambda[i,j])
    if(response_p[i,j] > 30) response_p[i,j] <- 30
  }
}

response_p <- as.data.frame(response_p)
colnames(response_p) <- c("x7")

#############################
#### NRM item parameters ####
#############################

a11 <- 0
a12 <- 1.00
a13 <- 1.00
a14 <- 1.50
a15 <- 1.50
a16 <- 2.00
a17 <- 1.25

c11 <- 0
c12 <- -0.30
c13 <- -0.75
c14 <- -0.50
c15 <- -0.25
c16 <- -0.80
c17 <- -1.00

a <- cbind(a11, a12, a13, a14, a15, a16, a17)
c <- cbind(c11, c12, c13, c14, c15, c16, c17)

#########################################
#### Generate 4 nominal response items ####
#########################################

ncatnom <- length(c)
nitemsnom <- nrow(c)

p <- matrix(c(rep(0)), nitemsnom, ncatnom)
resp <- matrix(c(rep(0)), ncatnom, nitemsnom)
resp1 <- matrix(c(rep(0)), nnom, nitemsnom)
```

```
for (i in 1:nnom){
  set.seed(i)
  theta <- rnorm(1,0,1)
  for (j in 1:nitemsnom){
    for (k in 1:ncatnom){
      p[j,k] <- exp(a[j,k]*theta + c[j,k])/
        (exp(a[j,1]*theta + c[j,1]) + exp(a[j,2]*theta + c[j,2]) +
           exp(a[j,3]*theta + c[j,3]) + exp(a[j,4]*theta + c[j,4]) +
           exp(a[j,5]*theta + c[j,5]) + exp(a[j,6]*theta + c[j,6]) +
           exp(a[j,7]*theta + c[j,7]))
    }
    resp[,j] <- rmultinom(1,1,p[j,])
    resp[,j] <- ifelse(resp[1,j] == 1, 0,
                       ifelse(resp[2,j] == 1, 5,
                              ifelse(resp[3,j] == 1, 10,
                                     ifelse(resp[4,j] == 1, 15,
                                            ifelse(resp[5,j] == 1, 20,
                                            ifelse(resp[6,j] == 1, 25,
                                            ifelse(resp[7,j] == 1, 30, NA)))))))
    resp1[i,] <- resp[1,]
  }
}

response_n <- as.data.frame(resp1)
colnames(response_n) <- c("x7")

################################
#### Generate All-0 People ####
################################

response_0 <- matrix(c(0), nmin, nitemsnom)
response_0 <- as.data.frame(response_0)
colnames(response_0) <- c("x7")

################################
#### Generate All-30 People ####
################################

response_30 <- matrix(c(30), nmax, nitemsnom)
response_30 <- as.data.frame(response_30)
colnames(response_30) <- c("x7")

########################################
#### Combine Neg Bin, NRM, 0, 30 data ####
########################################

response <- rbind(response_p, response_n, response_0, response_30)
```

```
table(response$x7)

################################
#### Generate Likert items ####
################################

nobs <- nnegbin + nnom
nitemsgrm <- 6
ncatgrm <- 5

a1 <- 1.75
c1 <- c(1, 0.25, -2, -4)
a2 <- 2.65
c2 <- c(0.5, -0.5, -0.8, -1.5)
a3 <- 1.5
c3 <- c(1.5, 0.2, -1.2, -3.4)
a4 <- 2.30
c4 <- c(-1, -2, -3, -4)
a5 <- 2.2
c5 <- c(1.5, 0.5, -0.4, -2)
a6 <- 3.2
c6 <- c(2, 0.75, 0.10, -0.75)

a <- c(a1, a2, a3, a4, a5, a6)
a <- matrix(a, nitemsgrm, length(a1), byrow=T)
c <- rbind(c1, c2, c3, c4, c5, c6)

GradedGen <- function(nobs,a,c){
  a <- as.matrix(a)
  c <- as.matrix(c)
  J <- nrow(c)
  K <- ncol(c)+1
  theta <- rnorm(nobs, mean=0, sd=1)
  response <- matrix(-1, nrow=nobs, ncol=J)
  for (i in 1:nobs){
    for (j in 1:J){
      temp <- runif(1,0,1);
      for (k in 1:(K-1)){
        if (temp > 1/(1+exp(-a[j,]%*%theta[i]-c[j,k]))){
          response[i,j] <- k-1
          break
        }
        else
          next
      }
      if (response[i,j] == -1)
        response[i,j] <- K-1
    }
```

```
  }
  return(list(response=response))
}

graded <- GradedGen(nobs, a, c)
graded <- as.data.frame(graded)
response_g <- graded
colnames(response_g) <- c("x1", "x2", "x3", "x4", "x5", "x6")

############################
#### Generate Min People ####
############################

response_0g <- matrix(c(0), nmin, nitemsgrm)
response_0g <- as.data.frame(response_0g)
colnames(response_0g) <- c("x1", "x2", "x3", "x4", "x5", "x6")

############################
#### Generate Max People ####
############################

response_30g <- matrix(c(4), nmax, nitemsgrm)
response_30g <- as.data.frame(response_30g)
colnames(response_30g) <- c("x1", "x2", "x3", "x4", "x5", "x6")

########################################
#### Combine Neg Bin, NRM, 0, 30 data ####
########################################

graded_all <- rbind(response_g, response_0g, response_30g)
response_all <- cbind(graded_all, response$x7)
response <- response_all
colnames(response) <- c("x1", "x2", "x3", "x4", "x5", "x6", "x7")

# Exact Class

response$e <- ifelse((response$x7 != 0 & response$x7 != 5 & response$x7 != 10 &
response$x7 != 15 & response$x7 != 20 &
response$x7 != 25 & response$x7 != 30), 1, 0)

# All Zero Response Pattern

response$all0 <- ifelse(response$x1 == 0 & response$x2 == 0 & response$x3 == 0 &
response$x4 == 0 & response$x5 == 0 &
 response$x6 == 0 & response$x7 == 0, 1, 0)

# 1 if person is NOT in max class, 0 if they could be
```

```
response$all30 <- ifelse(response$x1 == 4 & response$x2 == 4 & response$x3 == 4 &
response$x4 == 4 & response$x5 == 4 & response$x6 == 4 &
response$x7 == 30, 1, 0)

library(plyr)
patterns <- ddply(response, ~x1+x2+x3+x4+x5+x6+x7+e+all0+all30,
summarize, n=length(x4))
patterns <- as.data.frame(patterns)
colnames(patterns) <- c("x1", "x2", "x3", "x4", "x5", "x6",
"x7", "e", "all0", "all30", "r")
x1 <- patterns$x1
x2 <- patterns$x2
x3 <- patterns$x3
x4 <- patterns$x4
x5 <- patterns$x5
x6 <- patterns$x6
x7 <- patterns$x7
x <- cbind(x1,x2,x3,x4,x5,x6,x7)
r <- patterns$r
e <- patterns$e
all0 <- patterns$all0
all30 <- patterns$all30
r_030 <- patterns$r

all0[1] <- r[1]
all30[length(r)] <- r[length(r)]
r_030[1] <- 0
r_030[length(r)] <- 0

n <- sum(r)
data <- data.frame(I(x), r, e, all0, all30, r_030)
theta <- seq(-4,4,0.1)

########################################
#### Normal population distribution ####
########################################

Gaussian.pts <-function(mu,sigma,theta) {
  curve <- exp(-0.5*((theta - mu)/sigma)^2)
  curve <- curve/sum(curve)
}

########################################
#### Function for Neg Bin trace line ####
########################################

counts <- seq(0,max(x),by=1)
factlist <- factorial(counts)
```

```
trace.line.pts.negbin <- function(a,c,disp,theta) {
  mat <- matrix(0, nrow=length(a), ncol=length(theta))
  n1 <- length(counts)
  itemtrace <- lapply(seq_len(n1), function(X) mat)
  for (i in 1:length(a)) {
    sumtraces <- rep(0, length(theta))
    for (y in 1:(length(counts)-1)){
      itemtrace[[y]][i,] <-
        (gamma((y-1) + disp[i]^(-1))/(factlist[y]*gamma(disp[i]^(-1)))) *
        ((disp[i]^(-1))/(disp[i]^(-1) + exp(a[i]*theta + c[i])))^(disp[i]^(-1)) *
        ((exp(a[i]*theta + c[i]))/(disp[i]^(-1) + exp(a[i]*theta + c[i])))^(y-1)
      sumtraces <- sumtraces + itemtrace[[y]][i,]
    }
    itemtrace[[length(counts)]][i,] <- 1 - sumtraces
  }
  return(itemtrace)
}

#####################################
#### Function for NRM trace line ####
#####################################

num <- list()
mat <- matrix(0, nrow=nitemsnom, ncol=length(theta))
num <- lapply(seq_len(ncatnom), function(X) mat)
denom <- matrix(0, nrow=nitemsnom, ncol=length(theta))
itemtrace <- num

trace.line.pts.nrm <- function(a,c,theta) {
  itemtrace <- num
  for (j in 1:nitemsnom){
    for (k in 1:ncatnom){
      num[[k]][j, ] <- exp(a[j,k]*theta + c[j,k])
    }
    denom[j,] <- num[[1]][j,] + num[[2]][j,] + num[[3]][j,] + num[[4]][j,] +
    num[[5]][j,] + num[[6]][j,] + num[[7]][j,]
    itemtrace[[1]][j, ] <- num[[1]][j, ]/denom[j,]
    itemtrace[[2]][j, ] <- num[[2]][j, ]/denom[j,]
    itemtrace[[3]][j, ] <- num[[3]][j, ]/denom[j,]
    itemtrace[[4]][j, ] <- num[[4]][j, ]/denom[j,]
    itemtrace[[5]][j, ] <- num[[5]][j, ]/denom[j,]
    itemtrace[[6]][j, ] <- num[[6]][j, ]/denom[j,]
    itemtrace[[7]][j, ] <- num[[7]][j, ]/denom[j,]
  }
  return(itemtrace)
}
```

```
#####################################
#### Function for GRM trace line ####
#####################################

mat2 <- matrix(0, nrow=nitemsgrm, ncol=length(theta))
itemtrace_grm <- lapply(seq_len(ncatgrm), function(X) mat2)

trace.line.pts.grm <- function(a,c,theta) {
  for (j in 1:nitemsgrm){
    for (k in 0:ncatgrm-1){
      if (k == 0){
        itemtrace_grm[[k+1]][j,] <- 1 - (exp(a[j,]*theta +
c[j,k+1])/(1+exp(a[j,]*theta + c[j,k+1])))
      }
      if (k == 1){
        itemtrace_grm[[k+1]][j,] <- exp(a[j,]*theta + c[j,k])/(1+exp(a[j,]*theta +
c[j,k])) - exp(a[j,]*theta + c[j,k+1])/(1+exp(a[j,]*theta + c[j,k+1]))
      }
      if (k == 2){
        itemtrace_grm[[k+1]][j,] <- exp(a[j,]*theta + c[j,k])/(1+exp(a[j,]*theta +
c[j,k])) - exp(a[j,]*theta + c[j,k+1])/(1+exp(a[j,]*theta + c[j,k+1]))
      }
      if (k == 3){
        itemtrace_grm[[k+1]][j,] <- exp(a[j,]*theta + c[j,k])/(1+exp(a[j,]*theta +
c[j,k])) - exp(a[j,]*theta + c[j,k+1])/(1+exp(a[j,]*theta + c[j,k+1]))
      }
      if (k == 4){
        itemtrace_grm[[k+1]][j,] <- exp(a[j,]*theta + c[j,k])/(1+exp(a[j,]*theta +
c[j,k]))
      }
    }
  }
  return(itemtrace_grm)
}

#####################################
#### Function for the likelihood ####
#####################################

ll.negbinnrmgrm.ip <- function(p,testdata,theta) {
  nParmsPerNomItem <- 2*(ncatnom-1)
  nParmsPerNegBinItem <- 3
  nParmsPerGradedItem <- 5
  nParmsPi <- 3
  a_nom <- matrix(c(1), nitemsnom, ncatnom)
  c_nom <- matrix(c(0), nitemsnom, ncatnom)
  a_negbin <- matrix(c(1), nitemsnom, 1)
  c_negbin <- matrix(c(0), nitemsnom, 1)
```

```
disp_negbin <- matrix(c(0.5), nitemsnom, 1)
a_grm <- matrix(c(rep(2, nitemsgrm)), nitemsgrm, 1)
c_grm <- matrix(c(2, 1, 0, -1,
                  1, 0, -1, -2,
                  1, 0.5, -0.5, -1,
                  2.5, 1.5, 1, 0,
                  -1, -2, -2.5, -3,
                  0, -0.5, -1, -1.5), nitemsgrm, ncatgrm-1, byrow=T)
for (j in 1:nitemsnom) {
  a_nom[j,1] <- 0
  c_nom[j,1] <- 0
  for (k in 2:ncatnom){
    a_nom[j,k] <- p[(j-1) + (2*k) - 3]
    c_nom[j,k] <- p[(j-1) + (2*k) - 2]
  }
  a_negbin[j,1] <- p[(j-1) + nParmsPerNomItem + 1]
  c_negbin[j,1] <- p[(j-1) + nParmsPerNomItem + 2]
  disp_negbin[j,1] <- p[(j-1) + nParmsPerNomItem + 3]
  for (j in 1:nitemsgrm){
    for (k in 1:(ncatgrm-1)){
      a_grm[j,1] <- p[nParmsPerNomItem + nParmsPerNegBinItem + (5*j) - 4]
      c_grm[j,k] <- p[nParmsPerNomItem + nParmsPerNegBinItem + (5*j) - 4 + k]
    }
  }
  zpi_0 <- p[nParmsPerNomItem + nParmsPerNegBinItem +
  nitemsgrm*nParmsPerGradedItem + 1]
  zpi_e <- p[nParmsPerNomItem + nParmsPerNegBinItem +
  nitemsgrm*nParmsPerGradedItem + 2]
  zpi_r <- p[nParmsPerNomItem + nParmsPerNegBinItem +
  nitemsgrm*nParmsPerGradedItem + 3]
  pi_0 <- exp(zpi_0)/(1 + exp(zpi_0) + exp(zpi_e) + exp(zpi_r))
  pi_e <- exp(zpi_e)/(1 + exp(zpi_0) + exp(zpi_e) + exp(zpi_r))
  pi_r <- exp(zpi_r)/(1 + exp(zpi_0) + exp(zpi_e) + exp(zpi_r))
}
itemtrace_nrm <- trace.line.pts.nrm(a_nom, c_nom, theta)
itemtrace_negbin <- trace.line.pts.negbin(a_negbin, c_negbin, disp_negbin, theta)
itemtrace_grm <- trace.line.pts.grm(a_grm, c_grm, theta)
expected_nrm <- rep(0,length(testdata$r))
expected_negbin <- rep(0, length(testdata$r))
expected_grm <- rep(0, length(testdata$r))
for (i in 1:length(testdata$r)) {
  if (testdata$e[i]) {
    posterior_nrm <- 0
  } else {
    posterior_nrm <- Gaussian.pts(0,1,theta)
    for (item in 1:1) {
      x <- I(testdata$x[i,7])
      if (x == 0)
```

```
          posterior_nrm <- posterior_nrm*itemtrace_nrm[[1]][item,]
        if (x == 5)
          posterior_nrm <- posterior_nrm*itemtrace_nrm[[2]][item,]
        if (x == 10)
          posterior_nrm <- posterior_nrm*itemtrace_nrm[[3]][item,]
        if (x == 15)
          posterior_nrm <- posterior_nrm*itemtrace_nrm[[4]][item,]
        if (x == 20)
          posterior_nrm <- posterior_nrm*itemtrace_nrm[[5]][item,]
        if (x == 25)
          posterior_nrm <- posterior_nrm*itemtrace_nrm[[6]][item,]
        if (x == 30)
          posterior_nrm <- posterior_nrm*itemtrace_nrm[[7]][item,]
      }
      expected_nrm[i] <- sum(posterior_nrm)
    }
  }
  for (i in 1:length(testdata$r)) {
    posterior_negbin <- Gaussian.pts(0,1,theta)
    for (item in 1:1) {
      x <- I(testdata$x[i,7])
      posterior_negbin <- posterior_negbin*itemtrace_negbin[[x+1]][item,]
    }
    expected_negbin[i] <- sum(posterior_negbin)
  }
  for (i in 1:length(testdata$r)) {
    posterior_grm <- Gaussian.pts(0,1,theta)
    for (item in 1:6){
      x <- I(testdata$x[i,item])
      posterior_grm <- posterior_grm*itemtrace_grm[[x+1]][item,]
    }
    expected_grm[[i]] <- sum(posterior_grm)
  }
 l <- (-1)*(sum(testdata$all0*log(pi_0 + pi_e*expected_negbin*expected_grm +
 pi_r*expected_nrm*expected_grm)) + sum(testdata$all30*log((1-pi_0-pi_e-pi_r) +
 pi_e*expected_negbin*expected_grm + pi_r*expected_nrm*expected_grm)) +
 sum(testdata$r_030*(log(pi_r*expected_nrm*expected_grm +
 pi_e*expected_negbin*expected_grm))))
}

acoefs_nom <- matrix(c(1), nitemsnom, ncatnom-1)
ccoefs_nom <- matrix(c(-1), nitemsnom, ncatnom-1)
a_negbin <- matrix(c(1), nitemsnom, 1)
c_negbin <- matrix(c(0.5), nitemsnom, 1)
disp_negbin <- matrix(c(0.4), 1, 1)
a_grm <- matrix(c(rep(2, nitemsgrm)), nitemsgrm, 1)
c_grm <- matrix(c(2, 1, 0, -1,
                  1, 0, -1, -2,
```

```
                  1, 0.5, -0.5, -1,
                  2.5, 1.5, 1, 0,
                  -1, -2, -2.5, -3,
                  0, -0.5, -1, -1.5), nitemsgrm, ncatgrm-1, byrow=T)
zpi_0 <- 2
zpi_e <- 3
zpi_r <- 2
nParmsPerNomItem <- 2*(ncatnom-1)
nParmsPerNegBinItem <- 3
nParmsPerGradedItem <- 5
nParmsPi <- 3
p <- rep(0, nParmsPerNomItem*1 + nParmsPerNegBinItem*1 +
nParmsPerGradedItem*6 + nParmsPi)
for (j in 1:nitemsnom) {
  for (k in 1:(ncatnom-1)){
    p[(j-1) + (2*k) - 1] <- acoefs_nom[j,k]
    p[(j-1) + 2*k] <- ccoefs_nom[j,k]
  }
  p[(j-1) + nParmsPerNomItem + 1] <- a_negbin[j,1]
  p[(j-1) + nParmsPerNomItem + 2] <- c_negbin[j,1]
  p[(j-1) + nParmsPerNomItem + 3] <- disp_negbin[j,1]
}
for (j in 1:nitemsgrm){
  for (k in 1:(ncatgrm-1)){
    p[nParmsPerNomItem + nParmsPerNegBinItem + (5*j) - 4] <- a_grm[j,1]
    p[nParmsPerNomItem + nParmsPerNegBinItem + (5*j) - 4 + k] <- c_grm[j,k]
  }
  p[nParmsPerNomItem + nParmsPerNegBinItem +
  nitemsgrm*nParmsPerGradedItem + 1] <- zpi_0
  p[nParmsPerNomItem + nParmsPerNegBinItem +
  nitemsgrm*nParmsPerGradedItem + 2] <- zpi_e
  p[nParmsPerNomItem + nParmsPerNegBinItem +
  nitemsgrm*nParmsPerGradedItem + 3] <- zpi_r
}

system.time(result <- nlm(f=ll.negbinnrmgrm.ip,p=p,hessian=TRUE,
                          testdata=data,theta=theta,print.level=2,iterlim=1000))
```

References

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

Bartholomew, D. J., & Tzamourani, P. (1999). The goodness of fit of latent trait models in attitude measurement. *Sociological Methods and Research*, *27*, 525-546. doi: 10.1177/0049124199027004003

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

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

Bock, R. D. (1997). The nominal categories model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (p. 33-50). New York: Springer.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*(4), 665-678. doi: 10.1037/a0028111

Bockenholt, 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., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics*, *26*, 381-409. doi: 10.3102/10769986026004381

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Applications of a Mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*, 331-348. doi: 10.1111/j.1745-3984.2002.tb01146.x

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*(5), 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 (CDC). (1984-present). *Behavioral Risk Factor Surveillance System Survey Data*. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention.

Centers for Disease Control and Prevention (CDC). (1991-present). *Youth Risk Behavior Survey*.

Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). (1971-present). *National Health and Nutrition Examination Survey Questionnaire.* Hyattsville, MD: US Department of Health and Human Services, Centers for Disease Control and Prevention.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*, 133-148. doi: 10.1111/j.1745-3984.2005.00007

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

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

Dunson, D. B., & Herring, A. H. (2005). Bayesian latent variable models for mixed discrete outcomes. *Biostatistics*, *6*(1), 11-25. doi: 10.1093/biostatistics/kxh025

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

Fromme, K., Katz, E. C., & Rivet, K. (1997). Outcome expectancies and risk-taking behavior. *Cognitive Therapy and Research*, *21*(4), 421-442. doi: 10.1023/A:1021932326716

Hagenaars, J., & McCuthcheon, A. (2002). *Applied latent class analysis.* Cambridge, UK: 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*(410), 304-314. doi: 10.1080/01621459.1990.10476202

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

Jin, K., & Wang, W. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement*, *74*(1), 116-138. doi: 10.1177/0013164413498876

Katz, E., Fromme, K., & D'Amico, E. (2000). Effects of outcome expectancies and personality on young adults' illicit drug use, heavy drinking, and risky sexual behavior. *Cognitive Therapy and Research*, *24*(1), 1-22. doi: 10.1023/A:1005460107337

Klonsky, E. D., & Glenn, C. R. (2008). Psychosocial risk and protective factors. In M. K. Nixon & N. Heath (Eds.), *Self-injury in youth: The essential guide to assessment and intervention.* New York, NY: Routledge.

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

Lazarsfeld, P. F. (1959). Latent structure analysis. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 3). New York, NY: McGraw-Hill.

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*(2), 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

Maij-de Meij, A., Kelderman, H., & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement*, *32*(8), 611-631. doi: 10.1177/0146621607312613

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement*, *11*(3), 71-101. doi: 10.1080/15366367.2013.831680

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

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*(2), 195-215. doi: 10.1007/BF02295283

Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality and Quantity*, *42*, 779-794. doi: 0.1007/s11135-006-9067-x

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

Nock, M. K., Holmberg, E. B., Photos, V. I., & Michel, B. D. (2007). Self-Injurious Thoughts and Behaviors Interview: Development, reliability, and validity in an adolescent sample. *Psychological Assessment*, *19*(3), 309-317. doi: 10.1037/1040-3590.19.3.309

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Sage Publications, Inc.

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*(7), 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 W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (p. 449-463). New York: Springer.

Rost, J., Cartensen, C., & Von Davier, M. (1997). Applying the mixed rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (p. 324-332). Munster, Germany: Waxmann.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*(17). doi: 10.1002/j.2333-8504.1968.tb00153.x

Samuelson, K. M. (2008). Examining differential item functioning from a latent class perspective. In G. R. Hancock & K. M. Samuelson (Eds.), *Advances in latent variable mixture models* (p. 67-113). Charlotte, NC: Information Age Publishing.

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

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*. Chapman & Hall/CRC.

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

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567-577. doi: 10.1007/BF02295596

Thissen, D., & Wainer, H. (2001). *Test scoring*. Routledge.

Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to Likert-type items. *Journal of Educational and Behavioral Statistics*, *38*(5), 522-547. doi: 10.3102/1076998613481500

von Davier, M., & Rost, J. (2006). Mixture distribution item response models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics, Vol. 26: Psychometrics* (p. 643-661). Amsterdam: 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*(8), 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*(6), 671-692. doi: 10.3102/1076998610375838

Ward, J., Darke, S., & Hall, W. (1990). *The HIV Risk-Taking Behaviour Scale Manual* (Tech. Rep. No. 10). National Drug and Alcohol Research Centre.

Wedel, M., Bockenholt, 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

World Health Organization. (1998-present). *WHO Psychiatric Disability Assessment Schedule.* Geneva, WHO.

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

Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langenheine (Eds.), *Applications of latent trait and latent class models in the social sciences.* New York, NY: Waxmann.