THE ORDINAL COUNT FACTOR MODEL:
AN IMPROVED LATENT VARIABLE MODEL FOR ORDINAL COUNT ITEMS

Nathan David Markiewitz

A thesis submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Psychology and Neuroscience in the College of Arts and Sciences.

Chapel Hill
2017

Approved by:

Daniel J. Bauer

Kenneth A. Bollen

Patrick J. Curran

ABSTRACT

Nathan David Markiewitz: The Ordinal Count Factor Model: An Improved Latent Variable
Model for Ordinal Count Items
(Under the direction of Daniel J. Bauer)

Much of the measurement of human behaviors relies on the reporting of a rate of behavior. Common measures use items that ask participants to select between given intervals of counts—these items are called ordinal count items.

I present the ordinal count factor model (OCFM) as a latent variable model for ordinal count item responses in a single population and across multiple groups. OCFMs represent the underlying latent response as a count, instead of the logistic or normal distribution used by current latent variable models for ordinal data. In addition to representing the data generating process more faithfully, OCFMs allow for inferences on the metric of the underlying rate of behavior.

I evaluate the OCFM through two empirical examples using the Rutgers Alcohol Problem Index. These studies demonstrate that OCFMs may fit better than standard models, produce more precise factor scores, and may be fit using widely available, open-source software.

To Mom, Dad, Aaron, Sam, Dan, and Thunder

ACKNOWLEDGEMENTS

This work is what it is because of the tireless mentorship of Daniel Bauer. Through his coaching, I learned not only this method's potential but also my own. I also deeply appreciate the guidance provided by my other committee members, Kenneth Bollen and Patrick Curran, as well as Andrea Hussong. Their insights and advice taught me what quantitative psychology can be. I would be lost, however, without the indefatigable support of the other students in the L. L. Thurstone Psychometric Lab. Thank you for the coffee breaks, the R tutorials, the chats when the models were running. You make work fun. But above all, I would not be here without the endless support of my friends and family. Thank you for suffering terrible puns and reminding me that life is wider than science. Wherever I am, I know that I can count on you for anything.

TABLE OF CONTENTS

LIST OF TABLES ……………………………..……………..……………..………………..viii

LIST OF FIGURES………………………….……………..……………..………..……..……ix

INTRODUCTION ……………………..………………..……………..……………..………...1

Motivating Example ...…………………....………………..……………..…………6

The Rutgers Alcohol Problem Index…………………….…………………..…………7

The Real Experiences and Lives in the University Study ……………………..………..7

ORDINAL COUNT FACTOR MODELS IN A SINGLE POPULATION ……………..………9

The General Ordinal Count Factor Model ……………………..………………..………..9

The Ordinal Negative Binomial Factor Model ……………………..………..………..12

The Ordinal Beta Binomial Factor Model …………...……………….………..………14

Zero-Inflated Ordinal Count Factor Models …………...……………..…………...…..16

Identification of Ordinal Count Factor Models....……...……………..…………...…..17

Estimation of Ordinal Count Factor Models ....……...……………..…………...…..18

Feasibility Study...……...……………..…………...…..19

Study 1..……...……………..…………...… 22

Data...……....…...………..……………...…23

Method......................................……...………………...…………...…23

Results......................................……...………………...…………...…24

Discussion................................……...………………...…………...…33

ORDINAL COUNT FACTOR MODELS IN MULTIPLE POPULATIONS ....………...…36

Multiple Groups Ordinal Count Factor Model.......................………………………...…39

Specification for Multiple Groups............................………………………...…39

Identification for Multiple Groups ....………...…………...…40

Estimation for Multiple Groups................................………………………...…42

Study 2 ....………...………...………...…………………………...…43

Methods............................…………...………...……………...…………...…45

Results.............................… ..……………………...…...……...…………...…49

Discussion............................……….. ……………...…………………...…55

CONCLUSION.....................……………………...……….........………...…………...…58

APPENDIX A: REAL-U CROSSWALK FOR THE RAPI (FROM REAL-U CODEBOOK)....61

APPENDIX B: OCFMS IN A SINGLE POPULATION ...…………………………...……...…63

APPENDIX C: ORDINAL COUNT FACTOR MODELS IN MULTIPLE GROUPS.......…...…66

APPENDIX D: SCORE CORRELATIONS FOR ALL FOUR
MEASUREMENT CONDITIONS .....………………........………...…………………………71

REFERENCES ....………...……………………………………...……...…73

LIST OF TABLES

Table 1: Mean Relative Bias by Parameter and Model in Feasibility Study……………….……….20

Table 2: Linear CFA Item Parameter Estimates……………………………………….….…...25

Table 3: Graded Response Model Parameter Estimates……………………………….……....26

Table 4: Ordinal Count Factor Model Parameter Estimates………………………….…......27

Table 5: Factor Score Correlations…………………………………………….….….…...30

Table 6: Response Options by Scenario……………………………………………..…...44

Table 7: Multiple Group Ordinal Count Factor Model Estimates…………………………..…48

Table 8: Multiple Group Graded Response Model Parameter Estimates……………………..…49

Table 9: Multiple Group Factor Score Correlations Pooled over Scenarios………………..…50

Table 10: Multiple Group Factor Score Correlations for Participants
Who Received Scenario 1 Twice. ……………………………………..………………69

Table 11: Multiple Group Factor Score Correlations for Participants
Who Received Scenario 1 then Scenario 4…………………......………………..…………...…69

Table 12: Multiple Group Factor Score Correlations for Participants
Who Received Scenario 4 then Scenario 1…………………...…………………………......…70

Table 13: Multiple Group Factor Score Correlations for Participants
Who Received Scenario 4 Twice…………………..………..…………….……………......…70

LIST OF FIGURES

Figure 1: Parameter recovery plots for the four OCFMs described………………………………19

Figure 2: Plots of EAPs across models……………………………………………………...29

Figure 3: Plots of the standard errors of the GRM and the OCFM by their EAPs………..………..31

Figure 4: The difference between a participant's OCFM and GRM
standard errors plotted against their OCFM EAP……………………………….…………..31

Figure 5: Standard errors plotted against the EAP from the DIF-corrected OCFM………………52

Introduction

Psychology, education, and the allied health and social sciences regularly seek to explain the presence and frequency of behaviors through the constructs that underlie them. To this end, our models often include counts of behaviors—symptoms, correct responses, or peer interactions—that are related to an underlying construct of interest—disorder, ability, or social support. Thus measurement models in the behavioral sciences necessarily rely on observations of behavior.

Yet to observe a behavior is not as easy as it sounds, motivating the use of self-reports. Some behavioral phenomena may simply be impossible to observe directly, like cognition. Ethical concerns may render other behaviors impossible to observe directly without compelling the observer to intervene, such as substance abuse. Financial concerns also compel researchers to consider indirect measures of behavior. Perhaps most importantly, sometimes observing a behavior affects it, threatening internal validity. For these reasons and more, social and behavioral scientists often ask participants to report their behavior.

Many research questions concern the rate of a behavior, and the problems with self-reported rates are well documented. A researcher could instruct participants to report whether or not they have engaged in that behavior, to report a subjective rate of that behavior (e.g. never, sometimes, rarely), or to report the exact number, or raw count, of times they engaged in a behavior.[1] Unfortunately, self-reported raw counts of behavior often prove problematic to

[1] Of course, this exact number can only be reported if the behavior is discrete.

analyze in psychological research. In addition to sources of bias endemic to self-report, such as self-enhancement and pro-social responding, raw count items often display heaping, in which there are more reports of nice, round numbers than one would expect (Camarda, Eilers, & Gampe, 2008; Roberts & Brewer, 2001; Wright & Bray, 2003). For example, self-reported counts of drinks per night may have heaping around six because beer is often purchased in six-packs. The demographic literature is rife with examples, such as heaping of self-reported age around multiples of five. On one hand, the presence of heaping may reflect a true feature of the process under study that is not well-accounted for in traditional count models. On the other, one might wonder whether the heaping is an artifact of poor attention or memory. All these phenomena threaten the feasibility of using raw counts in behavioral and social science research.

Using ordinal count items allows researchers to avoid many of the problems posed by obtaining raw counts through self-report, at the potential expense of precision. An ordinal count item asks participants to select between given intervals of counts (e.g. 0, 1-2, 3-5, etc.). These items sidestep heaping by never allowing it to occur in the first place. In addition, ordinal count items can be placed alongside Likert-type items without changing the general layout of a survey. Given these advantages, ordinal count items have become widespread in the literature, with their use in large-scale longitudinal studies such as Monitoring the Future (Johnston, O'Malley, & Bachman, 2012) and their promotion by funding agencies such as the NIAAA (2003).

The most appropriate statistical model for analyzing ordinal count items remains unclear, even in regression. One approach is to treat the category numbers or interval midpoints as metrical and model them directly using ordinary least squares. However, McGinley & Curran noted a key problem with this approach is that neither the category numbers nor the midpoints of the intervals are continuous (2014). Additionally, count distributions can easily be skewed, while the linear model assumes that errors are normally distributed. These two qualities of an ordinal count response make it likely that the errors will be heteroscedastic, rendering standard errors and hypothesis tests unreliable without a robust estimator. Moreover, the use of a linear model can result in a model-implied negative count—an impossibility. Another option that is sometimes used is hypothesizing the existence of a continuous distribution underlying the ordinal responses. Current practice is to assume that the underlying distribution is normally distributed, and thresholds are estimated on the normal curve to demarcate the points at which the observed ordinal responses display a category shift (Bollen, 1989; Embretson & Reise, 2000). Yet a clear critique of these approaches is that normal and logistic distributions are continuous and symmetric whereas count distributions are discrete and usually skewed. That does not mean these models cannot generate ordinal count data—they could generate it and can approximate it—but they clearly lack full fidelity with the most plausible underlying structure for the data. In sum, the current underlying distributions used in modeling ordinal counts do not match our theory and prior knowledge that a count distribution, rather than a continuous distribution, actually underlies the responses. Despite the number of modeling options available for ordinal counts, none directly models an underlying count process, rendering it impossible to test whether an underlying count fits the data.

Past work by McGinley and collaborators proposed a solution for ordinal counts in regression models (2015). Paralleling the logic of cumulative logit and probit models, they specified an underlying distribution for the observed ordinal responses, but an underlying distribution that would be appropriate for a count variable rather than the normal or logistic curve. Since the underlying distribution is by definition unobserved, the observed data must provide sufficient information with which to estimate uniquely optimal parameter values. To identify the latent count distribution, they set the thresholds to be equal to those specified by the response options. For example, given the options 0, 1-2, 3-5, and 5+, the thresholds would be constrained to equal 1, 3, and 5. This assumption allows for the estimation of a count distribution underlying an ordinal outcome, more faithfully modeling the underlying process than using a linear, logit, or probit link function. By better specifying the latent process, ordinal count regression model allows researchers to make inferences on the metric of the underlying count. Previous approaches only allow researchers to make inferences on the latent logistic or normal distribution, which may or may not be easily interpretable with respect to the phenomena of interest. Thus, the use of an ordinal count response function allows practitioners to make more precise, grounded inferences on their actual outcome of interest: the count.

Although on its face an ordinal count model appears to have greater fidelity to the underlying process driving the observed response, this may not always be the case in practice. Cognitive psychology, however, has raised concerns regarding the validity of treating ordinal count items as if participants respond with a count. Research has identified a number of different strategies people use to respond to an item. Most relevant of these strategies to ordinal count items is enumeration. When using enumeration, participants count the number of times they remember having performed the behavior within the interval (or within a smaller interval, and then multiply). People are more likely to use other heuristics (e.g. rate-based estimation) when the behavior or experience, occurs often, is perceived as unimportant, or is vaguely defined in the item (Burton & Blair, 1991). It therefore is only sensible to use an underlying count model for ordinal count items if they are about infrequent, significant, and clearly defined behaviors. Domains in which ordinal count models are likely to be particularly useful include developmental psychopathology. For example, researchers of alcohol use disorders often ask about serious consequences of alcohol use, which are often infrequent, significant, and clearly defined. Thus, underlying count models for ordinal count items in psychology remain a theoretically grounded and potentially useful methodological contribution.

Although past work provides elegant methodology and well-tested theoretical framework to the problem of how best to analyze a single ordinal count outcome, many theories in psychology describe relationships between underlying constructs rather than a single outcome. The use of a multiple-item scale allows one to define the construct precisely through its different like its different observed manifestations, to partial out measurement error, and to more reliably measure the construct (Bollen, 1989). For example, researchers regularly conceptualize problem drinking as more than just the number of drinks one has on the average night. Problem drinking as a construct often refers to increased intake, both in frequency and quantity, as well as increased alcohol-related consequences and impairment (White & Labouvie, 1989). It is difficult to imagine a single item that incorporates all these aspects of problem drinking whereas it is easy to imagine a scale that does so. Even when a single item seems to represent the construct domain adequately, using only one item necessarily implies that it perfectly measures the construct. Using multiple items allows one to decompose the variability of the item responses into that of the construct and that of the error of measurement. Finally, the more items used to generate a person's measurement on the construct, in general, the more reliable the measurement becomes. For these reasons, it is best practice in the social and behavioral sciences to make use of multi-item scales for construct measurement.

Despite the clear utility of ordinal count models for multiple item scales, no one has yet extended the single outcome model. In this thesis, I do just that. I began by reviewing the current approaches to latent variable modeling with ordinal count items. I then introduce the ordinal count factor model (OCFM) as a novel model for scales that consist of ordinal items that are coarsened versions of either bounded or unbounded underlying counts. I will then compare the performance of OCFMs with that of the linear factor model and the graded response model, which is an ordinal factor model, in analyzing a common measure of problem alcohol use, the Rutgers Alcohol Problem Index (RAPI). Next, I extend the model to accommodate multiple groups, with a particular interest in how the model could facilitate multi-study integrative data analysis. As the thresholds are defined by the response options, the OCFM does not explicitly require the same response options to be used across studies—a significant advantage over traditional methods. To that end, I compare the results of using an OCFM to get commensurate measures of the RAPI for groups receiving different, experimentally perturbed variations of this measure, again as compared to the linear factor analysis and the graded response model. This work aims to evaluate whether these models, although useful in and for theory, have practical utility in behavior research.

Motivating Example

Counts of behavior are often a concern in the study of alcohol use and alcohol use disorders (AUDs). Whether it is the number of drinks one has a night or the number of times someone has blacked out, the rate of alcohol-related consequences is intimately connected to impairment and pathology (White & Labouvie, 1989). Consequences take on a special role in the study of alcohol use in adolescence. Given that AUD diagnostic criteria were developed for adults, the study of consequences might provide a more precise and developmentally appropriate measure of problem use than diagnosis (Winters, 1997; Martin & Winters, 1998). Focusing on consequences instead of diagnoses also allows researchers to chart the various pathways to being diagnosed with an AUD. Thus, the accurate measurement of alcohol-related consequences is vital not just for studying the development of alcohol use disorders across adolescence, but also for the early detection and intervention of pathological alcohol use.

The Rutgers Alcohol Problem Index

The first and most commonly used measure of the severity of consequences, the Rutgers Alcohol Problem Index (RAPI; White and Labouvie, 1989) might be best modeled using an OCFM. Through a series of 23 items, the RAPI assesses the overall severity of the consequences stemming from the participant's alcohol use over the past year.[2] Each item represents a potential consequence of alcohol use and has four response options: never, once or twice, between three and five times, and more than five times. Prior classical test theory research has demonstrated that this scale has high internal consistency (α=.92) and high test-retest reliability in paper and online administrations (*r*=.88) (Miller, 2002). Although work has been done on the RAPI using an item response theory framework, a non-systematic review suggests that items have always been collapsed to binary responses (i.e. never vs. at least once; Cohn, Hagman, Graff, & Noel, 2011; Earleywine, LaBrie, & Pedersen, 2008; Martens, Neighbors, Dams-O'Connor, Lee, & Larimer, 2007; Neal, Corbin, & Fromme, 2006). RAPI items are often collapsed in this way because endorsement of the upper categories tends to be low for severe consequences, rendering ordinal models difficult to estimate. Thus, considering its wide use in the literature and the dearth of research on its psychometric properties as an ordinal scale, the RAPI could be better understood through the use of an ordinal count factor model.

The Real Experiences and Lives in the University Study

The data for this evaluation come from the Real Experiences and Lives in the University Study (REAL-U) that focused on college student mental health and its measurement. Data was

[2] It is an open debate whether these behaviors are effect indicators or causal indicators (Arterberry, Chen, Vergés, Bollen, & Martens, 2015). I treat the RAPI items as effect indicators, consistent with much of the literature.

collected from a non-probability sample of undergraduate students at a major southeastern research university who had reported alcohol consumption at least once in their lifetime (N=854) over two visits. Every participant received a version of the RAPI at each visit.

I will use the lifetime RAPI (Scenario 1) in my empirical example of a single population OCFM, and a modified version of the lifetime RAPI (Scenario 4) for the evaluation of the multiple groups OCFM. In comparing the results from an OCFM to those of a linear factor model and a graded response model, I aim to not determine whether the OCFM produces a better measure of the intensity of alcohol use related consequences, but also demonstrate how the use of different models leads to similar or dissimilar inferences.

Ordinal Count Factor Models in a Single Population

To model an ordinal count, one must first relate the ordinal response options to an underlying count distribution. For simplicity's sake, one can divide underlying count distributions into two different types—unbounded and bounded. For example, classes missed due to alcohol use could be modeled as unbounded because the limit is large and, for some subjects, potentially unknowable[3]. However, number of days one drank alcohol in the past thirty days is bounded between zero and thirty. As it is straightforward to see that Psychology and behavioral sciences are concerned with both types of counts, I present an OCFM that can accommodate either kind of process. I first give the general specification of the OCFM, then note specific forms for particular bounded and unbounded count distributions with and without zero-inflation. I also discuss identification conditions for OCFMs. I then outline different options for fitting the model via maximum likelihood. I show that the OCFM performs well when fit to simulated data before testing the utility of the model in analyzing empirical data.

The General Ordinal Count Factor Model

Specification. There are two parts of an OCFM: the auxiliary threshold model and the latent factor model. The auxiliary threshold model includes the underlying count distribution and the function by which this probability mass function (PMF) is collapsed to an ordinal response. The latent factor model includes the latent factor(s) in the model and the link function that relates the latent factor(s) to the auxiliary threshold model. As I will show that the ideal link function

[3] This justification stems from the fact that a poison distribution can be derived from a binomial distribution where the number of trials approaches infinity (Ross, 2009).

depends on the auxiliary threshold model chosen, it is necessary to specify the threshold model before the factor model.

The first step in specifying the auxiliary threshold model is to decide what the underlying count distribution should be. Then one specifies that underlying distribution using its PMF:

$$P\left(Y_{ij}^{*} = y_{ij}^{*} \middle| \mu_{ij}, \boldsymbol{\gamma}_j\right) \tag{2}$$

where *i* indexes people, j indexes items, $P_j(\cdot)$ is the probability mass function of the underlying count, Y_{ij}^{*} is the underlying count random variable, y_{ij}^{*} is a unobserved realization of that random variable, μ_{ij} is the mean of the underlying random count distribution for that person *i* on item j, and $\boldsymbol{\gamma}_j$ is a vector containing any other item parameters needed to specify that item's probability mass function, where (for simplicity) these parameters are assumed to be invariant over persons.

With the count distribution specified, I now connect the known item thresholds to the count distribution. Each of the ordinal options is an interval of counts, so the probability of a response is the sum of the probability of each count in that interval. This relationship is represented graphically in Figure 2.

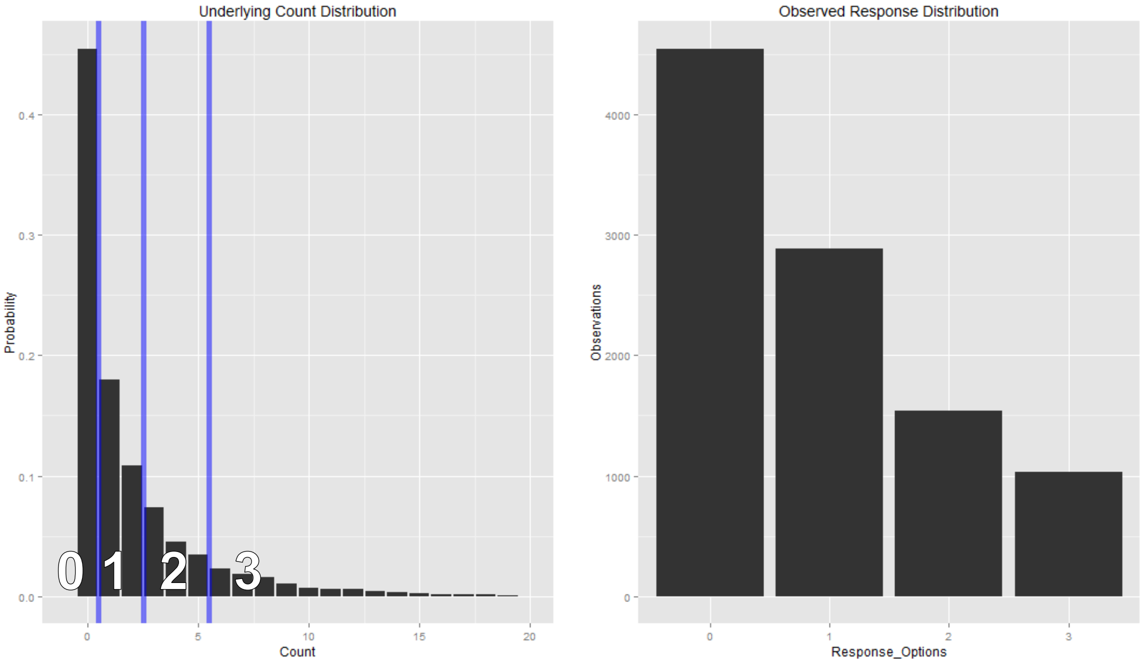

Figure 1: the situation where an underlying count is binned into categories of 0=never, 1=once or twice, 2=3-5 times, and 3 = 6 or more times.

The relationship can also be represented mathematically as

$$P\left(Y_{ij} = c_j \middle| \mu_{ij}\right) = \sum_{w\,=\,\min\,\{\mathrm{c_j}\}}^{\max\,\{\mathrm{c_j}\}} P\left(Y_{ij}^{*} = w \middle| \mu_{ij}, \boldsymbol{\gamma}_j\right) \tag{3}$$

where Y_{ij} is the observed ordinal count variable, c_j is a response category representing an interval of counts, min{·} is a function that returns the minimum count in the interval, and max {·} returns the maximum count in the interval, which may be infinity. The maximum and minimum count of each interval are known, having been determined through the selection of response options for the scale. That is, I do not estimate the thresholds; rather, I set them to be equal to the thresholds presented in the item. This approach departs from conventional IRT and FA models for categorical items, in which thresholds fall along an underlying normal or logistic distribution and their locations must be estimated (Bollen, 1989; Embretson & Reise, 2000). Having described the summation of the probabilities of each individual count, the auxiliary threshold model is now fully specified.

Given an auxiliary threshold model, it is possible to specify the latent factor model. This model may be written as

$$h(\mu_{ij}) = \beta_{0j} + \boldsymbol{\theta}_{\boldsymbol{i}}'\boldsymbol{\beta}_{\boldsymbol{1j}} \tag{1}$$

where $h(\cdot)$ is a link function, $\boldsymbol{\theta}_i$ is a vector of the *k* latent variables, β_{0ij} is the item specific intercept, and $\boldsymbol{\beta}_{1j}$ is a vector of item specific factor loadings. Since count distributions are strictly non-negative, a key aspect of the model specification is to select a link function that will return non-negative values over the domain of the linear combination of latent variables. Examples will be provided for specific versions of the OCFM below. As is typical of many latent variable models, the latent variables are assumed to be normally distributed , $\boldsymbol{\theta}_{\boldsymbol{i}} \sim N(\boldsymbol{\xi}, \boldsymbol{\Phi})$, where $\boldsymbol{\xi}$ is the mean vector and $\boldsymbol{\Phi}$ is the covariance matrix of the latent variables. The assumption of an underlying normal latent variable is considered reasonable regardless of the response distribution (Embretson & Reise, 2000). In this document, I scale the latent factors so that the means are zero and the diagonal elements of phi are one.

The Ordinal Negative Binomial Factor Model

With the general model as a reference, it is straightforward to specify an example of where the underlying count is unbounded. Unbounded count models are especially useful when the upper limit is unknown and could either vary across individuals or be likely much larger than the counts routinely observed. For example, there is theoretically a limit for the number of people someone calls in a week, but there is no way of knowing it.

The simplest unbounded count distribution is the Poisson distribution, which is described by a single rate parameter which is equal to both the random variable's mean and variance. A model based on an underlying Poisson distribution then cannot represent distributions where the variance is greater than the mean, a condition referred to as overdispersion. In behavioral data, the variance is often much greater than the mean, making the Poisson distribution rarely appropriate for modeling raw counts (McGinley et al., 2015). Given that one would expect an overdispersed count distribution if one collected a raw count, one should use an underlying overdispersed count distribution to model self-reported ordinal count data. The most common overdispersed, unbounded count distribution is the negative binomial, which I use below (Hilbe, 2011).

Specification. As with all OCFMs, I first specify the auxiliary threshold model and then specify the latent factor model. For the auxiliary threshold model, I select a negative binomial distribution[4], which includes a parameter α to model overdispersion. The PMF for each item can be represented as

$$P\left(Y_{ij}^{*} = y_{ij}^{*} \middle| \theta_i\right) = \frac{\left(y_{ij}^{*} + \alpha_j{}^{-1}\right)}{\Gamma\left(\alpha_j{}^{-1}\right)\Gamma\left(y_{ij}^{*} + 1\right)} \left(\alpha_j \mu_{ij}\right)^{y_{ij}^{*}} \left(1 + \alpha_j \mu_{ij}\right)^{-(y_{ij}^{*} + \alpha_j{}^{-1})} \tag{4}$$

where $\alpha > 0$ is the dispersion parameter, and μ_{ij} is the mean of the underlying random count distribution for that person and that item. To ensure that the mean is always positive, a log link function is used to connect the latent variable to each item:

[4] Specifically, we select an NB2 model, which does not force overdispersion to be constant (Hilbe, 2011)

$$ln(\mu_{ij}) = \beta_{0j} + \boldsymbol{\theta}_i'\boldsymbol{\beta}_{\mathbf{1j}} \tag{5}$$

One significant advantage of the OCFM model is that one can interpret the parameters on the metric of the raw variable, just as one would in a typical negative binomial regression model. The exponentiated intercept, exp (β_{0j}), is the expected value of the underlying count for the typical person (with $\boldsymbol{\theta} = \mathbf{0}$). The exponentiated slope for the *k*th latent variable, exp $(\beta_{1j,k})$, is the predicted multiplicative increase of the underlying count of the *j*th item for a one unit increase in that latent variable, holding the other latent variables constant. It is also possible to interpret α, although it is difficult to do so precisely. It is often sufficient to interpret it as simply a measure of how much larger the variance is than the mean.

Ordinal Beta Binomial Factor Model

Although some counts may have indeterminate bounds, others have clear ones. These bounds are often temporal (i.e. how many days in the past month). However, they could also be bounded for other reasons. For example, the count of a teen's classmates that have had a drink of alcohol is bounded by their total number of classmates. Although such bounds might vary across people, for simplicity's sake I only present models where the upper bound is invariant and known across subjects.

Specification. The most basic bounded count distribution is the binomial. However, much like the Poisson distribution, the binomial's mean and variance are constrained.[5] To model overdispersion with a raw bounded count, most methodologists use a beta binomial model,

[5] For the binomial, $\mu = np$ and $\sigma^2 = np(1-p)$, so the mean and variance are related such that $\sigma^2 = \mu\left(1 - \frac{\mu}{n}\right)$.

which uses a parameter $\tau,$ bounded between zero and one, to quantify overdispersion (Hilbe, 2013). The closer τ is to one, the more the mass of the distribution is pushed towards zero and the maximum count. The closer τ is to zero, the more the distribution resembles the standard binomial. Thus, I use a beta binomial for the underlying count, the PMF of which is

$$P\left(Y_{ij}^{*}=y_{ij}^{*}\middle|\boldsymbol{\theta_i}\right)=\binom{n_j}{y_{ij}^{*}}\frac{B\left(y_{ij}^{*}+\pi_{ij}\left(\frac{1}{\tau_j}-1\right),\left(\frac{1}{\tau_j}-1\right)(1-\pi_{ij})+n_j-y_{ij}^{*}\right)}{B\left(\pi_{ij}\left(\frac{1}{\tau_j}-1\right),\left(\frac{1}{\tau_j}-1\right)(1-\pi_{ij})\right)},y_{ij}^{*}=0,\dots,n_j \tag{6}$$

In the above equation, *B*(.) is the beta function, π_{ij} is the mean of the underlying beta distribution for that person and item, τ_j is the overdispersion parameter for that item, n_j is the upper bound of the count[6], and all other symbols have the same meanings as they did for the negative binomial model. As π_{ij} is bounded between zero and one, I use a logit link function to specify the latent factor model:

$$logit\left(\pi_{ij}\right)=\beta_{0j}+\boldsymbol{\theta_i'\beta_{1j}} \tag{7}$$

In the beta binomial OCFM, the parameters are again expressed in the metric of the underlying count, so π_{ij} can be interpreted as the probability of participating in a behavior (e.g. probability of drinking alcohol during a given day). For a given item *j,* the typical person will endorse

[6] so $\mu_{ij}=n_j\pi_{ij}$.

participating in that behavior $\frac{\mathrm{n_j}}{1+\exp(-\boldsymbol{\beta}_{0j})}$ times. Each entry in $\boldsymbol{\beta}_{1j}$ represents the predicted increase of the log-odds of performing the given behavior during a given occasion for a one unit increase of the corresponding latent variable. Finally, $\boldsymbol{\gamma}_j$ can be interpreted as a measure of the inertia of the behavior—that is, how much does having performed the behavior previously increase the probability of performing the behavior during a subsequent occasion. For example, one would expect that the probability of drinking another day of the week would be greater for someone who had already drank that week than someone who had not drank.

Zero-Inflated Ordinal Count Factor Models

Researchers fit zero-inflated models for multiple reasons (Hilbe, 2011). On one hand, one might note an extreme preponderance of zeroes in the responses. On the other hand, a theory might suggest that there are two groups in the population: those that never participate in the behavior and those that participate at various rates. To use fighting with relatives due to alcohol consumption as an example, there may be one group of people who never consume alcohol around relatives, and another group that regularly consumes alcohol around relatives, putting them at risk.

Specification. The PMF for a general zero-inflated OCFM is straightforward to represent as a mixture model:

$$P\left(Y_{ij}^* = 0 \middle| \boldsymbol{\theta}_i\right) = (1 - \eta_j) + \eta_j P_j\left(Y_{ij}^* = 0 \middle| \boldsymbol{\theta}_i\right) \tag{8}$$

$$P\left(Y_{ij}^* = y_{ij}^* \middle| \boldsymbol{\theta}_i\right) = \eta_j P_j\left(Y_{ij}^* = y_{ij}^* \middle| \boldsymbol{\theta}_i\right), \qquad y_{ij}^* = 1, \dots, n_j \tag{9}$$

where P_j is the probability mass function for the risk group, which might be a negative binomial or beta binomial distribution, as discussed above, and η_j is the probability for being in the at-risk group. Thus, zero-inflation models introduce one extra parameter for the underlying latent count distribution relative to their non-inflated counterparts, the mixing probability.

Identification of Ordinal Count Factor Models

It is vital to consider under what conditions an OCFM is identified; that is, how much information is needed and which assumptions are required to get unique estimates of the model parameters. For an OCFM to be identified, both the auxiliary threshold model and the latent variable model must be identified.

To identify the auxiliary threshold model, one needs to have enough information and adequate assumptions to uniquely estimate the parameters of the underlying count distribution. As I am only considering models with known thresholds, it is the number of intervals that determines which count distributions can be modeled. Through past work on ordinal factor analysis it is possible to exactly identify a response distribution with *q* parameters given *q* known response thresholds for the ordinal item (Bollen & Curran, 2006). Thus a sufficient, but not necessary condition to identify the auxiliary threshold model is for there to be no more than *q* parameters in its PMF. As such, the negative binomial and the beta binomial OCFM require three known thresholds, while their zero-inflated counterparts require four.

The identification of the latent variable model requires the usual constraints to set a scale for the factor. This can be done in one of two ways: set a scaling indicator or standardize the latent variable. Setting a scaling indicator means constraining the intercept of an item to zero and the loading of that item to one. Doing so allows for the mean and the variance of the latent variable to be directly estimated. Moreover, in the linear factor model, the scale of the latent variable is set such that a one unit increase of the latent variable corresponds to a one unit increase in the scaling indicator. However, in the case of the OCFM, these parameters are exponentiated, and that meaning is lost. As such, while one might use the scaling indicator approach to identify the latent factor model, it does not provide same interpretational clarity as in a linear factor model. Standardizing the latent variable to have a mean of zero and a variance of one makes interpreting OCFM parameters straightforward. This approach allows one to freely estimate every slope (β_{1j}) parameter. In doing so, one could interpret the slope parameter as the predicted increase in the transformed mean of the underlying response function for a one standard deviation increase in the latent variable. Moreover, one can freely estimate every intercept (β_{0j}) parameter, which is the transformed mean of the underlying response function for the typical person (i.e. $\theta = 0$). A potential drawback of this approach includes the potential for inadvertently constraining means and variances to be equal across time or groups. Nevertheless, given that the nonlinearity of the model complicates the scaling indicator approach, I implement this standardized latent variable constraint throughout the thesis.

Estimation of Ordinal Count Factor Models

The likelihood for an OCFM is

$$L = \int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \Pi_{i=1}^{N} \Pi_{j=1}^{J} \left[\Pi_{c=1}^{C_j} \left(\sum_{w\ =\min\{\mathrm{c}\}}^{\max\{\mathrm{c}\}} P\left(Y_{ij}^{*} = w \middle| \mu_{ij}, \boldsymbol{\gamma}_j\right) \Phi(\boldsymbol{\theta}) \right)^{I_{y_{ij}=c}} \right] d\boldsymbol{\theta} \tag{11}$$

where $\Phi(\cdot)$ is the multivariate normal probability density function, and $\mathrm{I}_{\mathrm{y}_{\mathrm{ij}}=c}$ is an indicator function. The log-likelihood follows clearly:

$$ll = \int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{N}\sum_{j=1}^{J} \left[\sum_{c=1}^{C_j} I_{y_{ij}=c_j} ln \left(\sum_{w\ =\min\{c\}}^{\max\{c\}} P\left(Y_{ij}^{*} = w \middle| \mu_{ij}, \boldsymbol{\gamma}_j\right) \Phi(\boldsymbol{\theta}) \right)\right] d\boldsymbol{\theta} \qquad (12)$$

To calculate the log likelihood for estimation, one must integrate over the range of the latent factor(s), a computationally demanding task for even a single latent variable. Instead of analytically solving for the integral, an approximation to the integral can be obtained using quadrature. Quadrature is available in many popular statistical software programs, although here I will focus on R as it allows users to easily define their own models (R Core Team, 2015). In SAS software, users can specify novel latent variable models using PROC NLMIXED, although I do not explore this option in depth. In R, the mirt package can be used to specify novel item response functions (Chalmers, 2012). All analyses for this thesis are done in R, and sample code for OCFMs in a single population can be found in Appendix B.

Feasibility Study

The feasibility study presented here demonstrates that, given data generated by an OCFM, one can recover the original parameters. The feasibility study consists of generating data from four different OCFMs: negative binomial, beta binomial, zero-inflated negative binomial, and zero-inflated beta binomial. The parameters for the generating model were obtained using real data to get plausible values. For the negative binomial model and the zero-inflated negative binomial model, I used seven items from the RAPI from REAL-U, and for the beta-binomial model and the zero-inflated beta-binomial model, I used four items from the CES-D depression scale from the 1994 wave of the National Longitudinal Survey of Youth (Bureau of Labor Statistics, U.S. Department of Labor., 2012; Radloff, 1977). I first fit the model to the real data to get the parameter estimates. I then examined the estimated values to rule out improper solutions.

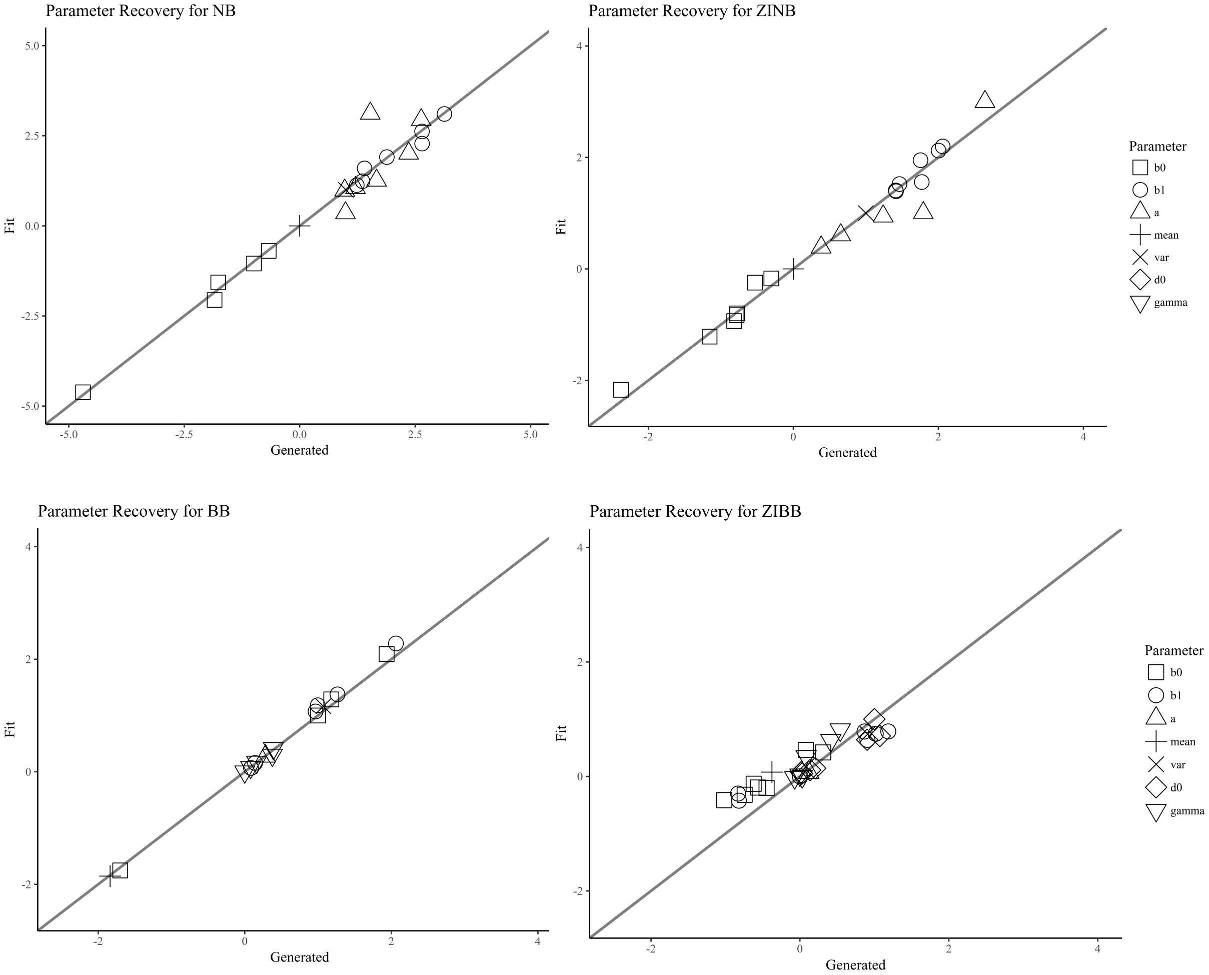

Figure 2: Parameter recovery plots for the four OCFMs described.

The parameter estimates were then used to generate a sample of 1,000 cases for each model, which were then fit with the correct model using random starting values. The results are represented in Figure 2, and relative bias is reported in Table 1

These results demonstrate that the OCFM can be successfully fit to data generated from an OCFM. Specifically, it shows that the estimation of an OCFM can be performed successfully—a nontrivial finding. Of course, this finding has little external validity, as it does not show that the OCFM is suitable to real data, but this feasibility study constitutes an important first step. Looking more closely, one can see that the parameter recovery for the, NB and ZINB model looks adequate. However, the zero-inflation parameter η seems to be severely negatively biased. As such, it seems like one should be cautious about interpreting the zero-inflation parameter in these models.

Table 1: Mean Relative Bias by Parameter and Model in Feasibility Study

Parameter	Negative Binomial	Zero-Inflated	Beta Binomial	Zero-Inflated
		Negative Binomial		Beta Binomial
β_0	<0.01	-0.11	0.03	-0.70
β_1	-0.03	0.02	0.11	-0.26
α	<0.01	-0.14		
η		-0.52		1.16
γ			-0.02	4.02

This result should not dissuade one from using an OCFM, although it may lead one to consider a NB model over a ZINB model. Now focusing on the BB and ZIBB models, it seems as if parameter recovery was excellent for the BB model, but less so for the ZIBB model—similar to the results of the unbounded count models. Nevertheless, the ZIBB model produced terribly biased estimates, which is deeply concerning. Given the small maximum count of seven, perhaps the ZIBB model was over-parametrized. Remember that the overdispersion parameter in a BB model pushes the responses towards the maximum and the minimum—potentially making the zero-inflation parameter difficult to estimate. Regardless of the relative bias of the parameters, I find that the coverage rates of the score confidence intervals were close to nominal levels (NB=0.96; ZINB=0.97; BB=0.96; ZIBB=0.94). Moreover, the correlations between the true scores and the estimated scores were acceptable (NB=0.79; ZINB=0.85; BB=0.86; ZIBB=0.76), especially considering the limited number of items. These results suggest that the OCFM can adequately reproduce the generating scores, as well as quantify the uncertainty in the scoring process. It should be clear that a single replication does not generalize to an entire family of models. Nevertheless, this simulation shows that OCFMs can be fit using readily available statistical software and, given a properly specified model, can adequately recover the generating parameters. Yet this simulation also suggests that one should not take specifying zero-inflation lightly—as it can lead to meaningfully biased parameter estimates.

Study 1

For the empirical demonstration, I will return to our motivating example of the RAPI in the REAL-U Study. I will focus our attention on the selection and interpretation of an appropriate OCFM for the original (unperturbed) version of the RAPI. Despite the theoretical benefits of OCFMs, it is important to consider whether they produce tangible benefits above and beyond the currently available methods. Through this empirical demonstration, I seek to accomplish two aims: to demonstrate how one should select between and interpret OCFMs and to compare the OCFM to traditional item response and factor analytic models. In addition, I hypothesize that:

1. the GRM, CFA, and OCFM will all produce scores that are highly correlated, but not to such an extent that model choice is trivial;
2. the OCFM will fit better than the GRM due to the fewer parameters that are estimated;
3. the OCFM scores will have smaller standard errors compared the standard errors of the GRM score for that person due to its increased parsimony.

Data

I use the lifetime version of the original RAPI from the REAL-U study. In both Study 1 and Study 2, I omit the fifth item: "Relatives avoided me." Across all four items and visit scenarios, no more than two people endorsed a category higher than the second. For the original RAPI specifically, no one endorsed category three, and only two people endorsed category two. Given this sparseness and the potential inapplicability of this item in a college population, this item was dropped for purposes of the empirical analysis.

Method

Fitting the Comparison Models. To provide a basis for comparison, I begin by fitting the graded response model and the linear normal model to the data. For the linear normal model, I recode the responses to represent the midpoint of the selected interval. As the final interval (5+) does not have a finite midpoint, there is no obvious number to which to recode the response. Acknowledging this limitation, I recode those responses as 7.5 so the interval between responses options increases by one each time (0, 1.5, 4, 7.5).

Selecting an Optimal Ordinal Count Factor Model. After the model converged, I reviewed the parameter estimates. As a logit larger than three indicates a probability close to 1, noticeably large zero-inflation model parameters indicate that the item does not support a sensible representation of zero-inflation. To that end, those items[7] were then fit with a standard negative binomial distribution. After inspection, this model was then selected as the OCFM for comparison.

Comparing the Models. The final section of our empirical demonstration compares the two traditional models and the final OCFM. All models were fit with N(0,1) scaling of latent factor and all item parameters estimated. After presenting the estimation results from all three models, I will compare the fit of the GRM and the OCFM using information criteria. As I cannot compare the fit of the linear CFA to either the GRM and the OCFM[8], I will report only the overall fit of the linear CFA. In addition to the fit of the model, I will also compare the quality of the scores generated for the factor. As EAPs are often preferred, (Thissen & Orlando, 2001), I will compare the correlation of EAPs across models. Finally, I will compare the precision of EAPS by examining the standard errors of scores across models. I will do this by testing whether the difference between a participant's OCFM score's standard error and their GRM standard error is less than zero. This approach takes into account that specific response patterns may have unique effects on scores. As such, while the standard errors may be produced by different models, they were produced by the same response patterns, and are thus at least roughly comparable.

Results

Estimation Results and Model Fit. The linear CFA was estimated using lavaan (Rosseel, 2011). The parameter estimates for the linear CFA are reported below. The overall chi-square test was rejected (X^2=1718, df=209, p<0.001). The CFI of 0.62 and the RMSEA of 0.133 indicate poor overall fit.

[7] Items 1, 2, 3, 4, 6, 7, 8, 9, 10, 16, 18, 19, 20, and 22

[8] A single polytomous item is represented as multiple dichotomous variables for the OCFM and the GRM.

The GRM was successfully estimated using mirt. Finally, a ZINB OCFM was estimated, and then items with $\eta > 3$, which suggested a lack of zero inflation, were dropped. The more parsimonious model did not fit significantly worse (LRT here), so it was retained.

Turning our attention to the information criteria, I can note that the AIC and the BIC for the OCFM is smaller than the GRM. Considering that the linear CFA did not fit well, and the OCFM (AIC=10901.87, BIC=11199.43) fit better than the GRM (AIC=10944.95, BIC=11298.8), selecting the OCFM is justifiable.

Interpretation. To demonstrate the interpretational clarity afforded by this model, I examine the item parameters of two items, item one "Got into fights with other people (friends, relatives, strangers)", and item eleven "Wanted to stop drinking but couldn't".

For the linear CFA models, we can interpret the intercepts to say that the typical person has gotten into 0.97 fights with other people because of their drinking and tried to stop drinking, but could not 0.06 times. For a one standard deviation increase in the latent variable we expect to observe an increase of 1.11 fights and 0.21 failed attempts to quit drinking. This model also improperly implies that the typical person one standard deviation below the mean on the latent variable should have negative fights and failed quit attempts. The linear model estimates and interpretation also may differ depending on the arbitrary selection of the final category midpoint score. As such, sensitivity analyses were performed, and I found that the choice of a smaller final category score (5.5) did not meaningfully change the results. For the GRM, we can interpret the thresholds to say that a typical person has around a 72% chance of having never been in a fight due to drinking, a 23% chance of having one or two fights, a 3% chance of having three to five fights, and a 2% chance of having more than five.

Table 2

Linear CFA Item Parameter Estimates

Item Stems	Loading	Intercepts	Residual Variance
1. Got into fights with other people (friends, relatives, strangers)	1.112	0.972	2.098
2. Went to work or school high or drunk	1.167	1.01	2.806
3. Caused shame or embarrassment to someone	1.24	1.058	1.631
4. Neglected your responsibilities	1.398	1.493	2.985
6. Felt that you needed more alcohol than you used to in order to get the same effect	1.129	1.172	3.489
7. Tried to control your drinking (tried to drink only at certain times of the day or in certain places, that is, tried to change your pattern of drinking)	1.06	0.861	2.893
8. Had withdrawal symptoms, that is, felt sick because you stopped or cut down on drinking	0.204	0.128	0.645
9. Noticed a change in your personality	0.513	0.476	1.66
10. Felt that you had a problem with alcohol	0.482	0.247	0.503
11. Wanted to stop drinking but couldn't	0.206	0.064	0.184
12. Suddenly found yourself in a place that you could not remember getting to	1.344	1.08	2.187
13. Passed out or fainted suddenly	0.554	0.405	0.977
14. Had a fight, argument, or bad feeling with a friend	1.141	0.978	1.69
15. Kept drinking when you promised yourself not to	0.622	0.416	0.976
16. Felt you were going crazy	0.692	0.269	0.789
17. Felt physically or psychologically dependent on alcohol	0.43	0.133	0.373
18. Was told by a friend, neighbor or relative to stop or cut down drinking	0.736	0.376	0.975
19. Not able to do your homework or study for a test	1.255	1.232	2.72
20. Missed out on other things because you spent too much money on alcohol	0.627	0.594	1.738
21. Missed a day (or part of a day) of school or work	1.048	0.923	2.136
22. Had a fight, argument, or bad feeling with a family member	0.347	0.17	0.386
23. Had a bad time	1.535	1.672	2.429

Table 3

Graded Response Model Parameter Estimates

Item Steps	Slope	First Threshold	Second Threshold	Third Threshold
1. Got into fights with other people (friends, relatives, strangers)	1.834	-0.96	-3.007	-4.21
2. Went to work or school high or drunk	1.913	-1.362	-2.928	-3.865
3. Caused shame or embarrassment to someone	2.231	-0.648	-3.246	-4.71
4. Neglected your responsibilities	1.994	-0.216	-2.265	-3.597
6. Felt that you needed more alcohol than you used to in order to get the same effect	1.489	-1.013	-2.178	-3.251
7. Tried to control your drinking (tried to drink only at certain times of the day or in certain places, that is, tried to change your pattern of drinking)	1.606	-1.758	-2.793	-3.581
8. Had withdrawal symptoms, that is, felt sick because you stopped or cut down on drinking	1.139	-3.898	-4.636	-5.224
9. Noticed a change in your personality	1.041	-1.947	-3.206	-4.177
10. Felt that you had a problem with alcohol	2.129	-3.208	-5.898	-6.97
11. Wanted to stop drinking but couldn't	2.030	-5.068	-7.127	-7.936
12. Suddenly found yourself in a place that you could not remember getting to	2.435	-1.172	-3.142	-4.557
13. Passed out or fainted suddenly	1.203	-1.925	-3.679	-5.052
14. Had a fight, argument, or bad feeling with a friend	1.828	-0.766	-2.95	-4.399
15. Kept drinking when you promised yourself not to	1.154	-1.93	-3.466	-5.029
16. Felt you were going crazy	2.00	-3.597	-4.783	-5.73
17. Felt physically or psychologically dependent on alcohol	2.318	-4.656	-6.569	-7.289
18. Was told by a friend, neighbor or relative to stop or cut down drinking	1.912	-2.793	-4.502	-5.49
19. Not able to do your homework or study for a test	1.844	-0.539	-2.712	-3.683
20. Missed out on other things because you spent too much money on alcohol	1.594	-1.932	-3.311	-4.9
21. Missed a day (or part of a day) of school or work	1.848	-1.162	-2.972	-4.346
22. Had a fight, argument, or bad feeling with a family member	2.002	-3.823	-5.551	-7.82
23. Had a bad time	2.029	0.358	-1.899	-3.689

Table 4

Ordinal Count Factor Model Parameter Estimates

Item Stems	Loading	Intercept	Overdispersion parameter	Zero-Inflation parameter
1. Got into fights with other people (friends, relatives, strangers)	1.623	0.979	0.976	
2. Went to work or school high or drunk	2.075	1.141	2.002	
3. Caused shame or embarrassment to someone	1.547	0.873	0.4	
4. Neglected your responsibilities	1.512	0.352	0.676	
6. Felt that you needed more alcohol than you used to in order to get the same effect	1.821	0.595	2.627	
7. Tried to control your drinking (tried to drink only at certain times of the day or in certain places, that is, tried to change your pattern of drinking)	2.17	1.015	5.153	
8. Had withdrawal symptoms, that is, felt sick because you stopped or cut down on drinking	2.509	3.086	47.693	
9. Noticed a change in your personality	1.317	1.272	6.7	
10. Felt that you had a problem with alcohol	2.001	3.021	1.687	
11. Wanted to stop drinking but couldn't	2.515	4.021	0.001	0.955
12. Suddenly found yourself in a place that you could not remember getting to	2.04	1.132	0.525	-2.374
13. Passed out or fainted suddenly	1.356	1.017	0.393	0.099
14. Had a fight, argument, or bad feeling with a friend	1.495	0.814	0.397	-2.319
15. Kept drinking when you promised yourself not to	1.152	0.684	0.271	0.429
16. Felt you were going crazy	2.621	3.454	6.047	
17. Felt physically or psychologically dependent on alcohol	2.423	3.501	0.001	0.526
18. Was told by a friend, neighbor or relative to stop or cut down drinking	2.157	2.626	3.039	
19. Not able to do your homework or study for a test	1.595	0.653	0.879	
20. Missed out on other things because you spent too much money on alcohol	1.796	1.598	2.902	
21. Missed a day (or part of a day) of school or work	1.712	1.096	1.092	-3.613
22. Had a fight, argument, or bad feeling with a family member	2.415	3.921	2.741	
23. Had a bad time	1.246	0.008	0.148	-2.412

Similarly, the typical person has a 99.3% chance of never failing a quit attempt, a 0.6% chance of failing once or twice, a 0.06% chance of failing three to five times, and around a 0.04% chance of failing more than five times. A one standard deviation increase in the latent variable leads to an increase in the underlying propensity to get into more fights due to drinking by 1.834 and the propensity to have more failed quit attempt by 2.03. As noted before, this model does not allow us to make inferences on the exact count of the behavior, but rather its underlying propensity.

For the OCFM, we can interpret the exponentiated intercept to say that the typical person will have gotten into 0.39 fights in their lifetime due to alcohol. The exponentiated slope 4.8 means that for a one standard deviation increase in problem alcohol use, we expect an increase in the number of lifetime fights by a factor of 4.8. For instance, for a person one standard deviation above the mean, we could expect .39*4.8 = 1.87 fights in their lifetime related to alcohol use. By the same reasoning, for a person one standard deviation below the mean we would expect .39/4.8 = .08 fights. The overdispersion parameter implies that there are likely other predictors of the number of reported fights than the latent variable of problem alcohol use. We cannot be sure whether that other source of variability is a specific factor or solely measurement error.

Turning our attention to the eleventh item, we can interpret its transformed zero-inflation parameter to say that the typical person has a 28% chance to be at risk for a failed quit attempt. Taking that risk into account, we interpret the exponentiated intercept to say we expect the typical person at-risk person to have failed a quit attempt 0.005 times. Each increase of one unit of the latent variable results in an increase in the expected count by a factor of exp(2.515) = 12.2, suggesting that this behavior, if a person is at risk for it, is extremely sensitive, but given the small intercept, is probably not sensitive in substantive terms. So, for a person one standard deviation above the mean, we expect them to have failed a quit attempt 0.06 times. An important observation to make with this item is that even though it may seem highly discriminating, it is only discriminating for those people at risk for the behavior. This caveat is not found in traditional ordinal models because zero inflation is not modeled.

Factor Score Correlations. EAPs were obtained through the respective R packages and then compared. Plots demonstrating the relationships between scores can be found in Figure 2. The curvilinear relationship between the scores from the CFA and those from the GRM and the OCFM is noticeable. This suggests that the CFA was not as capable[9] of modeling the nonlinear relationship between the factor and the responses as were the GRM and the OCFM. Furthermore, the linear relationship between the GRM and the OCFM suggests that the item response model of the GRM can approximate that of the OCFM.

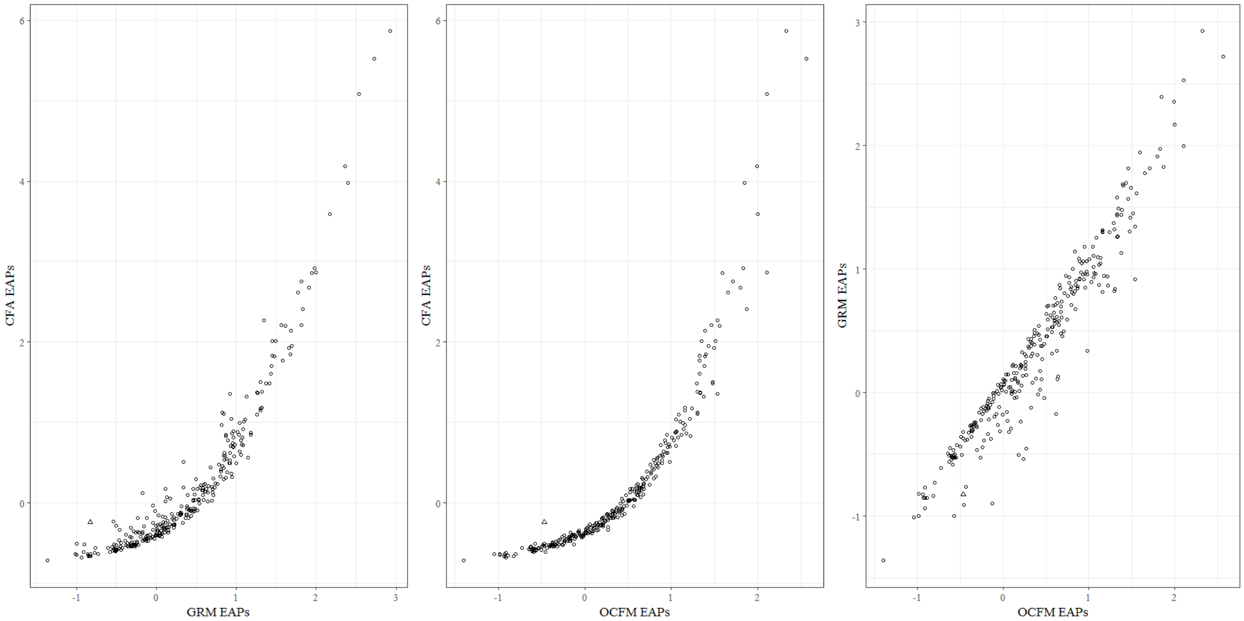

Figure 2: Plots of EAPs across models. The outlier is marked with a triangle.

[9] I performed a sensitivity analysis by coding the responses 0, 1.5, 3.5, and 5.5 as well as 0, 1, 2, and 3—the curvilinear trend remained in both cases.

These relationships are reflected in the correlations between scores. The scores are highly correlated between all three models but most correlated between GRM and OCFM. However, care should be taken with assuming that these scores approximate each other, even between the OCFM and the GRM. One might notice a small outlier, marked with a triangle, around the middle of the CFA vs. OCFM score plot, but conclude that the rest of the plots seem straightforward.

Table 5

Factor Score Correlations

	Linear CFA	GRM
GRM	0.860	
OCFM	0.833	0.982

Score Precision. Turning our attention to the standard errors, we can note that the GRM and the OCFM do not always provide smaller standard errors than the CFA (Figure 3, SE=0.31). For reference, the mean of the standard errors for the GRM is 0.36, and the mean of the standard errors for the OCFM is 0.35. When evaluating the difference in the standard errors using a t-test, the OCFM provided significantly smaller SEs than the GRM, but significantly larger SEs than the CFA.

An outlier is also clearly noted, the same from the OCFM vs CFA score scatterplot. This person responded that they had no consequences besides being told to cut down their drinking, which occurred more than five times, a highly unusual response pattern. Removing this outlier from the t-tests did not affect the results.

As the standard errors of the GRM and the OCFM vary across the level of the latent variable, I plotted them against the EAPs calculated from the OCFM in Figure 4. Here I find that the OCFM tends to provide more precise scores than the GRM as the level of the latent variable increases.

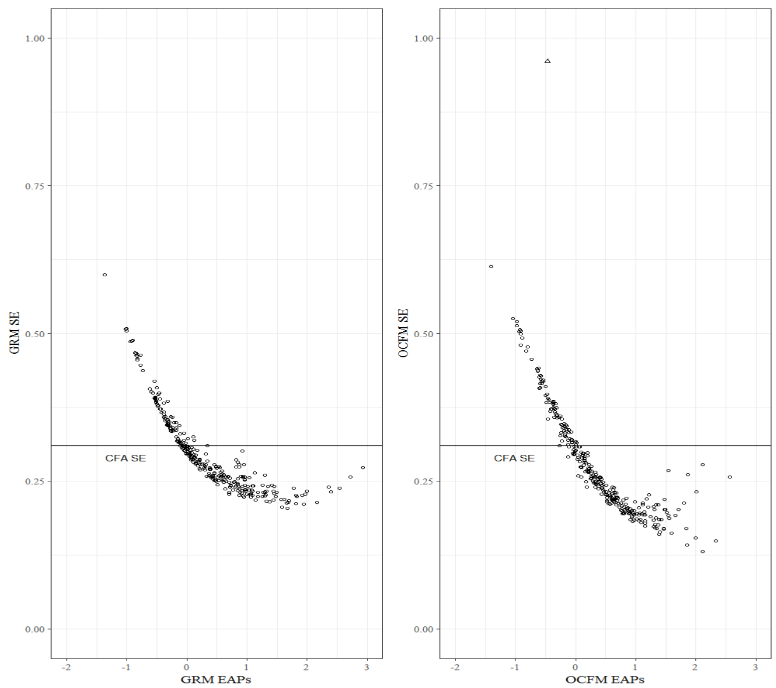

Figure 3: Plots of the standard errors of the GRM and the OCFM by their EAPs. The line at 0.31 in each of the models represent the standard errors of the CFA EAPs. Note the same outlier as before marked with a triangle.

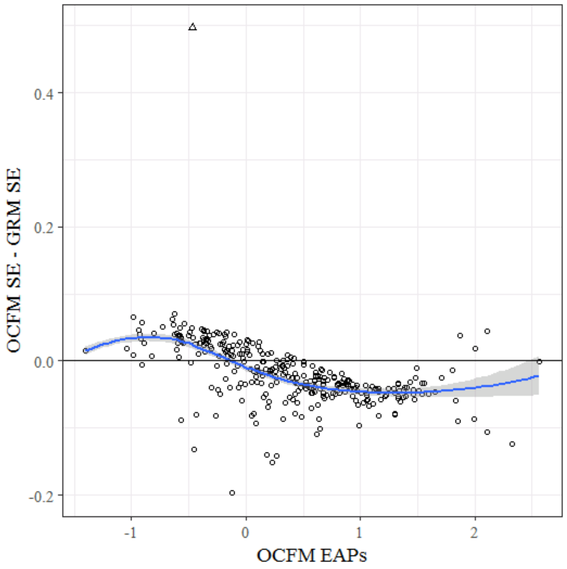

Figure 4: The difference between a participant's OCFM and GRM standard errors are plotted against their OCFM EAP. A loess line with its 95% confidence interval is included.

Discussion

This empirical example served two purposes. The first was to demonstrate that it is possible to fit an OCFM to real data. The second was to evaluate whether it is possible to get better fit with the OCFM than with a GRM.

The example clearly demonstrated that an OCFM can be fit to real data, with a flexible model for unbounded counts, the zero-inflated negative binomial, serving as the starting point in the model selection process. It was possible to select the most parsimonious underlying count for that specific item using an ad-hoc stepwise process. That said, even with an overdispersion parameter and many items being zero-inflated, the OCFM fit better than the GRM. Thus, one must weigh fit against parsimony and relative bias for the OCFM, at least when overdispersion is concerned. Future simulation studies should be performed to determine whether the model selection strategy I implemented here can lead to correct identification of the simplest possible count distribution for the data, and if so, whether the gain in parsimony is worth potential misspecification through improper constraints.

The model fit results suggest that the OCFM can fit better than the less restrictive GRM. This finding has larger implications than simply supporting the use of the OCFM. It demonstrates that measures that have been successfully modeled with a GRM may show improved fit under an OCFM. One might argue that the ability of the GRM to approximate the OCFM renders it unnecessary. On the contrary, this result suggests that a justifiable assumption can produce a measurement model that is not only more parsimonious, but is also more directly interpretable. Additional studies could reexamine these inventories to determine whether this result generalizes. Moreover, a simulation study should be performed to evaluate how well data generated under a GRM is fit by an OCFM. A variety of potential OCFM thresholds could be tested against the same generating GRM. On one hand, such a study may serve as a cautionary note against using an OCFM for all collapsed count inventories. On the other hand, it may demonstrate that despite the misspecified measurement model, the scores produced were still comparable. Those findings would provide the fitting propensity, and thus the specificity of the OCFM to an underlying count.

The emergence of the outlier in only the OCFM must be seriously considered—either as a sign of person misfit or model misspecification. Given the relatively narrow range of the other standard errors, its standard error suggests serious uncertainty. A reasonable hypothesis would be that the participant's response pattern did fit the GRM but did not fit the OCFM due to the additional constraints of the OCFM. From a substantive angle, it does seem odd that someone reports no observable negative consequences of alcohol use, but has been told to cut down on their drinking. On one hand, the participant's family may strongly discourage any use of alcohol. On the other hand, the participant may simply be in denial about the consequences of their alcohol use. That said, another explanation of this outlier may be that OCFM is misspecified—perhaps many these consequences should be seen as causal indicators. No matter what one believes is driving this outlier, its presence suggests that the OCFM may be used to detect aberrant response patterns and to make risky predictions that can be falsified.

The results of the first empirical demonstration clearly show the utility of the OCFM. It is straightforward to fit in open source, widely-distributed software. Real data is fit easily with an OCFM, providing interpretable parameters in the metric of the underlying count. The OCFM fits the data better than other models when evaluated with information criteria. Although the scores are similar to those produced by the GRM, they were generated using a theoretically motivated distribution and have greater precision in the OCFM. For all these reasons, I can support the use of an OCFM in the analysis of real data from a single population.

Ordinal Count Factor Models across Multiple Groups

Contemporary research in psychology, the social sciences, and allied health fields often tests whether the properties of a scale differ across multiple groups. For example, Ramirez and colleagues tested whether responses to items on the Mini-Mental State Examination differed as a function of the language the test was administered (2006). Most of these studies are concerned with whether item responses are independent of group membership after conditioning on the latent variable. That is, people at the same level of the latent variable may respond to an item differently based on other personal attributes. In the factor analysis literature, this phenomenon is referred to as a violation of measurement invariance (MI), and in the item response literature, it is known as differential item functioning (DIF).

A classic example of the need to examine measurement invariance comes from the measurement of depression (Steinberg & Thissen, 2006). Perhaps our research question was whether there was a difference in the means of depression for men and women. One would then encounter a hurdle: Given the same level of depression, women report more crying than men. Thus, for men, the crying item is more “difficult” to endorse than it is for women. Regardless of the source of this difference, if one does not allow for differences in that item’s parameters across men and women, then one would overestimate the depression scores of women and underestimate those of men. As the example demonstrates, if the performance of the scale depends on one’s group, then it is difficult to make meaningful comparisons across groups. The difference in one’s score could come from differential item functioning as surely as it could come from true underlying differences in the measured construct. These true underlying differences in the construct are often referred to as mean or variance *impact*. This limitation clearly threatens the internal validity of inferences made using those scores. It is critical then to consider testing for measurement invariance any time there are multiple groups under consideration.

Measurement invariance must also be considered when one wishes to integrate data across studies, as each study represents a distinct population and measures of the same constructs may differ to some degree between studies. Recently, modern demands for a more cumulative science of psychology have motivated the development of Integrative Data Analysis (IDA) to make inferences across multiple samples (Curran & Hussong, 2009). To ensure construct validity, the meaning of a construct must be constant across studies. This assumption implies measurement invariance, which is testable, most flexibly by moderated nonlinear factor analysis (MNLFA; Bauer, 2016; Bauer & Hussong 2009).

Measurement invariance requires that all item parameters are equivalent for score values to be directly comparable across groups (or in this case, studies). Only when the latent variables are on the same scale can one make inferences on group differences in the mean and variance or compare individual scores. With the GRM, assuming equality of the threshold parameters that separate the ordinal responses is really only sensible if each study uses the same response options. In the context of IDA, the use of identical measures is rare (Hussong, Curran, & Bauer, 2013a). Even if the measures are identical, differences in history, location, and other personal characteristics could violate the assumed equality of the thresholds *of the same items* across groups. To obtain the same response options across studies for these models, researchers often have to collapse the items into binary responses (no behavior vs. some behavior). Collapsing responses throws away valuable information and may reduce the power and the precision of our estimates. Thus, although current methods allow for IDA, the necessity of invariant response categories forces suboptimal analyses.

In addition to more faithfully modeling the underlying data generating process, OCFMs do not require invariant thresholds. First, it stands to note that the traditional multiple group approach used to fit latent variable models simply involves constraining and freeing parameters—something that is easily done for OCFMs. More importantly, under an OCFM, I treat the thresholds as known, which results in two important properties. First, I do not need to constrain the thresholds to be invariant across groups to identify the PMF. Second, as the PMFs of the items from both groups reference an underlying count, the use of an OCFM naturally results in the items having the same scale across populations. Thus, OCFMs are estimable and interpretable even when response options for the same item differ across studies. In preserving all available response options, the researcher stands to gain precision and power over traditional methods.

Given the advantages afforded by the use of a multiple groups OCFM, it is vital that such a model is formulated. I proceed by detailing how one can specify, identify, and estimate a multiple group OCFM. I then conclude with an empirical example which returns to the REAL-U data set to perform a mock integrative data analysis. I intend to demonstrate that the multiple groups OCFM can be fit in line with current psychometric methods, while at the same time determining whether there are advantages to the OCFMs unique specification.

Multiple Groups Ordinal Count Factor Model

A multiple groups model can be specified for groups given the same response scale or similar response scales with different ordinal count options. As such, the general approach will allow for thresholds to vary across populations, although the former case can be obtained by setting the thresholds to be the same across populations. As in the single population case, the general model for a multiple group OCFM consists of the auxiliary threshold model and the latent factor model. However, in the multiple group model, I allow each of these to vary as a function of population membership.

Specification for multiple groups. One first specifies the auxiliary threshold model by defining the underlying count random variable using its PMF:

$$P^{(g)}\left(Y_{ij}^{*(g)} = y_{ij}^{*(g)} \middle| \mu_{ij}^{(g)}, \boldsymbol{\Upsilon}_{j}^{(g)}\right) \tag{14}$$

where *g* indexes population membership. As before, the probability of the response option is the sum the probabilities for the counts in the corresponding interval

$$P\left(\mathrm{Y}_{\mathrm{ij}}^{(\mathrm{g})} = \mathrm{c}_{\mathrm{j}}^{(\mathrm{g})} \middle| \mu_{ij}^{(g)}\right) = \sum_{w\,=\,\min\{\mathrm{c}_{\mathrm{j}}^{(\mathrm{g})}\}}^{\max\{\mathrm{c}_{\mathrm{j}}^{(\mathrm{g})}\}} P^{(g)}\left(Y_{ij}^{*(g)} = w \middle| \mu_{ij}^{(g)}, \boldsymbol{\Upsilon}_{j}^{(g)}\right) \tag{15}$$

It should be mentioned that it is possible to include items which prompted responses in the form of raw counts. In that case, one needs to only model ordinal responses for those items without raw counts. For the raw count, the model still holds—it is as if each interval only includes a single count.

Then one specifies a non-negative link function from the latent variable to the mean of the underlying response function for each item.

$$h\left(\mu_{ij}^{(g)}\right) = \sum_{g=1}^{G} I_g(\beta_{0j}^{(g)} + \boldsymbol{\beta}_{1j}^{(g)}\boldsymbol{\theta}_{i}^{(g)}) \tag{13}$$

where G is the total number of groups, $\boldsymbol{\theta}_i^{(g)} \sim N(\boldsymbol{\xi}^{(g)}, \boldsymbol{\Phi}^{(g)})$, $\beta_{0j}^{(g)}$ is the j^{th} item's intercept for the g^{th} group, $\boldsymbol{\beta}_{1j}^{(g)}$ is the item's factor loadings for people in group g, I_g is an indicator function that is one for group g and zero for all other groups.

Turning our attention to OCFMs specifically, there are a number of ways DIF can manifest itself in these models. For example, in a ZINB OCFM, measurement invariance may be violated if the location, loading, or zero-inflation parameter differ across groups. To compare, complications arise with testing for DIF in a GRM with more than two categories because there are multiple location parameters. DIF testing in an OCFM does not require considering those issues, as there is only a single location parameter no matter the number of categories. This difference means that DIF testing in OCFMs does not become more onerous as the number of categories increases.

There are a number of ways to test for DIF in the general psychometric literature, and the uncertainty about the best method also applies to the OCFM. On one hand, one might start with a fully constrained model and use modification indices to free parameters between groups (Millsap, 2012). On the other hand, one might start with by freely estimated the parameters not constrained for identification and use Wald tests to select the invariant ones (Woods, Cai, & Wang, 2013). Thissen and colleagues would suggest constraining all items but one to be invariant across groups using LRTs to compare each item's model to the fully constrained model (Thissen, Steinberg, & Wainer, 1988).

These methods heretofore presented are not iterative, but iterative approaches exist and are often used when DIF may be widespread (Oort, 1998; Woods, 2009). One may begin with all items constrained to be equal across groups, fit models where one only frees one additional item, and test the models using an LRT. One would then select the item with the largest significant test statistic, free it in the next stage of the model, and repeat the process. The process ends when no new model rejects the LRT. I will implement this DIF selection procedure in my second empirical study. Of course, one can perform this process in the other direction: freeing all but a set of invariant items, and selecting an additional invariant item each time. Nevertheless, the relative benefits of different approaches to DIF testing continue to be the subject of debate and research in traditional psychometric models. Presumably, the relative performance of these approaches with the OCFM would not greatly differ from other contexts, such as linear CFA or GRM.

Identification for multiple groups. To properly identify OCFMs for multiple groups, one needs to ensure that one can obtain unique parameter estimates without using assumptions that incorrectly imply measurement invariance across the populations. Faced with these additional demands, one can still rely on the sufficient, but not necessary condition that the auxiliary threshold model and the latent factor model must be identified for the full model to be identified.

For the auxiliary threshold model to be identified, the rules for the single population model can be used. That is, one need at least q known thresholds to obtain unique estimates for q-1 parameters.

For the latent factor model to be identified, one needs to make sure that the scale for each latent variable is set. As mentioned earlier, I standardize the latent variable for the reference group to identify the model. Provided there is at least one invariant item, this constraint allows the mean and variance for the other two groups to be freely estimated. Moreover, this constraint does not require the user to specify *a priori* a specific invariant item—a necessary assumption made in the scaling indicator approach. Instead, this invariant item may be determined via data-driven DIF testing.

Although these conditions are necessary for the model to be identified, an identified model does not imply that the latent factors are comparable. For the latent variables to represent the same construct, measurement invariance is required. The necessary extent of this invariance is hotly debated. For direct comparison of factor means or scores across groups to be valid, invariance of both the factor loadings and intercepts is required (Millsap, 2012). It is possible, however, to make these comparisons under partial invariance, that is, when this factor loadings and intercepts are invariant for only a subset of the items (Byrne, Shavelson, & Muthén, 1989). In principal, only a single invariant item is necessary, but many find it unlikely that this item would be known *a priori* (Bollen, 1989). Without that knowledge, it is difficult to empirically select a single invariant item among a large number of items with DIF. As such, I agree that the presence of more invariant items is best to ensure measurement invariance and thus the comparability of factor means, variances, and scores (Kolen & Brennan, 2004). In practice, this implies that the fewer items that are identified as having DIF, the greater confidence one can have in the comparability of the factor means and scores across groups.

Estimation for multiple groups. As in the single population case, it is possible to estimate the model using maximum likelihood with the following likelihood:

$$L = \int_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \Pi_{g=1}^{G} \Pi_{i=1}^{N_g} \Pi_{j=1}^{J_g} \left[\Pi_{c=1}^{\mathrm{c}_\mathrm{j}^{(\mathrm{g})}} \left(\sum_{w\,=\,\min\{\mathrm{c}_\mathrm{j}^{(\mathrm{g})}\}}^{\max\{\mathrm{c}_\mathrm{j}^{(\mathrm{g})}\}} P^{(g)} \left(Y_{ij}^{*(g)} = w \middle| \mu_{ij}^{(g)}, \boldsymbol{\gamma}_j^{(g)} \right) \boldsymbol{\Phi}\,(\boldsymbol{\theta}) \right)^{I_{y_{ij}^{(g)}=c}} \right] d\boldsymbol{\theta} \qquad (16)$$

Taking the natural log, I get the log likelihood:

$$ll = \sum_{g=1}^{G} \int_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N_g} \sum_{j=1}^{J_g} \left[\sum_{c=1}^{\mathrm{C}_{\mathrm{j}}^{(\mathrm{g})}} I_{y_{ij}^{(g)}=c} \, ln \left(\sum_{w\,=\,\min\{\mathrm{c}_{\mathrm{j}}^{(\mathrm{g})}\}}^{\max\{\mathrm{c}_{\mathrm{j}}^{(\mathrm{g})}\}} P^{(g)} \left(Y_{ij}^{*(g)} = w \middle| \mu_{ij}^{(g)}, \boldsymbol{\gamma}_{j}^{(g)} \right) \phi(\theta) \right) \right] d\boldsymbol{\theta} \qquad (17)$$

Although the estimation of the model for multiple groups is more complex, it remains a direct extension of that for a single population, so researchers can use R to estimate the multiple groups OCFM. Code that demonstrates how to estimate such a model can be found in Appendix C.

Study 2

To demonstrate the utility of this approach, I return to the REAL-U study and the RAPI. In REAL-U, three perturbed versions of the RAPI were created with different stems. The original scale and the three perturbations were grouped into two batteries, Battery A and Battery B. Each participant was assigned a battery for each visit, resulting in four measurement conditions (i.e. first received A then received B, B then A, A then A, or B then B). Although there is a large number of potential comparisons, I will perform a mock integrative data analysis using two perturbed versions of the RAPI (Scenario 1, Battery A and Scenario 4, Battery B) from the first visit. As participants necessarily received either Battery A or Battery B, all participants can be included in this empirical example, maximizing the possible power (N=854). Compared to the Scenario 1 RAPI, the Scenario 4 RAPI has different instructions, but more importantly, different response options and item stems for all eighteen items common to both scenarios that can be found in Table 6. Note that all but one of the 23 original items were used in the Study 1. I continue to omit item 5 in Study 2. In addition, there are two different sets of response options used in Scenario 4. Of the six possible comparisons of the four versions of the RAPI collected, this comparison is the most extreme. Thus, generating commensurate measures across these two versions should be a challenge for both models.

Not only will this empirical example evaluate the potential benefits of OCFMs in integrative data analysis, but it will also investigate the sensitivity of young adults to the presentation of questions regarding the consequences of their alcohol use. In addition to the various perturbed item scales, the REAL-U data set uniquely provides us the opportunity to compare participants' scores across scales and time points. Unlike in a real integrative data analysis, I can estimate the test-retest reliability across measurement condition. For those who received the same battery at both visits, the correlation between their scores is a true test-retest reliability. For those who did not, and thus received different batteries, the correlation is an estimate of the stability of the scores after being perturbed by measurement. Comparing the two perturbed conditions to the two unperturbed conditions allows me to test whether the perturbation actually resulted in less reliable scores. Juxtaposing those comparisons across model allows me to determine whether the OCFM produces more reliable scores and whether the OCFM is more resistant to those perturbations than the GRM.

Finally, this study will allow for the evaluation of the precision of each model's scores. This precision, considered in the form of standard errors, should result in more power to detect meaningful relationship between the construct and others. As IDA is rarely the last modeling step, a scoring model that produces precise scores is valued. The ability of an OCFM to incorporate more information than a GRM should result in smaller standard errors. If so, that additional precision would underscore the utility of OCFMs in integrative data analysis.

Here I do not consider the linear factor analysis given its demonstrably inferior performance in Study 1. Thus, through this empirical demonstration, I seek to accomplish three aims: to demonstrate DIF detection in an OCFM, to assess the stability of each model's scores across time, RAPI version, and DIF modeling, and to determine whether the full data OCFM provides more precise factor scores than the harmonized data GRM. Specifically, I hypothesize that

1. there will be more DIF detected in the OCFM because the categories were not collapsed, thus providing more power than the binary outcomes of the GRM;
2. the DIF corrected models under each measurement condition will produce highly correlated scores, but these scores will not be similar enough to justify using the GRM over the OCFM;
3. even when DIF is included in both models, the OCFM scores will show higher test-retest reliability than the GRM scores due to the additional information in the model;
4. the OCFM score correlations will be more stable across different measurement occasions than those of the GRM due to the additional information in the model; and
5. the OCFM scores will have smaller standard errors than the GRM scores due to the additional information in the model.

Methods

Fitting the Graded Response Model. As data for a GRM must have ostensibly the same thresholds across groups for interpretable results, to use a GRM the response options must be collapsed into categories that have equal ranges, a process referred to as item harmonization (Hussong, Curran, & Bauer, 2013b). For the two RAPI scenarios under consideration, some but not all items can be harmonized in this way, and so I allow for imperfect harmonization to retain these items. To do so, I collapse the items from the first response scale into *never* vs. *at least once*. Those items from the second response scale cannot be collapsed as such, so I collapse them into *less than twice* vs. *more than twice*. This clearly suboptimal solution is detailed in Table 6.

Table 6

Response Options by Scenario

	Scenario 1 Response Options	Scenario 4 Response Options	Harmonized Response Options
Items 1, 4, 6, 7, 8, 14, 15, and 17	None 1-2 times 3-5 times More than 5 Times	Never Once Twice 3-5 times 6-9 times 10 or more times	None 1 or more times
Items 2, 3, 9, 10, 11, 12, 13, 16, and 18	None 1-2 times 3-5 times More than 5 Times	0-2 times 3-4 times 5-9 times 10 or more times	0-2 times 3 or more times

However, if these items had non-count response options, even such undesirable harmonization would be impossible. For example, if Scenario 4 had options from “strongly agree” to “strongly disagree”, it would be impossible to harmonize that scale with options from “not at all like me” to “almost exactly like me”. With the outcomes (imperfectly) harmonized, I fit the GRM. I then use sequential likelihood ratio tests to detect DIF across the two scenarios.

Fitting the Ordinal Count Factor Model. As the OCFM can be fit to groups with different known thresholds, the analysis proceeds without collapsing any response options. Given that the ZINB model was selected in Study 1, I refit that model in this study as well. I do not prune the zero-inflation component for any of the items so as to allow for differences across study. I perform sequential likelihood ratio tests to detect DIF. To do this, I begin with a completely invariant model. I then free each item and compare each model with the invariant model. If more than one model was significantly different, the item which produces the greatest improvement in log-likelihood is freed. This process continues until freeing an item does not significantly improve fit.

Scoring. After both models have been specified, I obtain EAPs from the models with DIF and without DIF. I score the second visit data using the estimates obtained from the models fit to the first visit data; that is, I do not re-estimate the model for the second visit. In total, I will score participants using four models (i.e. DIF included/excluded and OCFM/GRM).

Comparison. I begin by evaluating the first hypothesis—that the OCFM will detect more DIF than the GRM. I perform no test, given the exploratory nature of this study. Nevertheless, I will report which items demonstrate DIF for both items and comment on the result qualitatively. Note that there is no comparison of goodness-of-fit between the models. The collapsing of the data necessary to fit the GRM (but not OCFM) means that the two models are fit to different data. Yet if the OCFM has more items with DIF, then the first hypothesis is supported.

To evaluate my second hypothesis, I then calculate correlations between the DIF corrected OCFM and GRM scores across all four measurement scenarios and two visits for a total of eight correlations. If the correlations are extremely high (i.e. ~.98) then perhaps one might argue that the scores are roughly equivalent. If the correlations are not high at all (i.e. <.70) then the validity of either model may be thrown into question. For my hypothesis to be supported, the majority of the correlations should lie between .70 and .98, suggesting that, although the scores linearly approximate each other, they do not approximate each other well enough to render the choice of the model inconsequential.

I then consider whether the correlations across visits are the same for the DIF-corrected GRM and OCFM. For the measurement conditions of the same battery across both visits, this correlation estimates the test-retest reliability, and thus this comparison tests whether the test-retest reliability is higher for the OCFM than the GRM. For those conditions with different batteries, this comparison tests whether test-retest reliability is higher when the measurement changes. This procedure will test my third hypothesis.

Having considered the effect of model on test-retest reliability, I move on to consider whether the OCFM is more resistant to measurement perturbations than the GRM, the fourth hypothesis. To test this, I compare the test-retest correlations of the perturbed measurement conditions to that of the unperturbed measurement conditions for each DIF-corrected model. If the correlations are significantly different for the scores from the GRM but not for the scores from the OCFM, then my fourth hypothesis will be supported. Otherwise, if both are significant, then the measurement perturbation significantly worsened the reliability of the scores. If both are non-significant, then it is possible that both models adequately controlled for the measurement perturbations.

Finally, I turn to the standard errors. I first plot the standard errors of the GRM scores and those of the OCFM against the OCFM scores. I then fit a linear meta-model to the difference between a participant's OCFM standard error and their GRM standard error, using the OCFM score as a predictor. This model allows me to test whether the OCFM standard errors for this model are meaningfully smaller than those of the GRM, and whether that difference depends on the person's level of the latent variable. If average difference between the OCFM and the GRM standard errors is significant, then I find support for my fifth and final hypothesis.

Results:

I completed sequential DIF testing for both the ZINB OCFM and GRM to evaluate my first hypothesis. Consistent with my hypothesis, the OCFM identified more items with DIF than the GRM. Although both models identified DIF in items 1, 2, 3, 4, 7, 12, 14, and 15, the GRM also indicated DIF in items 16 and 18, whereas the OCFM also indicated DIF in items 9, 11, and 12. This widespread DIF should not be seen as surprising in an integrative data analytic setting. Nevertheless, both of the models shared four invariant items (6, 8, 10, 17) and a few items unique to that model. This number of invariant items is considered sufficient to provide partial measurement invariance across the two studies (Kolen & Brennan, 2004). The difference in DIF identification, though in the direction predicted, does not seem large enough to support the first hypothesis. As such, although sequential DIF testing identified enough invariant items to link the two mock studies, it did not provide enough evidence to make a claim that the OCFM identifies more DIF items than the GRM in general.

Before considering the rest of my hypotheses, I present the model estimates. Final parameter estimates for the OCFM can be found in Tables 7 and those for the GRM can be found in Table 8. As the mean and variance of Scenario 1 was standardized in both models, I can examine whether the mock study design resulted in mean or variance impact. Note that because the subjects were randomly assigned to a measurement battery, there should be no impact. The OCFM estimate of the mean and variance of those in Scenario 4 is -0.237 and 1.257, and the GRM was similar at -0.195 and 1.357, respectively. Although I had no hypothesis about the result of this manipulation, the fact that it exists even when correcting for DIF is potentially concerning. Yet the purpose of this study is not to evaluate the internal validity of IDA; rather, it is to evaluate the efficacy of the multiple group OCFM as compared to traditional methods. Moreover, this mean and variance impact could be successfully modeled, and thus adjusted for, by including study impact—a common practice (Curran & Hussong, 2009). As both models found similar mean and variance impact, the validity of the model comparison is not threatened.

With the models estimated and described, I turn my attention to the factor scores, the focus of hypotheses 2, 3, and 4. I estimated EAPs for the eight models, using the visit 1 model estimates to estimate scores at each visit. Correlations between the scores pooled across measurement conditions can be found in Table 9. Tables for each measurement condition can be found in Appendix D.

Table 9

Multiple Group Factor Score Correlations Pooled over Scenarios

			Visit 1				Visit 2		
			GRM		OCFM		GRM		OCFM
			No DIF	*DIF*	*No DIF*	*DIF*	*No DIF*	*DIF*	*No DIF*
Visit 1	*GRM*	*DIF*	0.934						
	OCFM	*No DIF*	0.909	0.964					
		DIF	0.924	0.961	0.996				
Visit 2	*GRM*	*No DIF*	**0.684**	0.764	0.755	0.749			
		DIF	0.637	**0.709**	0.700	0.696	0.949		
	OCFM	*No DIF*	0.694	0.780	**0.807**	0.797	0.938	0.859	
		DIF	0.662	0.744	0.782	**0.772**	0.857	0.729	0.965

Test-retest correlations within model are bolded

Table 7

Multiple Group Ordinal Count Factor Model Estimates

Item Stems	Loading		Intercept		Overdispersion parameter		Zero-Inflation parameter	
Scenario	S1	S4	S1	S4	S1	S4	S1	S4
1. Got into fights with other people (friends, relatives, strangers)	-0.9	-1.94	1.56	1.44	1.15	0.37	-10.76	-0.20
2. Went to work or school high or drunk	-1.0	-0.93	2.01	1.07	2.26	0.83	-11.26	-7.36
3. Caused shame or embarrassment to someone	-0.83	-0.49	1.50	0.79	0.45	0.26	-11.40	-9.77
4. Neglected your responsibilities	-0.17	-1.42	1.32	1.64	0.97	1.74	-11.98	-0.70
6. Felt that you needed more alcohol than you used to in order to get the same effect	-0.50	-0.50	1.62	1.62	1.70	1.70	-2.08	-2.08
7. Tried to control your drinking (tried to drink only at certain times of the day or in certain places, that is, tried to change your pattern of drinking)	-1.33	-1.39	2.39	1.49	4.20	0.05	-10.1	-0.45
8. Had withdrawal symptoms, that is, felt sick because you stopped or cut down on drinking	-3.89	-3.89	2.43	2.44	20.07	20.07	-7.64	-7.64
9. Noticed a change in your personality	-1.42	0.10	1.44	1.04	5.85	0.80	-8.88	-8.52
10. Felt that you had a problem with alcohol	-2.97	-2.98	2.09	2.09	0.43	0.43	-1.84	-1.84
11. Wanted to stop drinking but couldn't	-4.54	-2.51	2.60	1.88	0.00	0.00	0.55	-9.07
12. Suddenly found yourself in a place that you could not remember getting to	-1.24	0.57	2.08	1.07	0.46	0.21	-2.82	-2.11
13. Passed out or fainted suddenly	-1.0	-0.01	1.42	1.21	0.34	0.19	0.07	-1.23
14. Had a fight, argument, or bad feeling with a friend	-0.80	-2.49	1.46	1.77	0.50	2.48	-2.58	-6.61
15. Kept drinking when you promised yourself not to	-0.94	-3.45	1.20	2.64	0.55	1.67	0.08	-0.37
16. Felt you were going crazy	-2.9	-2.96	2.42	2.42	1.58	1.59	-0.28	-0.28
17. Felt physically or psychologically dependent on alcohol	-4.72	-4.72	3.03	3.04	0.62	0.62	-0.55	-0.55
18. Was told by a friend, neighbor or relative to stop or cut down drinking	-3.02	-3.03	2.20	2.21	1.89	1.89	-8.70	-8.70

Table 8

Multiple Group Graded Response Model Parameter Estimates

Item Stems	Loading		Intercept	
Scenario	S1	S4	S1	S4
1. Got into fights with other people (friends, relatives, strangers)	1.531	1.421	-0.905	-2.795
2. Went to work or school high or drunk	1.764	1.637	-1.326	-3.615
3. Caused shame or embarrassment to someone	1.97	1.786	-0.633	-3.508
4. Neglected your responsibilities	1.502	1.338	-0.177	-2.24
6. Felt that you needed more alcohol than you used to in order to get the same effect	1.465	1.465	-1.086	-1.086
7. Tried to control your drinking (tried to drink only at certain times of the day or in certain places, that is, tried to change your pattern of drinking)	2.052	1.395	-2.045	-2.028
8. Had withdrawal symptoms, that is, felt sick because you stopped or cut down on drinking	1.634	1.634	-4.367	-4.367
9. Noticed a change in your personality	1.336	1.336	-1.959	-1.959
10. Felt that you had a problem with alcohol	2.298	2.298	-3.355	-3.355
11. Wanted to stop drinking but couldn't	3.317	3.317	-7.028	-7.028
12. Suddenly found yourself in a place that you could not remember getting to	2.199	2.199	-1.225	-1.225
13. Passed out or fainted suddenly	1.187	2.056	-1.963	-2.551
14. Had a fight, argument, or bad feeling with a friend	1.464	1.566	-0.725	-2.766
15. Kept drinking when you promised yourself not to	1.164	2.138	-1.99	-4.252
16. Felt you were going crazy	1.919	2.313	-3.585	-5.871
17. Felt physically or psychologically dependent on alcohol	2.653	2.653	-5.149	-5.149
18. Was told by a friend, neighbor or relative to stop or cut down drinking	2.152	2.621	-3.037	-6.671

Let us first consider my second hypothesis, the one focused on whether the GRM and the OCFM would produce meaningfully different score estimates. The DIF-adjusted score correlations range from .94 to .97 at the first visit to 0.73 to .94 at the second. These correlations are large for the first visit, and thus may suggest that the scores are roughly comparable. However, the lack of a very strong correlation at the second visit suggests that such a claim is tenuous. Thus, the OCFM and the GRM, although highly correlated at the first visit, may produce scores that may lead to different inferences when included in models—in accordance with my second hypothesis.

Next, I examine the stability of the factor score estimates across time as affected by model and condition, the stability of which I predicted to be higher for the OCFM as compared to the GRM. To do so, for each measurement condition I tested whether the OCFM's test-retest reliability was higher than the GRM's test-retest reliability (Steiger, 1980). All differences favored the OCFM, and all differences besides that for those who received Scenario four twice were insignificant ($p_{11} = 0.93, p_{14} = 0.1, p_{41} = 0.08, p_{44} = 0.01$). Although the OCFM scores were more highly correlated than the GRM scores, only when one battery was presented twice was that difference significant. Thus, I do not find strong support for my third hypothesis.

Having considered whether the OCFM results in higher test-retest reliability across all measurement conditions relative to the GRM, I now turn my attention to whether the reliability of the OCFM scores were more resistant to measurement perturbations. I performed the same tests for the correlations before, comparing each perturbed condition (i.e. received Scenario 1 then Scenario 4 and received Scenario 4 then Scenario 1) to each unperturbed one (i.e. only received Scenario 1 and only received Scenario 4) fit with the same model. Thus, I compared the correlation of each model's scores between visit one and visit two for people who received the same battery to those who received different batteries. None of the comparisons made were significant (min(p)=.68). A lack of a significant difference does not imply that the model perturbation had the same effect for both models. However, it does suggest that accounting for DIF in the multiple group model successfully represented those measurement differences. Given this performance from both models, I do not find support for my fourth hypothesis.

Finally, I turn my attention to the standard errors of the scores for both models. A plot of the standard errors by OCFM EAPs (Figure 5) suggests that the OCFM produces scores with smaller standard errors, regardless of the visit or measurement condition. Plotting the standard errors against the GRM EAPs result in a similar pattern, although the GRM EAPs demonstrate a clear ceiling for scores that are either all zeros or all ones.

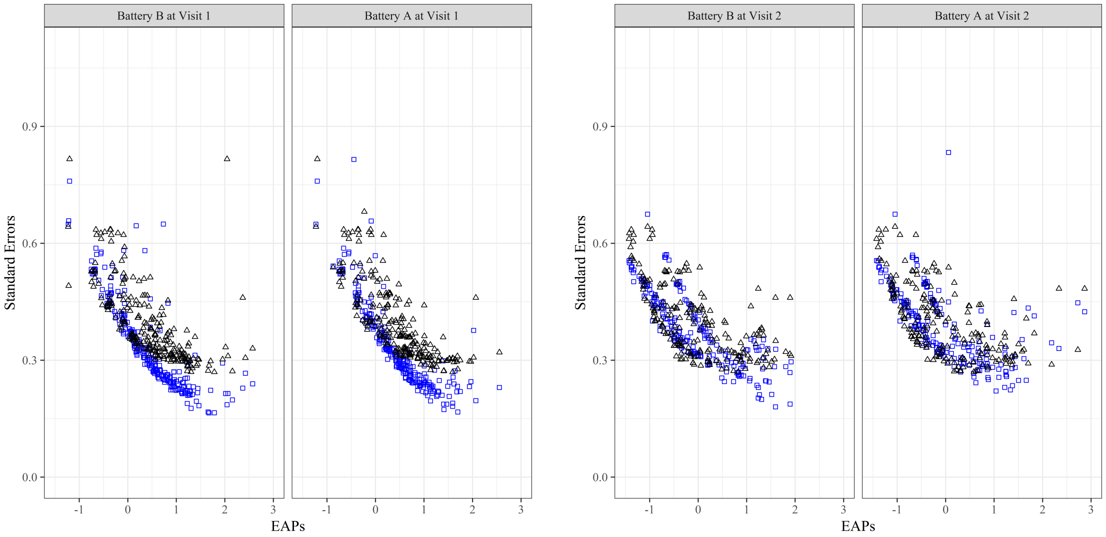

Figure 5: Standard errors plotted against the EAP from the DIF-corrected OCFM. Squares represent standard errors from the OCFM; triangles represent standard errors from the GRM. Battery A means participants received Scenario 1 and Battery B means participants received Scenario 2.

A linear regression meta-model supports these suspicions with the outcome as the difference in standard error between the OCFM and the GRM pooled across condition. I fit two meta-models—one for each visit. Not only is there a significant difference between the standard errors of the OCFM and GRM ($\beta_0 = -0.036, p < 0.001$), but this difference increases as OCFM EAP increases in level ($\beta_1 = -0.066, p < 0.001$) at visit one. In the model for the second visit, the OCFM EAP predicts a decrease in the difference between the standard errors ($\beta_1 = 0.024, p < 0.001$) at visit 2. The result for visit one should not be surprising, as the OCFM incorporates more information on those with more extreme values on the scale. However, the result for visit 2 complicates that interpretation. The OCFM scores for visit 1 were 0.4 larger than those at visit 2 ($t = 20.8, p < 0.001$). As such, it still might hold that the OFM incorporates more information at higher levels of the latent variable, but the second visit simply does not have the range to detect it. Therefore, this result not only supports the fifth hypothesis, that the OCFM produces more precise scores overall, but also suggests that this increase in precision can be magnified at higher levels of the latent variable.

Discussion

The results of the second empirical example supports the use of the multiple groups OCFM, especially in integrative data analytic settings. This support comes from three main findings: the ability to fit the model, the relative similarity of OCFM scores to GRM scores, and the significant gains in precision that result from using an OCFM over a GRM.

Regardless of a model's properties, it is vital that one can fit it in a straightforward way using accessible software. The second empirical example demonstrates just that. Although the estimation was computationally demanding, it is possible to fit such a model in R. As such, there is little reason why researchers who already use psychometric models cannot use an OCFM. The results that support the first hypothesis, that the multiple group OCFM may provide more power to detect DIF, should also make the OCFM attractive to psychometricians.

The second main finding lies in the similarity of OCFM scores to GRM scores. It may seem counterintuitive that the strong correlation between OCFM scores and GRM scores supports the use of the OCFM. The results that support of the second hypothesis, that the scores would be highly, but not perfectly related, might lead the reader to assume that the OCFM does not really change the scores. Moreover, the results contrary to hypotheses three and four suggest that the OCFM does not produce significantly more stable scores across time (i.e. higher test-retest reliability) or across measurement condition than a GRM, although the correlations between OCFM scores remain consistently higher. That said, let us consider a counterfactual situation: that in which the OCFM scores at the same visit and measurement condition were only weakly related to the GRM scores. Such a result would imply that OCFM orders people in a meaningfully different way than the GRM does. One would then have to be cautious about using an OCFM, because it could produce different results than a GRM. Yet the OCFM produces scores that are very similar to the GRM, generating confidence that the additional assumption does not meaningfully change the ordering of participants. That said, the correlation is not perfect, which means that the use of an OCFM still may change inferences, albeit potentially in a less dramatic way. This balance, between perfect resemblance and noticeable divergence, manages both to validate the multiple group OCFM and to support its use.

The third major finding, that of the increased precision of scores in the OCFM, cannot be overstated. The smaller standard errors of the OCFM scores, supporting the fifth hypothesis, underscore its potential utility in traditional multiple group analysis and in integrative data analysis. It seems probable that the additional response categories allow the multiple groups OCFM to produce smaller standard errors than the GRM. It should be noted that these additional responses do not require an assumption beyond those made by a single group OCFM. Rather, the ability to incorporate these additional assumptions follow from the very definition of an OCFM. Thus, the added precision not only supports the use of OCFMs in integrative data analysis, but also in traditional multiple group models.

Nevertheless, a key question in multiple group OCFMs is in how to interpret DIF, which remains ambiguous, but more interpretable than in the GRM. If one group has a larger intercept, it could mean that group engages in more of that behavior than the other. On the other hand, it could also mean that group interpreted the description of that behavior differently, which led them to count additional instances as that behavior. A larger slope may mean that the behavior increases faster for that group, or that as the latent variable increases, people in that group simply remember more instances of that behavior. Either interpretation is valid, underscoring the need to explore such results through additional qualitative and quantitative research. This ambiguity does not threaten the utility of OCFMs, rather it underscores the relative utility of an OCFM over a GRM. Different thresholds for groups are difficult to interpret, especially if those thresholds are counts. One might say that it is simply more difficult for one group to endorse higher levels of that behavior, but what does that mean substantively, let alone clinically? Thus, the question of DIF in a multiple group OCFM demonstrates how the OCFM produces more interpretable inferences than traditional approaches.

Conclusion

Ordinal Count Factor Models represent a next step in advancing the measurement of behavioral counts. Not only do they more faithfully represent the data generating process, they require as many parameters as a linear CFA and often fewer parameters than a GRM with the same number of response options. Of course, such a model is not the cure-all. Fitting certain models may produce extremely biased parameter estimates, and the use of an OCFM does nothing beyond traditional psychometric models in illuminating the latent structure of the indicators. However, their ability to model DIF in traditional and integrative data analysis settings represents a clear advantage for the type of research being currently performed in the behavioral and health sciences. Moreover, the code developed through the course of this document aims to make these modeling approaches available to applied researchers across disciplines. Yet the impact of this thesis can be summarized in its two main results: theoretical development of the general single group and multiple group OCFMs, and empirical evaluation of the single group OCFM through Studies 1 and 2.

In this paper, I theoretically develop the OCFM while grounding it in both cognitive psychology and psychometrics. Cognitive psychology has shown that under certain conditions, people respond to ordinal count items by counting up past incidents. These conditions match many experiences and behaviors researched in developmental psychopathology. Psychometrics provides a framework to model these behaviors or experiences as being caused by a set of latent variables. With the introduction of an ordinal count response function, this framework provides a set of tests and approaches that render the OCFM usable in myriad ways. This general form of the OCFM thus represents a novel extension to the latent variable modeling literature while remaining motivated by it.

In Studies 1 and 2, I interrogate the utility of the single and multiple groups OCFM. First and foremost, the ability to fit these models in open-source, widely available software cannot be understated. This convenience, in combination with their parsimony, suggests that these models can used in applied research. Moreover, in these studies the OCFM fit better than standard methods and generated smaller standard errors. Finally, in the multiple groups case, I demonstrate just how powerful the assumption of a latent count is in integrative data analysis. The OCFM creates invariance where there previously was not, and allows for sparse or empty categories too, all from the assumption of a latent count. To review, the invariance comes from the assumption that the responses are on the metric of the underlying count, allowing for the estimation of the model without collapsing categories. As the thresholds for each category are known through the assumption of a count, we do not need to collapse thresholds for the model to be identified, unlike in a GRM (Ostini & Nering, 2006). For these reasons, the empirical studies provide clear evidence that the OCFM should be strongly considered—if not preferred—when fitting ordinal count items.

The use of empirical data examples clearly limits the generalizability of these inferences. First, the generating models and their parameter values are unknown. As such, it is impossible to estimate relative bias and other finite sample properties under a variety of conditions, including a misspecified model. Using the aforementioned model fit statistics as standards is weak, as it tells us not whether the data comes from the model, or even whether the model fits in an absolute sense—barring those statistics used to evaluate the linear CFA. Nevertheless, these findings remain a promising first step in understanding and applying the OCFM to real data.

Of course, more work should be done on the OCFM to broaden its utility and determine its properties. A primary issue is generating standard errors for the model parameters analytically or through another method like the bootstrap. I conducted a preliminary evaluation of the profile likelihood confidence intervals within mirt using the feasibility study models. The approach within mirt was not sufficient, and often generated confidence intervals with coverage rates of less than 50%. Given such a poor outcome, I did not use the profile likelihood confidence intervals. A more theoretical issue is the performance of the OCFM when there is, in fact, an underlying GRM. On one hand, OCFM may generate roughly equivalent scores with smaller standard errors. On the other hand, the OCFM may lead to biased inferences—regardless if the goal is only to obtain factor scores and not necessarily interpret the measurement model. Other finite sample properties, such as the benefits of treating counts as bounded or using a Poisson instead of a negative binomial PMF may also prove useful. Finally, determining the best practices for fitting an OCFM in PROC NLMIXED remains an open and important question.

The OCFM represents a clear extension to psychometrics, one that can improve research from public health to psychology. An assumption that there is an underlying count distribution, one congruent with the data and psychological theory, empowers the user to fit more parsimonious models and find invariance in places traditionally deemed hopeless. For those reasons and more, such an assumption may be not only justifiable, but also prudent.

APPENDIX A: REAL-U CROSSWALK FOR THE RAPI (FROM REAL-U CODEBOOK)

	Scenario 1 (Battery A)	Scenario 4 (Battery B)
Directions	Different things happen to people while they are drinking ALCOHOL or because of their drinking. Indicate how many times each of the following things happen to you WITHIN THE PAST YEAR/ AT SOME POINT IN YOUR LIFE.	Different things happen to people while they are drinking ALCOHOL or because of their drinking. Indicate how many times each of the following things happen to you WITHIN THE PAST YEAR/ AT SOME POINT IN YOUR LIFE.
Response Scale	None (0), 1-2 times (1), 3-5 times (2), More than 5 Times (3)	Never (0), Once (1), Twice (2), 3-5 times (3), 6-9 times (4), 10 or more times (5) (for unhighlighted items); 0-2 times (0), 3-4 times (1), 5-9 times (2), 10 or more times (3) (for highlighted items)
#		
1	Got into fights with other people (friends, relatives, strangers)	Gotten into physical fights when drinking
2	Went to work or school high or drunk	Gone to class or a job when drunk
3	Caused shame or embarrassment to someone	Made others ashamed by your drinking behavior or something you did when drinking
4	Neglected your responsibilities	Neglected your obligations, your family, or your work for two or more days in a row because you were drinking
5	Relatives avoided you	Family members rejected you because of your drinking
6	Felt that you needed more alcohol than you used to in order to get the same effect	Needed to drink more and more to get the effect you want
7	Tried to control your drinking (tried to drink only at certain times of the day or in certain places, that is, tried to change your pattern of drinking)	Tried to cut down or quit drinking or using alcohol Have you tried to cut down or quit drinking or using alcohol or other drugs?
8	Had withdrawal symptoms, that is, felt sick because you stopped or cut down on drinking	Felt sick, shaky or depressed when you stopped drinking
9	Noticed a change in your personality	Acted in a very different way or did things you normally would not do because of your drinking
10	Felt that you had a problem with alcohol	Thought you might have a drinking problem
11	Wanted to stop drinking but couldn’t	Tried unsuccessfully to stop drinking
12	Suddenly found yourself in a place that you could not remember getting to	Awakened the morning after some drinking the night before and could not remember a part of the evening.
13	Passed out or fainted suddenly	Passed out after drinking

14	Had a fight, argument, or bad feeling with a friend	Drinking created problems between you and a near relative or close friend
15	Kept drinking when you promised yourself not to	Could not stop drinking without difficulty after one or two drinks
16	Felt you were going crazy	Your drinking made you feel out of control even when you were sober
17	Felt physically or psychologically dependent on alcohol	Thought you were dependent on alcohol
18	Was told by a friend, neighbor or relative to stop or cut down drinking	Near relative or close friend worried or complained about your drinking
19	Not able to do your homework or study for a test	
20	Missed out on other things because you spent too much money on alcohol	
21	Missed a day (or part of a day) of school or work	
22	Had a fight, argument, or bad feeling with a family member	
23	Had a bad time	

APPENDIX B: ORDINAL COUNT FACTOR MODELS IN A SINGLE POPULATION

Loading Packages

```
library(mirt)

library(ggplot2)
```

Making the Item Responses

```
# NB2

name <- 'NB'
par  <- c(b0=1, b1=0, alpha=1)
est  <- c(T,T,T)
P_nb_rapi <- function(par, Theta, ncat){
  b0    <- par[1]
  b1    <- par[2]
  alpha <- par[3]
  mu    <- exp(b0 + b1*Theta)
  P0 <- pnbinom(0, size=1/alpha, mu=mu)
  P1 <- pnbinom(2, size=1/alpha, mu=mu) - pnbinom(0, size=1/alpha, mu=mu)
  P2 <- pnbinom(5, size=1/alpha, mu=mu) - pnbinom(2, size=1/alpha, mu=mu)
  P3 <- 1 - pnbinom(5, size=1/alpha, mu=mu)
  return(cbind(P0, P1, P2, P3))
}

NB <- createItem(name, par=par,est=est,P=P_nb_rapi,   lbound=c(-Inf,-Inf,0.00
001))


# OZINB

name <- 'ZINB'
par  <- c(b0=1, b1=1, alpha=1, gam0=0)
est  <- c(T,T,T,T)
P_zinb <- function(par, Theta, ncat){
  b0    <- par[1]
  b1    <- par[2]
  alpha <- par[3]
  gam0  <- par[4]

  mu    <- exp(b0 + b1*Theta)
  pp    <- 1 / (1+exp(gam0))

  P1 <- pp * (pnbinom(2, size=1/alpha, mu=mu) - pnbinom(0, size=1/alpha, mu=m
u))
```

```
  P2 <- pp * (pnbinom(5, size=1/alpha, mu=mu) - pnbinom(2, size=1/alpha, mu=m
u))


  P3 <- pp * (1 - pnbinom(5, size=1/alpha, mu=mu))

  ret <- cbind(1-P1-P2-P3, P1, P2, P3)
  ret <- ifelse(ret > (1-1e-7), (1-1e-7), ret)
  ret <- ifelse(ret < 1e-7, 1e-7, ret)
  return(ret)
}

ZINB <- createItem(name, par=par, est=est, P=P_zinb, lbound=c(-Inf,-Inf,0.000
01,-Inf))

# Beta Binomial

name <- 'BB'
par  <- c(b0=1, b1=1, gamma=.5)
est  <- c(T,T,T)
P_bb <- function(par, Theta, ncat){
  b0    <- par[1]
  b1    <- par[2]
  gamma <- par[3]
  mu    <- 1 / (1+exp(-1 *(b1*Theta + b0)))
  ig    <- 1/gamma
  #0, 1-2, 3-4, 5-7

  P0 <- beta((0 + mu*(ig-1)), ((1-mu)*(ig-1) + 7 - 0))/beta((mu*(ig-1)), ((1-
mu)*(ig-1)))


  P1 <- 7  * beta((1 + mu*(ig-1)), ((1-mu)*(ig-1) + 7 - 1))/beta((mu*(ig-1)),
((1-mu)*(ig-1))) +
        21 * beta((2 + mu*(ig-1)), ((1-mu)*(ig-1) + 7 - 2))/beta((mu*(ig-1)),
((1-mu)*(ig-1)))


  P2 <- 35 * beta((3 + mu*(ig-1)), ((1-mu)*(ig-1) + 7 - 3))/beta((mu*(ig-1)),
((1-mu)*(ig-1))) +
        35 * beta((4 + mu*(ig-1)), ((1-mu)*(ig-1) + 7 - 4))/beta((mu*(ig-1)),
((1-mu)*(ig-1)))

  ret <- cbind(P0, P1, P2, 1-P0-P1-P2)
  ret <- ifelse(ret > (1-1e-8), (1-1e-8), ret)
  ret <- ifelse(ret < 1e-8, 1e-8, ret)
  return(ret)
}
```

```r
BB <- createItem(name, par=par, est=est, P=P_bb, lbound=c(-Inf,-Inf,0.00001))

# Zero-Inflated Beta Binomial
name <- 'ZIBB'
par  <- c(b0=1, b1=1, gamma=.5, d0=0)
est  <- c(T,T,T,T)
P_zibb <- function(par, Theta, ncat){
  b0    <- par[1]
  b1    <- par[2]
  gamma <- par[3]
  d0    <- par[4]

  mu    <- 1 / (1+exp(-1 *(b1*Theta + b0)))
  pp    <- 1 / (1+exp(d0))
  ig    <- 1/gamma
  #0, 1-2, 3-4, 5-7

  P0 <- (1-pp)+(pp * beta((0 + mu*(ig-1)), ((1-mu)*(ig-1) + 7 - 0))/beta((mu*(ig-1)), ((1-mu)*(ig-1))))
  P1 <- pp * (7  * beta((1 + mu*(ig-1)), ((1-mu)*(ig-1) + 7 - 1))/beta((mu*(ig-1)), ((1-mu)*(ig-1))) +
              21 * beta((2 + mu*(ig-1)), ((1-mu)*(ig-1) + 7 - 2))/beta((mu*(ig-1)), ((1-mu)*(ig-1))))
  P2 <- pp * (35 * beta((3 + mu*(ig-1)), ((1-mu)*(ig-1) + 7 - 3))/beta((mu*(ig-1)), ((1-mu)*(ig-1))) +
              35 * beta((4 + mu*(ig-1)), ((1-mu)*(ig-1) + 7 - 4))/beta((mu*(ig-1)), ((1-mu)*(ig-1))))

  ret <- cbind(P0, P1, P2, 1-P0-P1-P2)
  ret <- ifelse(ret > (1-1e-8), (1-1e-8), ret)
  ret <- ifelse(ret < 1e-8, 1e-8, ret)
  return(ret)
}

ZIBB <- createItem(name, par=par, est=est, P=P_zibb, lbound=c(-Inf,-Inf,0.00001,-Inf))

#Fitting the model

fit <- mirt(data,
            number of factors,
            rep('Name', number of items) ,
            customItems=list(Name=Name))
```

APPENDIX C: ORDINAL COUNT FACTOR MODELS IN MULTIPLE GROUPS

The changes in the code from Appendix B are easily applied across different types of count model. As such, I only demonstrate how to fit models for the unbounded count, using the cut-points for Scenario 1 and Scenario 3 from REAL-U.

Loading Packages and Formatting Data

```
library(mirt)

library(ggplot2)

data <-merge(data_studya,data_studyb,all=T)
```

I merge the data so that I can estimate both measurement models at the same time. In creating our items, I need to do two additional things. The first is create two types of items with different cut points. The second is to specify which latent variable loads on which item. The *mirt* package automatically constrains the covariance of these latent variables to be zero, thus identifying the latent variable model. It also constrains the mean and variance of eta for both groups to one and zero—one can change this by directly editing the model object, which we do below.

Making the Item Responses Never (0), Once (1), Twice (2), 3-5 times (3), 6-9 times (4), 10 or more times (5) (for unhighlighted items); 0-2 times (0), 3-4 times (1), 5-9 times (2), 10 or more times (3) (for highlighted items)

```
# NB2_pop1

name <- 'NB'
par  <- c(b0=1, b1=0, alpha=1)
est  <- c(T,T,T)
P_nb_rapi_1 <- function(par, Theta, ncat){
  b0    <- par[1]
  b1    <- par[2]
  alpha <- par[3]

 Theta <- Theta[1]


  mu    <- exp(b0 + b1*Theta)
  P0 <- pnbinom(0, size=1/alpha, mu=mu)

  P1 <- pnbinom(1, size=1/alpha, mu=mu)
  P2 <- pnbinom(2, size=1/alpha, mu=mu) - pnbinom(0, size=1/alpha, mu=mu)
  P3 <- pnbinom(5, size=1/alpha, mu=mu) - pnbinom(2, size=1/alpha, mu=mu)

  P5 <- pnbinom(9, size=1/alpha, mu=mu) - pnbinom(5, size=1/alpha, mu=mu)
  P6 <- 1 - pnbinom(9, size=1/alpha, mu=mu)
  return(cbind(P0, P1, P2, P3, P4, P5, P6))
}
```

```
NB_1 <- createItem(name, par=par,est=est,P=P_nb_rapi_1,   lbound=c(-Inf,-Inf,
0.00001))

# NB2_pop2_highlighted

name <- 'NB'
par  <- c(b0=1, b1=0, alpha=1)
est  <- c(T,T,T)
P_nb_rapi_2_h <- function(par, Theta, ncat){
  b0    <- par[1]
  b1    <- par[2]
  alpha <- par[3]

Theta <- Theta[2]


  mu    <- exp(b0 + b1*Theta)
  P0 <- pnbinom(2, size=1/alpha, mu=mu)
  P1 <- pnbinom(4, size=1/alpha, mu=mu) - pnbinom(0, size=1/alpha, mu=mu)
  P2 <- pnbinom(9, size=1/alpha, mu=mu) - pnbinom(4, size=1/alpha, mu=mu)
  P3 <- 1 - pnbinom(9, size=1/alpha, mu=mu)
  return(cbind(P0, P1, P2, P3))
}

NB_2_h <- createItem(name, par=par,est=est,P=P_nb_rapi_2_h,   lbound=c(-Inf,-
Inf,0.00001))

# NB2_pop2_unhighlighted

name <- 'NB'
par  <- c(b0=1, b1=0, alpha=1)
est  <- c(T,T,T)
P_nb_rapi_2_u <- function(par, Theta, ncat){
  b0    <- par[1]
  b1    <- par[2]
  alpha <- par[3]

 Theta <- Theta[2]


  mu    <- exp(b0 + b1*Theta)
  P0 <- pnbinom(0, size=1/alpha, mu=mu)

  P1 <- pnbinom(1, size=1/alpha, mu=mu)
  P2 <- pnbinom(2, size=1/alpha, mu=mu) - pnbinom(0, size=1/alpha, mu=mu)
  P3 <- pnbinom(5, size=1/alpha, mu=mu) - pnbinom(2, size=1/alpha, mu=mu)

  P5 <- pnbinom(9, size=1/alpha, mu=mu) - pnbinom(5, size=1/alpha, mu=mu)
  P6 <- 1 - pnbinom(9, size=1/alpha, mu=mu)
  return(cbind(P0, P1, P2, P3, P4, P5, P6))
}

NB_2_u <- createItem(name, par=par,est=est,P=P_nb_rapi_2_u,   lbound=c(-Inf,-
Inf,0.00001))



# OZINB_pop1

name <- 'ZINB'
par  <- c(b0=1, b1=1, alpha=1, gam0=0)
est  <- c(T,T,T,T)
P_zinb_1 <- function(par, Theta, ncat){
  b0    <- par[1]
  b1    <- par[2]
  alpha <- par[3]
  gam0  <- par[4]

  Theta <- Theta[1]


  mu    <- exp(b0 + b1*Theta)
  pp    <- 1 / (1+exp(gam0)))

  P1 <- pp*(pnbinom(1, size=1/alpha, mu=mu))
  P2 <- pp*(pnbinom(2, size=1/alpha, mu=mu) -pnbinom(0, size=1/alpha, mu=mu))
  P3 <- pp*(pnbinom(5, size=1/alpha, mu=mu) -pnbinom(2, size=1/alpha, mu=mu))

  P5 <- pp*(pnbinom(9, size=1/alpha, mu=mu) -pnbinom(5, size=1/alpha, mu=mu))
  P6 <- pp*(1 - pnbinom(9, size=1/alpha, mu=mu))
  ret <- cbind(1-P1-P2-P3-P4-P5-P6, P1, P2, P3, P4, P5, P6))
  ret <- ifelse(ret > (1-1e-7), (1-1e-7), ret)
  ret <- ifelse(ret < 1e-7, 1e-7, ret)
  return(ret)
}


# OZINB_pop2_h

name <- 'ZINB'
par  <- c(b0=1, b1=1, alpha=1, gam0=0)
est  <- c(T,T,T,T)
P_zinb_2_h <- function(par, Theta, ncat){
  b0    <- par[1]
  b1    <- par[2]
  alpha <- par[3]
  gam0  <- par[4]

Theta <- Theta[2]
```

```r
  mu    <- exp(b0 + b1*Theta)
  pp    <- 1 / (1+exp(gam0))

  P1 <- pp * (pnbinom(4, size=1/alpha, mu=mu) - pnbinom(2, size=1/alpha, mu=m
u))


  P2 <- pp * (pnbinom(9, size=1/alpha, mu=mu) - pnbinom(4, size=1/alpha, mu=m
u))


  P3 <- pp * (1 - pnbinom(9, size=1/alpha, mu=mu))

  ret <- cbind(1-P1-P2-P3, P1, P2, P3)
  ret <- ifelse(ret > (1-1e-7), (1-1e-7), ret)
  ret <- ifelse(ret < 1e-7, 1e-7, ret)
  return(ret)
}

# OZINB_pop2_h

name <- 'ZINB'
par  <- c(b0=1, b1=1, alpha=1, gam0=0)
est  <- c(T,T,T,T)
P_zinb_2_u <- function(par, Theta, ncat){

  b0    <- par[1]
  b1    <- par[2]
  alpha <- par[3]
  gam0  <- par[4]
  Theta <- Theta[2]


  mu    <- exp(b0 + b1*Theta)
  pp    <- 1 / (1+exp(gam0))

  P1 <- pp*(pnbinom(1, size=1/alpha, mu=mu))
  P2 <- pp*(pnbinom(2, size=1/alpha, mu=mu) -pnbinom(0, size=1/alpha, mu=mu))
  P3 <- pp*(pnbinom(5, size=1/alpha, mu=mu) -pnbinom(2, size=1/alpha, mu=mu))

  P5 <- pp*(pnbinom(9, size=1/alpha, mu=mu) -pnbinom(5, size=1/alpha, mu=mu))
  P6 <- pp*(1 - pnbinom(9, size=1/alpha, mu=mu))
  ret <- cbind(1-P1-P2-P3-P4-P5-P6, P1, P2, P3, P4, P5, P6))
  ret <- ifelse(ret > (1-1e-7), (1-1e-7), ret)
  ret <- ifelse(ret < 1e-7, 1e-7, ret)
  return(ret)
}


ZINB_1 <- createItem(name, par=par, est=est, P=P_zinb_1, lbound=c(-Inf,-Inf,0
.00001,-Inf))
```

```
ZINB_2_h <- createItem(name, par=par, est=est, P=P_zinb_2_h, lbound=c(-Inf,-I
nf,0.00001,-Inf))

ZINB_2_u <- createItem(name, par=par, est=est, P=P_zinb_2_u, lbound=c(-Inf,-I
nf,0.00001,-Inf))


#This is the same in both models

IDA_Model <-mirt(data, mirt(dep[,4:10],

            2,

           c( rep('NB_1',7), rep('NB_2',7))  ,

            customItems=list(NB_1=NB_1, NB_2=NB_2),

      pars=”values”

            technical=list(

            removeEmptyRows=TRUE))

# Freeing the Mean and Variance of Eta for Study 2

IDA_Model[23,9] <-TRUE

IDA_Model[26,9] <-TRUE

mirt(data, mirt(dep[,4:10],

            model= IDA_Model,

           c( rep('NB_1',7), rep('NB_2',7))  ,

            customItems=list(NB_1=NB_1, NB_2=NB_2),

            technical=list(

            removeEmptyRows=TRUE))
```

APPENDIX D: SCORE CORRELATIONS FOR ALL FOUR MEASUREMENT CONDITIONS

Table 10

Multiple Group Factor Score Correlations for Participants Who Received Scenario 1 Twice

		Visit 1					Visit 2		
			GRM		OCFM		GRM		OCFM
			No DIF	*DIF*	*No DIF*	*DIF*	*No DIF*	*DIF*	*No DIF*
Visit 1	*GRM*	*DIF*	0.936						
	OCFM	*No DIF*	0.920	0.962					
		DIF	0.934	0.958	0.997				
Visit 2	*GRM*	*No DIF*	**0.651**	0.735	0.725	0.709			
		DIF	0.624	**0.695**	0.682	0.669	0.942		
	OCFM	*No DIF*	0.653	0.732	**0.744**	0.726	0.944	0.856	
		DIF	0.618	0.696	0.719	**0.700**	0.861	0.714	0.965

Test-retest correlations within model are bolded.

Table 11

Multiple Group Factor Score Correlations for Participants Who Received Scenario 1 then Scenario 4

		Visit 1					Visit 2		
			GRM		OCFM		GRM		OCFM
			No DIF	*DIF*	*No DIF*	*DIF*	*No DIF*	*DIF*	*No DIF*
Visit 1	*GRM*	*DIF*	0.930						
	OCFM	*No DIF*	0.920	0.974					
		DIF	0.933	0.968	0.996				
Visit 2	*GRM*	*No DIF*	**0.674**	0.758	0.771	0.763			
		DIF	0.625	**0.702**	0.714	0.706	0.953		
	OCFM	*No DIF*	0.703	0.795	**0.810**	0.801	0.956	0.878	
		DIF	0.680	0.771	0.786	**0.778**	0.898	0.770	0.970

Test-retest correlations within model are bolded.

Table 12

Multiple Group Factor Score Correlations for Participants Who Received Scenario 4 then Scenario 1

			Visit 1				Visit 2		
			GRM		OCFM		GRM		OCFM
			No DIF	*DIF*	*No DIF*	*DIF*	*No DIF*	*DIF*	*No DIF*
Visit 1	*GRM*	*DIF*	0.941						
	OCFM	*No DIF*	0.923	0.976					
		DIF	0.937	0.972	0.996				
Visit 2	*GRM*	*No DIF*	**0.715**	0.792	0.785	0.782			
		DIF	0.640	**0.716**	0.707	0.705	0.956		
	OCFM	*No DIF*	0.712	0.794	**0.816**	0.812	0.941	0.860	
		DIF	0.700	0.768	0.792	**0.790**	0.863	0.739	0.962

Test-retest correlations within model are bolded.

Table 13

Multiple Group Factor Score Correlations for Participants Who Received Scenario 4 Twice

			Visit 1				Visit 2		
			GRM		OCFM		GRM		OCFM
			No DIF	*DIF*	*No DIF*	*DIF*	*No DIF*	*DIF*	*No DIF*
Visit 1	*GRM*	*DIF*	0.932						
	OCFM	*No DIF*	0.875	0.945					
		DIF	0.896	0.945	0.995				
Visit 2	*GRM*	*No DIF*	**0.688**	0.759	0.739	0.740			
		DIF	0.641	**0.705**	0.692	0.695	0.946		
	OCFM	*No DIF*	0.695	0.783	**0.847**	0.839	0.912	0.838	
		DIF	0.640	0.726	0.819	**0.808**	0.805	0.682	0.962

Test-retest correlations within model are bolded.

REFERENCES

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Wiley.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective* (Vol. 467). John Wiley & Sons.

Bureau of Labor Statistics, U.S. Department of Labor. (2012). *National Longitudinal Survey of Youth 1979 cohort, 1979-2010 (rounds 1-24)*. Columbus, OH: The Center for Human Resource Research.

Burton, S., & Blair, E. (1991). Task Conditions, Response Formulation Processes, and Response Accuracy for Behavioral Frequency Questions in Surveys. *Public Opinion Quarterly*, *55*(1), 50–79. https://doi.org/10.1086/269241

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456.

Camarda, C. G., Eilers, P. H. C., & Gampe, J. (2008). Modelling general patterns of digit preference. *Statistical Modelling*, *8*(4), 385–401. https://doi.org/10.1177/1471082X0800800404

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1–29.

Cohn, A. M., Hagman, B. T., Graff, F. S., & Noel, N. E. (2011). Modeling the severity of drinking consequences in first-year college women: an item response theory analysis of the Rutgers Alcohol Problem Index. *Journal of Studies on Alcohol and Drugs*, *72*(6), 981–990.

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, *14*(2), 81–100. https://doi.org/10.1037/a0015914

Earleywine, M., LaBrie, J. W., & Pedersen, E. R. (2008). A brief Rutgers Alcohol Problem Index with less potential for bias. *Addictive Behaviors*, *33*(9), 1249–1253. https://doi.org/10.1016/j.addbeh.2008.05.006

Elizabeth T Miller, D. J. N. (2002). Test-retest reliability of alcohol measures: Is there a difference between internet-based assessment and traditional methods? Psychology of Addictive Behaviors, 16, 56-63. *Psychology of Addictive Behaviors : Journal of the Society of Psychologists in Addictive Behaviors*, *16*(1), 56–63. https://doi.org/10.1037/0893-164X.16.1.56

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Retrieved from http://doi.apa.org/psycinfo/2000-03918-000

Hilbe, J. M. (2011). *Negative Binomial Regression* (2nd ed.). Cambridge: Cambridge University Press.

Hilbe, J. M. (2013). Beta Binomial Regression. Retrieved from http://works.bepress.com/cgi/viewcontent.cgi?article=1073&context=joseph_hilbe

Hussong, A. M., Curran, P. J., & Bauer, D. J. (2013a). Integrative Data Analysis in Clinical Psychology Research. *Annual Review of Clinical Psychology*, *9*, 61–89. https://doi.org/10.1146/annurev-clinpsy-050212-185522

Hussong, A. M., Curran, P. J., & Bauer, D. J. (2013b). Integrative Data Analysis in Clinical Psychology Research. *Annual Review of Clinical Psychology*, *9*, 61–89. https://doi.org/10.1146/annurev-clinpsy-050212-185522

Johnston, L., O'Malley, P., & Bachman, J. (2012). *Monitoring the Future National Results on Adolescent Drug Use: Overview of Key Findings, 2011*. Institute for Social Research: The University of Michigan: Ann Arbor.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. Springer. Retrieved from http://link.springer.com/content/pdf/10.1007/978-1-4939-0317-7.pdf

Martens, M. P., Neighbors, C., Dams-O'Connor, K., Lee, C. M., & Larimer, M. E. (2007). The factor structure of a dichotomously scored Rutgers Alcohol Problem Index. *Journal of Studies on Alcohol and Drugs*, *68*(4), 597–606.

Martin, C. S., & Winters, K. C. (1998). Martin, C. S., & Winters, K. C. (1998). Diagnosis and assessment of alcohol use disorders among adolescents. , 22(2), 95. *Alcohol Research and Health*, *22*(2). Retrieved from http://pubs.niaaa.nih.gov/publications/arh22-2/95-106.pdf

McGinley, J. S., & Curran, P. J. (2014). Validity concerns with multiplying ordinal items defined by binned counts: An application to a quantity-frequency measure of alcohol use. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *10*(3), 108–116. https://doi.org/10.1027/1614-2241/a000081

McGinley, J. S., Curran, P. J., & Hedeker, D. (2015). A novel modeling framework for ordinal data defined by collapsed counts. *Statistics in Medicine*, *34*(15), 2312–2324. https://doi.org/10.1002/sim.6495

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge. Retrieved from https://books.google.com/books?hl=en&lr=&id=EXmsAgAAQBAJ&oi=fnd&pg=PR2&dq=millsap+factor+invariance&ots=XX4zbpO5py&sig=3RMXssKgB17SFP-_A8dsNr3g3ew

Neal, D. J., Corbin, W. R., & Fromme, K. (2006). Measurement of alcohol-related consequences among high school and college students: application of item response models to the Rutgers Alcohol Problem Index. *Psychological Assessment*, *18*(4), 402–414. https://doi.org/10.1037/1040-3590.18.4.402

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(2), 107–124. https://doi.org/10.1080/10705519809540095

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Sage. Retrieved from https://books.google.com/books?hl=en&lr=&id=wS8VEMtJ3UYC&oi=fnd&pg=PR5&dq=collapsing+thresholds+empty+categories+graded+response+model&ots=m8yJBsv4cD&sig=lAD09uAgiWQOvZAlKBZerUer97A

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

Radloff, L. S. (1977). The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement*, *1*(3), 385–401. https://doi.org/10.1177/014662167700100306

Recommended alcohol questions. (2003). Retrieved March 20, 2016, from http://www. niaaa.nih.gov/research/guidelines-and-resources/recommended-alcohol-questions

Roberts, J. M., & Brewer, D. D. (2001). Measures and tests of heaping in discrete quantitative distributions. *Journal of Applied Statistics*, *28*(7), 887–896. https://doi.org/10.1080/02664760120074960

Ross, S. (2009). *A First Course in Probability 8th Edition*. Pearson.

Rosseel, Y. (2011). lavaan: an R package for structural equation modeling and more Version 0.4-9 (BETA). *Retrieved from*. Retrieved from http://byrneslab.net/classes/lavaan_materials/lavaanIntroduction4-9.pdf

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods*, *11*(4), 402–415. https://doi.org/10.1037/1082-989X.11.4.402

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. Retrieved from http://psycnet.apa.org/psycinfo/2001-01226-002

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. Retrieved from http://psycnet.apa.org/psycinfo/1988-97024-010

White, H. R., & Labouvie, E. W. (1989). Towards the assessment of adolescent problem drinking. *Journal of Studies on Alcohol*, *50*(1), 30–37.

Winters, K. (1997). Assessment of Alcohol and Other Drug Use Behaviors Among Adolescents. In J. P. Allen & Megan Columbus (Eds.), *Assessing alcohol problems: A guide for clinicians and researchers*. DIANE Publishing.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*(1), 1–27.

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, *73*(3), 532–547.

Wright, D. E., & Bray, I. (2003). A mixture model for rounded data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *52*(1), 3–13. https://doi.org/10.1111/1467-9884.00338