Statistics
in Medicine

Research Article

Received 15 July 2013, Accepted 13 March 2015 Published online 8 April 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6495

A novel modeling framework for ordinal data defined by collapsed counts

James S. McGinley,[a*†] Patrick J. Curran[b] and Donald Hedeker[c]

Adolescent alcohol use is a serious public health concern. Despite advances in the theoretical conceptualization of pathways to alcohol use, researchers are limited by the statistical techniques currently available. Researchers often fit linear models and restrictive categorical models (e.g., proportional odds models) to ordinal data with many response categories defined by collapsed count data (0 drinking days, 1–2 days, 3–6 days, etc.). Consequently, existing models ignore the underlying count process, resulting in disjoint between the construct of interest and the models being fitted. Our proposed ordinal modeling approach overcomes this limitation by explicitly linking ordinal responses to a suitable underlying count distribution. In doing so, researchers can use maximum likelihood estimation to fit count models to ordinal data as if they had directly observed the underlying discrete counts. The usefulness of our ordinal negative binomial and ordinal zero-inflated negative binomial models is verified by simulation studies. We also demonstrate our approach using real empirical data from the 2010 National Survey of Drug Use and Health. Results show the benefit of the proposed ordinal modeling framework compared with existing methods. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: ordinal data; count data; grouped counts; collapsed counts; ordinal-count; zero inflation

1. Introduction

Substance use is one of the most commonly occurring health-risk behaviors during adolescence. According to Monitoring the Future, in 2012, 30% of eight graders, 54% of 10th graders, and 69% of 12th graders reported drinking alcohol, and 19% of eight graders, 37% of 10th graders, and 49% of 12th graders reported illicit drug use at least once in their life [1]. Adolescent drinking not only has a high monetary cost to society but also causes morbidity, driving accidents, risky sexual behavior, and even death [2, 3]. Research suggests that adolescent substance use has many significant long-term effects. For example, adolescent substance use has been linked to mental health problems such as increased internalizing symptomatology during adulthood [4]. Fortunately, tremendous advances have recently been made in the theoretical conceptualization and empirical evaluation of pathways to substance use during adolescence. Despite these recent gains, researchers are often limited by the statistical techniques that are currently available to test specific research hypotheses.

Adolescent substance use research is largely focused on outcomes that are, by definition, discrete counts. For example, researchers often focus on the quantity (number of drinks consumed) and frequency (number of drinking occasions) of substance use as well as the level of impairment (number of symptoms). There are several well-developed statistical methods for analyzing count data including the Poisson, negative binomial (NB), hurdle, and zero-inflated models. However, for reasons such as the need to reduce participant burden and limit error in cognitive recall, these outcomes are frequently measured using ordinal scales [5]. The ordinal scales applied in practice typically consist of 5–12 response categories that represent collapsed discrete counts. For example, two large studies assessing adolescent substance use, Monitoring the Future and Health and Behavior of School-Aged Children, inquire about

[a] *McGinley Statistical Consulting, LLC, North Huntingdon, PA, U.S.A.*
[b] *Department of Psychology, University of North Carolina at Chapel Hill, Chapel Hill, NC, U.S.A.*
[c] *Department of Public Health Sciences, University of Chicago, Chicago, IL, U.S.A.*
Correspondence to: James S. McGinley, McGinley Statistical Consulting, LLC, 610 Vincent Drive, North Huntingdon, PA 15642 U.S.A.
[†] *E-mail: jim@mcginleystatconsult.com*

the past 30-day alcohol use on ordinal scale (e.g., 1 = 0 occasion, 2 = 1–2 occasions, 3 = 3–5 occasions, and 4 = 6–9 occasions) [6, 7]. Similar collapsing procedures are recommended by the National Institute on Alcohol Abuse and Alcoholism and are thus widely used across many research applications [8].

Although assessing counts as binned ordinal responses appears advantageous from a measurement perspective (e.g., ease of burden and improved recall), it introduces significant complexities in statistical modeling. Researchers commonly fit linear models or proportional odds (PO) models to these ordinal data, but these models often explicitly violate statistical assumptions such as normally distributed errors and PO. Existing techniques also fail to account for the underlying count process, making substantive interpretation of the model parameters unclear. For example, the PO model examines the effect of covariates over the cumulative odds across response categories even though the counts (not the categories) are of actual substantive interest. Similarly, standard linear models are fitted to ordinal category scores and substantively interpreted as if they were the actual count construct of interest. However, these ordinal scores have little substantive meaning because they are arbitrarily assigned numbers used to represent ranges of collapsed counts. As a result, there is disjoint between the statistical models being used and the theoretical constructs of interest, which impacts researchers' ability to test theory in a reliable and valid manner [9].

Furthermore, a defining feature of adolescent substance use data is excess zero values beyond what would be expected by standard count distributions. These data may be conceptualized in the mixture framework. For instance, the large proportion of zeros may be theorized as arising from two populations, one group of non-drinkers who never drank alcohol and a second group of drinkers who did not drink over the assessed time frame [10]. Again, models for zero inflation are well developed for count data, but few methods have been proposed for ordinal data consisting of binned counts. The existing statistical methods for zero-inflated ordinal data such as the zero-inflated proportional odds (ZIPO) model do not assume known cut-points but instead link the ordinal response to an underlying logistic distribution through the threshold concept and ignore any possible underlying count distribution [11, 12]. Consequently, these models assume the incorrect underlying distribution and lack parsimony because they require the estimation of several intercepts.

For more than four decades, statisticians and economists alike have proposed methods for handling grouped Poisson count data [13–16]. However, these techniques have not been analytically expressed or empirically evaluated from an underlying latent response variable framework [17, 18]. As a result, these models have not been integrated into a broader analytical framework, and this, in turn, has limited dissemination and use in practice. Our goal here is to propose a general ordinal modeling framework that explicitly links the ordinal responses to a suitable count distribution through the known cut-points; these cut-points are known because they are the values that define the range of counts within each ordinal response. This overcomes the limitations associated with existing techniques (e.g., violation of model assumptions such as PO and normally distributed errors), and importantly, the proposed framework is consistent with the true underlying count construct of substantive interest. We illustrate the framework by describing and demonstrating ordinal negative binomial (ONB) and ordinal zero-inflated negative binomial (OZINB) models for ordinal data with underlying counts. Furthermore, this framework logically extends to other count distributions such as Poisson and alternative models that accommodate truncation, heterogeneous dispersion, and zero inflation. Thus, the ordinal-count framework offers quantitative and substantive advantages over existing methods so that researchers can test research hypotheses in ways previously not possible.

We begin by introducing the ordinal-count framework with an emphasis on linking ordinal responses to underlying NB and zero-inflated negative binomial (ZINB) distributions. We focus on these distributions because they are consistent with adolescent substance use data compared with the more restricted count distributions (e.g., Poisson and zero-inflated Poisson). Next, we describe a simulation study that assesses the performance of our proposed ordinal-count models compared with NB and ZINB count models fitted to the complete underlying count data. We then present an empirical example of our ordinal-count modeling framework using the data from the 2010 National Survey on Drug Use and Health [19]. Finally, we discuss the results and the unique contributions of the proposed ordinal-count modeling framework for substantive research.

2. Ordinal-count models

2.1. *Ordinal negative binomial model*

In substance use research, the goal is often to predict some underlying count outcome (e.g., quantity and frequency of use or level of symptomatology) as a function of a set of covariates. However, researchers often measure these count outcomes with ordinal measures that collapse discrete counts into a series of response categories with known cut-points, which are defined by the ordinal response scale [20]. Under these circumstances, we can assume that underlying the ordinal response is a discrete count distribution, in this case, the NB distribution. The NB distribution is justified for adolescent substance use data because the variance of the underlying count outcome is usually much larger than the mean. Next, we describe the process of linking the ordinal responses to the underlying count distribution.

In order to model the ordinal response as a function of an underlying NB distribution, assume that underlying the ordinal outcome, Y_i for person $i (i = 1, \dots, n)$, is an unobserved count latent variable, Y_i^*. This unobserved count latent variable may follow any count distribution, but in this paper, we focus on the NB distribution. The probability mass function (PMF) for the NB distribution in terms of Y_i^* conditional on our covariates, $\boldsymbol{x}_i$, is

$$f\left(y_i^*\right) = P\left(Y_i^* = y_i^* | \boldsymbol{x}_i\right) = \frac{\Gamma(y_i^* + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i^* + 1)} \left(\alpha \mu_I\right)^{y_i^*} \left(1 + \alpha \mu_i\right)^{-\left(y_i^* + \alpha^{-1}\right)}, \quad y_i^* = 0, 1, 2, \ldots \tag{1}$$

where $E\left(Y_i^*\right) = \mu_i$, $Var\left(Y_i^*\right) = \mu_i + \alpha \mu_i^2$, and α is the dispersion parameter. We can use the log function to link our linear predictor to the mean of Y_i^*:

$$\log\left(\mu_i\right) = \boldsymbol{x}_i' \boldsymbol{\beta}, \tag{2}$$

where $\boldsymbol{x}_i$ is a $p \times 1$ vector of covariates (typically including '1' as the first element for the intercept) and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients.

The cumulative distribution function (CDF) for the NB distribution is simply the sum of the PMFs such that

$$F\left(y_i^*\right) = \sum_{v=0}^{y_i^*} f(v), \tag{3}$$

where the cumulative probability is evaluated at y_i^*.

The next step is linking the ordinal outcome, Y_i, to the unobserved latent variable, Y_i^*. This is accomplished by using the fixed and known cut-points defined by the ordinal measure such that

$$Y_i = c \text{ if } \kappa_{c-1} < Y_i^* \leqslant \kappa_c, \tag{4}$$

where κ_c is the count number that defines the upper bound of ordinal response category $c (c = 1, 2, \dots, M)$. We can express the probability of observing a response in category c as the function of the cumulative probabilities from the underlying Y_i^* distribution.

$$P(Y_i = c | \boldsymbol{x}_i) = P\left(\left(\kappa_{c-1} < Y_i^* \leqslant \kappa_c\right) | \boldsymbol{x}_i\right) = F(\kappa_c) - F(\kappa_{c-1}). \tag{5}$$

Here, $F(\kappa_c)$ and $F(\kappa_{c-1})$ designate the CDFs evaluated at the known upper count numbers for categories c and c-1 for a distribution with a mean of μ_i and dispersion of α. Finally, letting $(y_{i1}, \dots, y_{ic})$ represent the binary reference codes indicative of the response for subject i (e.g., $y_{ic} = 1$ if $Y_i = c$ and 0 otherwise), we can express the likelihood function for the ordinal data following an underlying count distribution as

$$L_{ONB} = \prod_{i=1}^{n} \left[\prod_{c=1}^{M} \left[F\left(\kappa_c\right) - F\left(\kappa_{c-1}\right) \right]^{y_{ic}} \right]. \tag{6}$$

With slight modification, we can also adjust for truncation or use another count distribution such as the Poisson distribution. These are direct extensions, so we do not detail them further here (see Hilbe [21], for examples of potential counts models).

Our proposed approach differs substantially from the current best practice of the PO model because we explicitly link the ordinal responses to the underlying NB CDF as opposed to the logistic CDF. Additionally, our proposed ONB model examines the effect of covariates as if we were directly modeling the latent count responses, whereas the PO model examines the effect of covariates over the cumulative odds across response categories.

2.2. *Ordinal zero-inflated negative binomial model*

In many research settings where count data are collected, such as adolescent substance use, we observe excess zeros relative to the NB distribution. Researchers frequently address zero inflation using the ZINB model. The ZINB model assumes a mixture distribution that generates one binary outcome (modeled through logistic regression) and one count outcome (modeled through count regression). Consequently, the observed zeros arise from two unique sources. First, zeros arise because individuals are not at risk for a given outcome. Second, zeros arise because individuals who are at risk do not experience the outcome over the assessment period. For example, adolescents may report no drinking over the past 30 days because (1) they are non-drinkers or (2) they are drinkers, but they did not drink in the past 30 days. These zero-inflated models may be of both quantitative and substantive interests to researchers with ordinal data with underlying counts. Assuming the ordinal data have a zero response category, we can readily incorporate zero-inflated models into our ordinal modeling framework.

Lambert [22] formally proposed the zero-inflated Poisson model for count data, which logically extends to the ZINB [23]. Like the ONB, we can map these well-developed count models onto the ordinal data through Y_i^*. For example, consistent with Lambert's [22] count modeling framework,

$$\begin{array}{ll} Y_i^* \sim 0 & \text{with probability } \pi_i, \\ Y_i^* \sim f\left(y_i^*\right) & \text{with probability } \left(1-\pi_i\right), \end{array} \tag{7}$$

where $f\left(y_i^*\right)$, again, denotes the PMF for the NB distribution. The PMF can be expressed as

$$\begin{array}{l} P\left(Y_i^*=0\right)=\pi_i+\left(1-\pi_i\right)f(0) \\ P\left(Y_i^*=y_i^*\right)=\left(1-\pi_i\right)f\left(y_i^*\right), \quad y_i^*=1,2,..., \end{array} \tag{8}$$

where $f(0)$ is simply the probability that $Y_i^*=0$ based on the NB distribution. We can model the binary zero process through a logistic model,

$$\text{logit}\left(\pi_i\right)=\boldsymbol{w}_i'\boldsymbol{\gamma}, \tag{9}$$

where $\boldsymbol{w}_i$ is a $q\times1$ vector of covariates (typically including '1' as the first element for the intercept) and $\boldsymbol{\gamma}$ is a $q\times1$ vector of regression coefficients. The count process is modeled as in Equation (2). The covariates for the count and zero processes do not need to be the same, as shown through the unique vectors $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$. The procedure for linking the ordinal outcome, Y_i, to the unobserved latent variable, Y_i^*, is precisely the same as in Equations (4) and (5). Letting $(y_{i1},\dots,y_{ic})$, again, represent the binary reference codes indicative of the response for subject i ($y_{ic}=1$ if $Y_i=c$ and 0 otherwise), we can express the likelihood function for zero-inflated ordinal models as

$$L_{OZINB}=\prod_{i=1}^{n}\left[\prod_{c=1}^{M}\left[\left(1-\pi_i\right)\left(F\left(\kappa_c\right)-F(\kappa_{c-1})\right)+I(Y_i=1)\pi_i\right]^{y_{ic}}\right], \tag{10}$$

where $I(Y_i=1)$ is an indicator function that equals one when $Y_i=1$ (e.g., when the first response category representing zero is selected) and zero otherwise. The OZINB is similar to fitting the zero-inflated model as if we had the true underlying counts. Alternative approaches to modeling zero-inflated ordinal data such as the ZIPO model do not account for this underlying count process.

We have defined a novel modeling framework for ordinal data consisting of underlying counts with specific emphasis on NB and ZINB models. Now, we turn to estimation.

2.3. *Estimation*

The proposed ordinal-count models can be fitted using standard maximum likelihood estimation through optimizing the respective log-likelihood functions corresponding to Equations (6) and (10). We provide code in our online appendix to fit these models using PROC NLMIXED both in SAS and in R [24, 25].

3. Simulation study

We conducted a simulation study using SAS 9.3 to evaluate the proposed ONB and OZINB models under known population conditions. We generated count data from an NB distribution and a ZINB distribution. We created ordinal data by collapsing the counts into ordinal responses corresponding to 5 pt (0 = '0', 1 = '1–4', 2 = '5–10', 3 = '11–20', 4 = '21+'); 7 pt (0 = '0', 1 = '1–2', 2 = '3–5', 3 = '6–9', 4 = '10–19', 5 = '20–39', 6 = '40+'); and 10 pt (0 = '0', 1 = '1', 2 = '2–3', 3 = '4–5', 4 = '6–9', 5 = '10–15', 6 = '16–22', 7 = '23–29', 8 = '30–39', 9 = '40+') scales consistent with those observed in practice [5–8]. We fitted NB and ZINB models to the open-ended count data, while the proposed ONB and OZINB models were fitted to ordinal data defined by collapsed counts. We examined sample sizes of 250, 500, and 1000. For both the NB and ZINB distributions, we completed 1000 replications at each sample size. We included two standard normal continuous covariates x_{1i} and x_{2i} with a correlation of 0.3. We used raw bias, standardized bias, root mean square error, and 95% CI coverage probabilities to evaluate model performance over the replications.[‡] We also compared the proposed ordinal-count models with existing models including the PO, linear, and ZIPO models by examining empirical power and relative efficiency.

We defined empirical power as the proportion of statistically significant effects using an alpha level of 0.05. Relative efficiency was computed as the ratio of the efficiencies for predictions from the existing models compared with the proposed ordinal-count models (e.g., $\frac{EFF_{PO}}{EFF_{ONB}}, \frac{EFF_{Lin}}{EFF_{ONB}}$). Hence, if the relative efficiency was greater than 1, the proposed models were more efficient than the existing models. Efficiency was computed as $\sum_i \left(\hat{Y}_i - \mu_i\right)^2$ for each of the converged replications. Here, μ_i denotes the $E\left[Y_i|\boldsymbol{x}_i\right]$ based on the true population-generating model. In the linear model, $\hat{Y}_i$ was the predicted value of the ordinal-count outcome scored as category numbers (0, 1, 2, etc.). For the ordinal models, $\hat{Y}_i$ was $\sum_c Y_{ic}\hat{p}_{ic}$, where $\hat{p}_{ic}$ was the predicted probability of person i being in category c. To retain focus, we present a subset of the simulation results. The complete results are available in our online supplemental appendix at http://www.unc.edu/~curran/manuscripts.htm

Model convergence rates for the NB, ONB, and linear models were high (e.g., 99.6% or higher across all models and conditions). The convergence rates for the PO model were high with five response categories across all sample sizes (99.6% or higher). The PO model convergence rates were also respectable with 7 and 10 response categories when sample size was greater than or equal to 500 (95.5% or higher). However, convergence rates were lower for the PO model when $N = 250$, and there were 7 or 10 response categories (7 pt: 83.8%; 10 pt: 72.8%).

Table I displays the parameter recovery results for the NB simulation with the open-ended counts and 7-pt response scale. The simulation results showed that the ONB and NB models both recovered the assigned parameter values adequately across all three sample sizes as indicated by the small biases and root mean square errors and close 95% coverage rates. Importantly, aside from slightly larger standard errors for the ONB model, the performance of the proposed ONB was virtually the same as the correct population count NB model. These findings also generalized to the 5-pt and 10-pt response scales such that the greater the number of response categories, the smaller the standard errors. This suggested that our proposed ONB model fitted to the collapsed counts was functionally identical to the standard NB model fitted directly to the counts themselves.

Table II displays the empirical power results from the NB simulation. The proposed ONB model had very similar levels of empirical power compared with the NB model fitted to the open-ended counts. Empirical power for ONB model also increased with more response categories. The ONB had systematically higher empirical power rates than both the standard linear model and the PO model. Results also showed that the ONB was more efficient than the PO and linear models across all conditions. For example, in the 10-pt response scale, relative efficiencies for the PO model versus the ONB model ranged from 1.12 to 1.18, and the relative efficiencies for the linear model versus the ONB model ranged from 1.11 to 1.15 across the sample size conditions. Taken together, the ONB model displayed clear benefits over existing models for ordinal-count data.

The model convergence rates were acceptable for the ZINB and OZINB models and improved with increased response categories and subjects (ZINB: 93.2–99.1% and OZINB 5 pt: 90.7–98.9%, 7 pt: 92.3–99.1%, and 10 pt: 92.6–99.4%). In contrast, the ZIPO model often did not converge to a proper solution (5 pt: 48.8–57.5%, 7 pt: 21.6–49.8%, and 10 pt: 10.8–51.5%). Because of this instability in estimation,

[‡] For reference, standard bias of +/−40% can negatively impact efficiency, coverage, and error rates [26].

Table I. Parameter recovery for the simulation with the negative binomial model fitted to open-ended counts and the ordinal negative binomial model fitted to the 7-pt ordinal data.

		NB – counts						ONB – 7-pt scale					
	True	Est	SE	Bias	SB	RMSE	Cov	Est	SE	Bias	SB	RMSE	Cov
							$N = 250$						
β_0 (intercept)	1.20	1.18	0.13	−0.02	−17.81	0.13	0.95	1.18	0.14	−0.02	−16.13	0.14	0.95
$\beta_1(x_{1i})$	−0.20	−0.20	0.15	0.00	−2.23	0.15	0.94	−0.20	0.15	0.00	−2.51	0.15	0.95
$\beta_2(x_{2i})$	0.15	0.15	0.15	0.00	−0.35	0.15	0.93	0.15	0.15	0.00	−0.14	0.15	0.94
α (dispersion)	4.00	3.96	0.48	−0.04	−8.91	0.48	0.94	3.96	0.49	−0.04	−7.62	0.49	0.94
							$N = 500$						
β_0 (intercept)	1.20	1.19	0.09	−0.01	−10.78	0.09	0.96	1.19	0.10	−0.01	−8.42	0.10	0.95
$\beta_1(x_{1i})$	−0.20	−0.20	0.10	0.00	−1.30	0.10	0.95	−0.20	0.10	0.00	−3.56	0.10	0.94
$\beta_2(x_{2i})$	0.15	0.15	0.10	0.00	−1.27	0.10	0.94	0.15	0.11	0.00	−0.18	0.11	0.95
α (dispersion)	4.00	3.97	0.33	−0.03	−9.54	0.33	0.95	3.98	0.34	−0.02	−7.22	0.34	0.95
							$N = 1000$						
β_0 (intercept)	1.20	1.20	0.07	0.00	−6.32	0.07	0.94	1.20	0.07	0.00	−4.96	0.07	0.94
$\beta_1(x_{1i})$	−0.20	−0.20	0.07	0.00	−2.32	0.07	0.94	−0.20	0.07	0.00	−3.65	0.07	0.94
$\beta_2(x_{2i})$	0.15	0.15	0.07	0.00	3.84	0.07	0.94	0.15	0.07	0.00	5.00	0.07	0.94
α (dispersion)	4.00	3.97	0.23	−0.03	−11.88	0.24	0.95	3.97	0.24	−0.03	−10.99	0.24	0.95

NB, negative binomial; ONB, ordinal negative binomial; Est, average estimate; SE, empirical standard error; Bias, raw bias; SB, standardized bias; RMSE, root mean square error; Cov, coverage for 95% CI.

Table II. Empirical power from the simulation with the negative binomial model fitted to open-ended counts and the ordinal negative binomial, proportional odds, and linear models fitted to ordinal data.

	NB	ONB			PO			Linear		
	Count	5 pt	7 pt	10 pt	5 pt	7 pt	10 pt	5 pt	7 pt	10 pt
					$N = 250$					
β_0 (intercept)	1.00	1.00	1.00	1.00	—	—	—	1.00	1.00	1.00
$\beta_1(x_{1i})$	0.30	0.28	0.28	0.29	0.19	0.21	0.21	0.25	0.26	0.26
$\beta_2(x_{2i})$	0.19	0.17	0.18	0.19	0.13	0.13	0.13	0.17	0.16	0.17
					$N = 500$					
β_0 (intercept)	1.00	1.00	1.00	1.00	—	—	—	1.00	1.00	1.00
$\beta_1(x_{1i})$	0.52	0.49	0.51	0.52	0.30	0.30	0.31	0.42	0.41	0.46
$\beta_2(x_{2i})$	0.34	0.30	0.33	0.33	0.20	0.20	0.20	0.28	0.27	0.29
					$N = 1000$					
β_0 (intercept)	1.00	1.00	1.00	1.00	—	—	—	1.00	1.00	1.00
$\beta_1(x_{1i})$	0.82	0.78	0.81	0.81	0.52	0.52	0.53	0.71	0.71	0.73
$\beta_2(x_{2i})$	0.58	0.55	0.56	0.57	0.36	0.35	0.36	0.49	0.49	0.50

NB, negative binomial; ONB, ordinal negative binomial; PO, proportional odds.

Values represent the proportion of significant effects using an alpha of 0.05.

we do not present the results from the ZIPO model. These results can be found in the supplemental online appendix.

Table III displays the parameter recovery results for the ZINB simulations for open-ended counts and the 7-pt response scale. Results showed that the proposed OZINB and population-generating ZINB models performed similarly across the three sample sizes with the exception of slightly larger standard errors for the OZINB model compared with the ZINB model. As expected, both models performed better as the sample size increased. At the smallest sample size of $N = 250$, the parameters from the logistic portion of the model $(\gamma_0, \gamma_1, \gamma_2)$ showed elevated bias and larger standard errors, and the 95% coverage on the logistic intercept (γ_0), the count intercept (β_0), and the dispersion parameter (α) was low. At the largest sample size $N = 1000$, the standardized bias for the parameters from the logistic portion of the model was slightly high, and the 95% coverage on the logistic intercept (γ_0), the count intercept (β_0), and the dispersion parameter (α) was still lower than expected. Although not shown in this table, as the number of response categories increased (e.g., 5-pt response scale versus 10-pt response scale), OZINB model performance improved, and results paralleled those from the ZINB fitted to the open-ended counts.

The OZINB model was characterized by empirical power rates that were generally smaller than the ZINB model fitted to the open-ended counts (Simulation Appendix Table V). However, these differences were negligible when the ordinal data had a large number of response categories. In sum, the simulation results suggested that the OZINB model can accurately recover an underlying zero-inflated count process even when the data are collected as ordered categories.

4. Adolescent substance use example

We next demonstrate the utility of the proposed models using real empirical data. We used data assessing the frequency of alcohol use over the past 30 days from the 2010 National Survey on Drug Use and Health [19]. The alcohol frequency data were open-ended counts ranging between 0 and 30 days, but we created ordinal responses by collapsing the counts into six response categories consistent with those collected in practice (0 = '0 day', 1 = '1–2 days', 2 = '3–5 days', 3 = '6–9 days', 4 = '10–19 days', and 5 = '20–30 days'). This analytic strategy is advantageous over simply demonstrating the ordinal-count models on existing ordinal data because we could directly compare the parameter estimates produced from the ONB and OZINB models fitted to the ordinal data with the NB and ZINB models fitted to the open-ended counts. We are thus able to clearly assess the extent to which our method recovers the underlying count process. For comparison, we also fitted the linear, PO, and ZIPO models to the ordinal

Table III. Parameter recovery for the simulation with the zero-inflated negative binomial model fitted to open-ended counts and the ordinal zero-inflated negative binomial model fitted to the 7-pt ordinal data.

		ZINB – counts						OZINB – 7-pt scale					
	True	Est	SE	Bias	SB	RMSE	Cov	Est	SE	Bias	SB	RMSE	Cov
							$N = 250$						
β_0 (intercept)	1.20	1.20	0.33	0.00	0.42	0.33	0.85	1.20	0.33	0.00	0.79	0.33	0.85
$\beta_1(x_{1i})$	−0.30	−0.31	0.22	−0.01	−3.92	0.22	0.93	−0.31	0.23	−0.01	−4.35	0.23	0.94
$\beta_2(x_{2i})$	0.20	0.20	0.21	0.00	1.03	0.21	0.93	0.20	0.23	0.00	1.45	0.23	0.93
α (dispersion)	2.70	2.58	1.52	−0.12	−8.15	1.52	0.74	2.59	1.59	−0.11	−6.81	1.60	0.74
γ_0 (Inflated intercept)	0.10	−0.02	0.95	−0.12	−12.84	0.96	0.84	−0.03	0.98	−0.13	−13.37	0.99	0.85
γ_1(Inflated x_{1i})	0.30	0.35	0.49	0.05	10.58	0.49	0.96	0.35	0.50	0.05	10.29	0.51	0.97
γ_2(Inflated x_{2i})	−0.15	−0.19	0.48	−0.04	−8.13	0.48	0.98	−0.20	0.49	−0.05	−10.12	0.49	0.98
							$N = 500$						
β_0 (intercept)	1.20	1.18	0.27	−0.02	−5.69	0.27	0.87	1.19	0.28	−0.01	−5.02	0.28	0.85
$\beta_1(x_{1i})$	−0.30	−0.30	0.14	0.00	−0.90	0.14	0.94	−0.30	0.15	0.00	−2.09	0.15	0.95
$\beta_2(x_{2i})$	0.20	0.20	0.14	0.00	1.73	0.14	0.95	0.20	0.15	0.00	2.11	0.15	0.95
α (dispersion)	2.70	2.78	1.36	0.08	5.63	1.36	0.81	2.80	1.48	0.10	6.80	1.49	0.80
γ_0 (Inflated intercept)	0.10	−0.04	0.88	−0.14	−16.12	0.89	0.87	−0.07	0.95	−0.17	−17.72	0.96	0.86
γ_1(Inflated x_{1i})	0.30	0.35	0.40	0.05	13.70	0.40	0.97	0.36	0.42	0.06	14.48	0.42	0.97
γ_2(Inflated x_{2i})	−0.15	−0.16	0.35	−0.01	−3.08	0.35	0.97	−0.16	0.36	−0.01	−2.04	0.36	0.97
							$N = 1000$						
β_0 (intercept)	1.20	1.19	0.19	−0.01	−5.80	0.19	0.93	1.18	0.21	−0.02	−9.67	0.21	0.91
$\beta_1(x_{1i})$	−0.30	−0.29	0.10	0.01	5.46	0.10	0.94	−0.30	0.11	0.00	2.92	0.11	0.94
$\beta_2(x_{2i})$	0.20	0.20	0.10	0.00	2.47	0.10	0.95	0.20	0.10	0.00	3.56	0.10	0.96
α (dispersion)	2.70	2.77	0.97	0.07	7.07	0.97	0.87	2.83	1.13	0.13	11.59	1.14	0.86
γ_0 (Inflated intercept)	0.10	0.04	0.50	−0.06	−11.58	0.50	0.91	0.00	0.62	−0.10	−15.93	0.63	0.90
γ_1(Inflated x_{1i})	0.30	0.33	0.20	0.03	14.99	0.20	0.97	0.34	0.24	0.04	18.55	0.24	0.96
γ_2(Inflated x_{2i})	−0.15	−0.17	0.19	−0.02	−10.59	0.19	0.96	−0.17	0.22	−0.02	−10.59	0.22	0.97

ZINB, zero-inflated negative binomial; OZINB, ordinal zero-inflated negative binomial; Est, average estimate; SE, empirical standard error; Bias, raw bias; SB, standardized bias; RMSE, root mean square error; Cov, coverage for 95% CI.

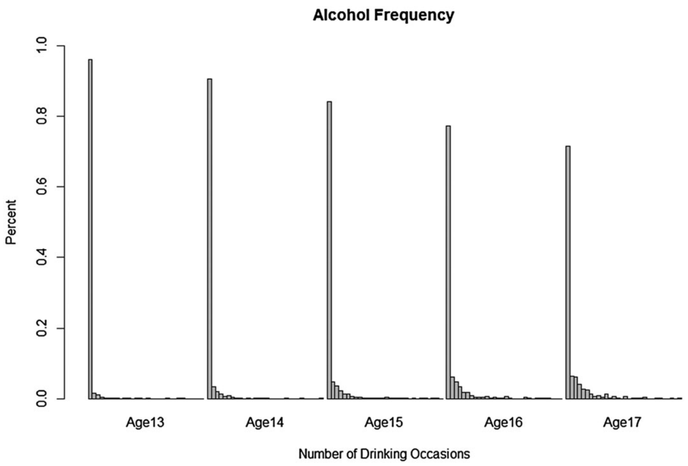

Figure 1. Open-ended alcohol frequency counts stratified by age.

data. However, parameter estimates from these models cannot be directly compared with the ordinal-count models because they function on fundamentally different metrics.

The National Survey on Drug Use and Health is a nationwide survey that aims to monitor trends, consequences, levels, and patterns of substance use and abuse. The subsamples used for our analyses consisted of 14,051 adolescents ranging in age from 13 to 17 years with a mean (sd) age of 15.07(1.41) and were 51% males and 38% minority (minority being defined as non-White). The outcome of interest was the frequency of alcohol use over the past 30 days. Figure 1 shows the distribution of open-ended count stratified by age, which suggests that zero inflation is a valid concern. Additional covariates included in our analyses were a binary lifetime major depressive episode (MDE) indicator (coded 0 = no lifetime MDE, 1 = lifetime MDE) and a composite peer substance use variable consisting of the mean of four items asking about how many fellow students use substances (0 = 'None of them' to 3 = 'All of them'). Age was centered at 13 years old for all analyses (e.g., age13 = age-13).

Our analytic goal was to test the unique effect of depression on the frequency of alcohol use above and beyond the influence peer use. First, we used single process models (e.g., NB, ONB, PO, and linear models) to test whether lifetime MDE predicts increased drinking frequency. Second, we used dual process models (e.g., ZINB, OZINB, and ZIPO models) to test the effect of depression on the probability of being a non-drinker and, for adolescent drinkers, the effect of depression on the frequency of alcohol use. We used the existing count NB and ZINB models fitted to the open-ended counts as the 'gold standard' and compared the results with those obtained from the proposed ordinal-count models and other commonly used models.

Table IV shows the results for the NB model fitted to the alcohol frequency counts and the ONB, PO, and linear models fitted to the six category ordinal data. The ONB, PO, and linear models led to similar substantive findings compared with the results from the NB model. In all models, the results indicated that increased age predicted more frequent drinking, that men drank more frequently than women, and that minorities drank less frequently than whites ($p < 0.001$ for all effects). As expected, having peers that use substances predicted significantly greater frequency of alcohol use. Above and beyond the influence of demographic characteristics and peer use, lifetime MDE predicted increased the frequency of alcohol use.

It is important to recognize that the parameter estimates from the proposed ONB were highly similar to those obtained fitting the NB model to open-ended counts, which is precisely the objective of the ONB model. In fact, all parameter estimates from the ONB were within the 95% CIs corresponding to the parameter estimates from the estimates produced by the count NB model. This further suggests that our proposed ONB is potentially highly useful for understanding the underlying count process when only binned categories are observed.

As for the PO model, the score test did indicate a possible violation of the PO assumption; $\chi^2(20) = 38.58$, $p = 0.008$. We recognize that the score statistic can be overly influenced by the large sample size, but it suggested that less parsimonious non-PO (or partial proportional) models may be needed. Either of these ordinal models would make substantive interpretations much more difficult because the effect of one or more covariates would vary across the cumulative logits. It was also interesting that the standard

Table IV. Results for the negative binomial model fitted to the alcohol frequency counts and the ordinal negative binomial, proportional odds, and linear models fitted to the ordinal data drawn from the National Survey on Drug Use and Health.

	Count	Ordinal		
	NB	Proposed ONB	PO	Linear
	Est(SE)	Est(SE)	Est(SE)	Est(SE)
β_0 : (intercept)	−3.20(0.09)	−3.36(0.09)	—	−0.16(0.02)
β_1 : (age13)	0.37(0.02)	0.39(0.02)	0.33(0.02)	0.07(0.01)
β_2 : (male)	0.18(0.06)	0.20(0.06)	0.21(0.05)	0.05(0.01)
β_3 : (minority)	−0.27(0.06)	−0.25(0.06)	−0.31(0.05)	−0.07(0.01)
β_4 : (peer use)	1.29(0.05)	1.36(0.06)	1.20(0.05)	0.27(0.01)
β_5 : (MDE)	0.40(0.08)	0.43(0.08)	0.39(0.06)	0.11(0.02)
α : (dispersion)	8.53(0.25)	8.60(0.21)	—	—
σ^2	—	—	—	0.59(0.01)
τ_1 : intercept 1	—	—	−4.04(0.09)	—
τ_2 : intercept 2	—	—	−4.89(0.09)	—
τ_3 : intercept 3	—	—	−5.93(0.10)	—
τ_4 : intercept 4	—	—	−6.60(0.11)	—
τ_5 : intercept 5	—	—	−8.09(0.16)	—

NB, negative binomial; ONB, ordinal negative binomial; PO, proportional odds; age13 is age centered at 13 years old (e.g., age-13). The parameter estimates from the PO and linear models are not directly comparable with those from the NB and ONB models because they operate in different metrics. MDE, major depressive episode.

linear model fitted to the ordinal data produced predictions that fell outside of the range of the observed data. More specifically, the intercept, which represents the predicted value for non-depressed 13-year-old White girls that reported no peer substance use, was negative (e.g., −0.16). This prediction was not valid because the ordinal data response categories were scored as integers from 0 to 5.

Although the significance tests (e.g., $p < .05$) are similar between the ONB, PO, and linear models, it is important to highlight the interpretational differences. Consider the point estimates for the depression effect, which is 0.43 for the ONB model, 0.39 for the PO model, and 0.11 for the linear model. In the proposed ONB model, the results can be interpreted in several ways. For instance, an interpretation consistent with standard linear models is, controlling for all other covariates, the log of the expected number of drinking days is 0.43 units larger for individuals with an MDE compared with individuals without an MDE. Another useful interpretation for the ONB model involves incidence rate ratios. For example, controlling for the other covariates, adolescents having an MDE increase their expected number of drinking occasions in the past 30 days by over 50% (e.g., $e^{.43} = 1.54$).

Conversely, the interpretation of the PO model is in terms of cumulative logits, or log odds. Thus, the interpretation is that, controlling for the other covariates, the expected log odds of falling into a higher response category for individuals with an MDE are 0.39 units greater than those for individuals without an MDE. The standard linear model's interpretation is more opaque because the dependent alcohol frequency variable represents the category numbers from the ordinal response variable ranging from 0 to 5. The resulting interpretation is that the expected value of alcohol frequency is 0.11 units greater for individuals with an MDE compared with individuals without an MDE. Clearly, the ONB, PO, and standard linear models differ substantially in how they treat the substantive construct of interest for hypothesis testing.

Table V shows the results for the ZINB model fitted to alcohol frequency counts and the proposed OZINB and ZIPO models fitted to the ordinal data. We found that our proposed OZINB model produced similar substantive results to those obtained fitting a ZINB model to the open-ended counts, but the ZIPO model results were substantially different. For the ZINB and the OZINB, age, peer substance use, and depression status significantly predicted the decreases in the log odds of being a non-drinker, whereas the ZIPO produced no significant covariate effects. Furthermore, the parameter estimates from the logistic portion of the model were similar between OZINB and ZINB models, whereas the estimates from the ZIPO were markedly discrepant. The substantive findings from the second portion of the model were similar across all three models. Results indicated that older individuals, men, Whites, and individuals with greater reported peer substance use had significantly greater frequency of alcohol use among drinkers.

Table V. Results for the zero-inflated negative binomial model fitted to alcohol frequency counts and the ordinal zero-inflated negative binomial and zero-inflated proportional odds models fitted to the ordinal data drawn from the National Survey on Drug Use and Health.

		Ordinal	
	ZINB Est(SE)	OZINB Est(SE)	ZIPO Est(SE)
β_0 : (intercept)	−0.22(0.15)	−0.56(0.16)	—
β_1 : (age13)	0.08(0.03)	0.09(0.03)	0.29(0.08)
β_2 : (male)	0.20(0.06)	0.23(0.06)	0.31(0.07)
β_3 : (minority)	−0.24(0.06)	−0.25(0.06)	−0.36(0.07)
β_4 : (peer use)	0.44(0.06)	0.49(0.07)	1.52(0.23)
β_5 : (MDE)	0.09(0.08)	0.12(0.08)	0.27(0.17)
α : (dispersion)	2.65(0.33)	3.78(0.36)	—
γ_0 : (*Inf* intercept)	3.40(0.13)	3.37(0.15)	0.72(1.16)
γ_1 : (*Inf* age13)	−0.40(0.03)	−0.44(0.04)	−0.17(0.10)
γ_2 : (*Inf* peer use)	−1.39(0.11)	−1.60(0.13)	−0.13(0.41)
γ_3 : (*Inf* MDE)	−0.53(0.12)	−0.65(0.15)	−0.32(0.19)
τ_1 : intercept 1	—	—	−3.29(0.78)
τ_2 : intercept 2	—	—	−4.40(0.70)
τ_3 : intercept 3	—	—	−5.56(0.68)
τ_4 : intercept 4	—	—	−6.27(0.67)
τ_5 : intercept 5	—	—	−7.80(0.68)

ZINB, zero-inflated negative binomial; OZINB, ordinal zero-inflated negative binomial; age13 is age centered at 13 years old (e.g., age-13). The parameter estimates from the ZIPO model are not directly comparable with those from the ZINB and OZINB models because they operate in different metrics. MDE, major depressive episode.

However, unlike our single process models (NB, ONB, PO, and linear), depression did not significantly predict increased frequency of alcohol use. Thus, the proposed OZINB permitted a more rigorous evaluation of the subtle relationship between depression and substance use, which is highly debated in applied research [27–30]. Similar to the NB and ONB models, the parameter estimates between the ZINB and OZINB were quite close. The only OZINB parameter estimates that did not fall in the 95% CIs for the corresponding ZINB parameter estimates were the dispersion parameter (α) and the intercept parameter from the count process (β_0). In sum, the proposed OZINB model recovered the underlying count process in a manner that is impossible using the existing ZIPO model.

5. Discussion

We have introduced a novel modeling framework for ordinal data that represent ranges of underlying counts. We demonstrated how ordinal responses can be explicitly linked to an underlying count construct of substantive interest through cumulative probabilities. We believe that this ordinal-count modeling framework offers both quantitative and substantive advantages over currently available methods. While standard models often violate assumptions and lack clear substantive interpretations, the ordinal modeling framework that we have outlined overcomes many of these statistical limitations and offers rich substantive interpretations.

In our simulations, the ONB and OZINB models performed similarly to their count model counterparts across conditions. This indicated that, assuming the underlying construct of interest truly follows a count distribution (e.g., NB), our proposed ordinal-count models perform well. The simulations also indicated that the proposed ordinal-count models outperformed existing models with regard to the rate of model convergence, empirical power, and relative efficiency. These findings were further buttressed by the comparable performance of the proposed ordinal-count models and the standard count models in the empirical adolescent alcohol use example. Importantly, the proposed OZINB model was able to obtain similar parameter estimates and substantive conclusions compared with the count ZINB, while this was not the case for the ZIPO model. These discrepant results were important because they illustrated that the ZIPO model can result in different patterns of effects compared with the ZINB and OZINB models. From substantive standpoint, the OZINB model is also more consistent with the actual construct of interest (e.g., number of drinking days) than the ZIPO model. Taken together, the simulation and empirical analysis demonstrated that ordinal-count models offer utility above and beyond existing ordinal modeling techniques if the true construct of interest is a count, not ordinal response categories.

Limitations and future directions

Our work here considered two potential ordinal-count model applications. We focused on ONB and OZINB models because they aligned with our substantive goal of assessing adolescent alcohol use. However, there is a host of other possible models that we did not consider including Poisson models, NB models with heterogeneous dispersion, Poisson and NB Hurdle models, and more (see Hilbe [21], for a review of potential count models). One limitation of our ordinal-count modeling framework is that it assumes that the cut-points are fixed and known. If these cut-points are not known, our modeling strategy cannot be implemented. Furthermore, ordinal-count models face limitations similar to their standard count model counterparts. For instance, if the count distribution is misspecified, the accuracy of results will be impacted and more complicated models such as models for zero inflation require relatively large sample sizes.

We also did not consider the impact of measurement error and unreliability in participant recall, which likely arises in practice. Indeed, these are important issues to address in future research. Although it does not undermine the findings of the current study, future research should also investigate the performance of a wider array of ordinal-count models under various conditions (e.g., missing data, different cut-point selections, misspecification of the latent response variable distribution, and alternative generating distributions), develop goodness-of-fit tests, and extend the models in meaningful ways (e.g., inclusion of random effects). Despite these potential limitations, we believe that our ordinal-count framework offers researchers quantitative and substantive advantages compared with existing techniques that can progress substance use research in meaningful ways.

Acknowledgements

This research was supported by grants F31DA033688 awarded to the first author and R01DA15398 co-awarded to the second author and Dr Andrea Hussong.

References

1. Johnston LD, O'Malley PM, Bachman JG, Schulenberg JE. The rise in teen marijuana use stalls, synthetic marijuana use levels, and use of 'bath salts' is very low University of Michigan News Service: Ann Arbor, MI. Available from: http://www.monitoringthefuture.org. [Accessed on June 14 2013].
2. Miller TR, Levy DT, Spicer RS, Taylor DM. Societal costs of underage drinking. *Journal of Studies on Alcohol and Drugs* 2006; **67**:519–528.
3. United States Department of Health and Human Services. *The surgeon general's call to action to prevent and reduce underage drinking*. USDHHS, Office of the Surgeon General: Washington, DC, 2007.
4. Trim RS, Meehan BT, King KM, Chassin L. The relation between adolescent substance use and young adult internalizing symptoms: findings from a high-risk longitudinal sample. *Psychology of Addictive Behaviors* 2007; **21**:97–107.
5. Dawson DA. Methodological issues in measuring alcohol use. *Alcohol Research and Health* 2003; **27**(1):18–29.
6. Currie C, Zanotti C, Morgan A, Currie D, de Looze M, Roberts C, Samdal O, Smith ORF, Barnekow V (eds). Social determinants of health and well-being among young people. Health Behaviour in School-aged Children (HBSC) study: international report from the 2009/2010 survey. WHO Regional Office for Europe, (Health Policy for Children and Adolescents, No. 6), Copenhagen, 2012.
7. Johnston LD, O'Malley PM, Bachman JG. *Monitoring the Future National Results on Adolescent Drug Use: Overview of Key Findings, 2011*. Institute for Social Research, The University of Michigan: Ann Arbor, 2012.
8. National Institute on Alcohol Abuse and Alcoholism. *Recommended alcohol questions*, 2003. Available from: http://www.niaaa.nih.gov/research/guidelines-and-resources/recommended-alcohol-questions [Accessed on June 3 2013].
9. Curran PJ, Willoughby MT. Implications of latent trajectory models for the study of developmental psychopathology. *Development and Psychopathology* 2003; **15**(3):581–612.
10. Buu A, Li R, Tan X, Zucker RA. Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. *Statistics in Medicine* 2012; **31**(29):4074–4086.
11. Kelley ME, Anderson SJ. Zero inflation in ordinal data: incorporating susceptibility to response through the use of a mixture model. *Statistics in Medicine* 2008; **27**:3674–3688.
12. Harris MN, Zhao X. A zero-inflated ordered probit model, with an application to modeling tobacco consumption. *Journal of Econometrics* 2007; **141**:1073–1099.

13. Carter WH, Jr., Bowen JV, Jr., Myers RH. Maximum likelihood estimation from grouped Poisson data. *Journal of the American Statistical Association* 1971; **141**(334):351–353.
14. Carter WH, Myers RH. Maximum likelihood estimation from linear combinations of discrete probability functions. *Journal of the American Statistical Association* 1973; **68**(341):203–206.
15. Moffatt PG. Grouped Poisson regression models: theory and an application to public house visit frequency. *Communications in Statistics-Theory and Methods* 1995; **24**(11):2779–2796.
16. Moffatt PG, Peters SA. Grouped zero-inflated count data models of coital frequency. *Journal of Population Economics* 2000; **13**(2):205–220.
17. Agresti A. *Categorical Data Analysis*. John Wiley & Sons: New York, 2002.
18. Koran J, Hancock GR. Using fixed thresholds with grouped data in structural equation modeling. *Structural Equation Modeling* 2010; **17**(4):590–604.
19. United States Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. Center for Behavioral Health Statistics and Quality. *National Survey on Drug Use and Health, 2010*. Inter-university Consortium for Political and Social Research: Ann Arbor, MI, 2012.
20. McGinley JS, Curran PJ. Validity concerns with multiplying ordinal items defined by binned counts: an application to a quantity-frequency measure of alcohol use, 2012. Under Review.
21. Hilbe J. *Negative Binomial Regression*. Cambridge University Press: New York, 2011.
22. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**(1): 1–14.
23. Greene WH. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. *Working Paper*, Department of Economics, Stern School of Business, New York University: New York, 1994.
24. Core Team R. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2012.
25. SAS Institute Inc. *SAS/STAT 9.2 User's Guide*. SAS Institute Inc: Cary, NC, 2008.
26. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* 2001; **6**:330–351.
27. Chassin L, Pillow DR, Curran PJ, Molina BS, Barrera M. Relation of parental alcoholism to early adolescent substance use: a test of three mediating mechanisms. *Journal of Abnormal Psychology* 1993; **102**(1):3–19.
28. Cooper ML, Frone MR, Russell M, Mudar P. Drinking to regulate positive and negative emotions: a motivational model of alcohol use. *Journal of Personality and Social Psychology* 1995; **69**:990–1005.
29. Hallfors DD, Waller MW, Bauer D, Ford CA, Halpern CT. Which comes first in adolescence—sex and drugs or depression? *American Journal of Preventive Medicine* 2005; **29**(3):163–170.
30. Hussong AM, Curran PJ, Chassin L. Pathways of risk for accelerated heavy alcohol use among adolescent children of alcoholic parents. *Journal of Abnormal Child Psychology* 1998; **26**(6):453–466.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.