EVALUATING CURRENT PRACTICES IN MEASURING AND MODELING ADOLESCENT ALCOHOL FREQUENCY DATA

James S. McGinley

A thesis submitted to the faculty of 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.

Chapel Hill
2011

Approved by:

Patrick J. Curran, Ph.D.

Daniel J. Bauer, Ph.D.

Andrea M. Hussong, Ph.D.

ABSTRACT

JAMES S. MCGINLEY: Evaluating Current Practices in Measuring and Modeling Adolescent Alcohol Frequency Data
(Under the direction of Patrick Curran)

Substance use is a significant health risk behavior from both developmental and public health perspectives. In recent years, there has been substantial growth in the theoretical conceptualization of pathways to substance use during adolescence. However, in order to test these developmental theories researchers must be able to validly measure and model substance use. This project evaluated the current standard practices in measuring and modeling adolescent alcohol frequency data. Using a simulation study and empirical demonstration, I investigated the degree to which the quantitative characteristics of ordinal measures and ordinal scoring approaches impact researchers' ability to draw valid inferences from standard linear models. My results showed that ordinal alcohol frequency measures interacted with scoring approaches to substantially reduce statistical power and led to different patterns of effects. There was no clearly superior ordinal scoring approach and, in some conditions, the performance of scoring approaches depended on which measure was used.

ACKNOWLEDGEMENTS

I would like to thank my wife, Melissa, and my family for continuously supporting me throughout this research project. Also, I would like to thank my advisor, Dr. Patrick Curran, and the rest of the L.L. Thurstone Psychometric Laboratory for giving me critical feedback that helped to improve this project.

DEDICATION

This is dedicated to Melissa.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

Evaluating Current Practices in Measuring and Modeling Adolescent Alcohol Frequency Data

Adolescent substance use is a widespread concern in the United States. The Monitoring the Future (MTF) study, a national school-based survey, found that in 2009, 36.6% of 8th graders, 59.1% of 10th graders, and 72.3% of 12th graders reported drinking alcohol at least once in their lifetime and 17.4% of 8th graders, 38.6% of 10th graders, and 56.5% of 12th graders reported getting drunk at least once in their lifetime (Johnston, O'Malley, Bachman, & Schulenberg, 2009). In 2001 alone, it was estimated that underage drinking cost the United States 61 billion dollars (Miller, Levy, Spicer, & Taylor, 2006). More concerning are the non-monetary consequences of adolescent substance use. In the short-term, adolescent substance use is associated with morbidity, driving accidents, risky sexual behavior, and even death (USDHHS, 2007). There is also evidence that substance use in adolescence has negative long-term biological effects such as disruptions in neuropsychological development and performance (Tapert, Caldwell, & Burke, 2005) and female pubertal development (Emanuele, Wezeman, & Emanuele, 2002). Several studies have found a relationship between adolescent substance use and lower educational attainment, difficulties transitioning from adolescence into young adulthood (Hussong & Chassin, 2002), and psychological problems in adulthood (Trim, Meehan, King, & Chassin, 2007). Clearly, adolescent substance use is a significant public health concern.

Recently, there has been tremendous growth in the theoretical conceptualization and empirical evaluation of pathways to substance use during adolescence. These theoretical models of adolescent substance use range from broader deviance proneness, biological, and internalizing models to more complex integrative models (e.g., Scheier, 2010; Schulenberg &

Maslowsky, 2009). Deviance proneness models, which operate off the tenet that substance use occurs along with the development of general conduct problems, and biological models, which posit that adolescents with family histories of drug abuse and dependence are at greater risk for substance use, have been well supported by prior research (Chassin, Hussong, & Beltran, 2009; Chassin, Presson, Pitts, & Sherman, 2000). On the other hand, empirical tests of the relationship between internalizing symptomatology and the etiology of substance use in adolescence remain inconclusive with some studies reporting a significant association between them (Chassin, Pillow, Curran, Molina, & Barrera, 1993; Cooper, Frone, Russell, & Mudar, 1995), while others have failed to find such a link (Hallfors, Waller, Bauer, Ford, & Halpern, 2005; Hussong, Curran, & Chassin, 1998). This tremendous growth in adolescent substance use research over the past decade and a half needs to continue well into the future.

However, in order to empirically evaluate any of these theories, researchers must be able to validly and reliably measure and model the substance use outcomes of interest. Substance use measures examine alcohol or other drugs such as marijuana, cocaine, and heroin. Dimensions of substance use commonly examined include abuse, dependence, consequences, and frequency and quantity of use. It has been well documented that many adolescent substance use measures lack sufficient psychometric properties (Leccese & Waldron, 1994). In part, this psychometric deficiency is caused by the complex nature of adolescent substance use measurement. For instance, it is difficult to determine the reliability and validity of instruments that assess multiple substances (e.g., alcohol, cigarettes, marijuana, heroin, etc.) and dimensions (e.g., abuse, dependence, quantity/frequency). Although methods for assessing these different substances share much in common, my project focused strictly on frequency of alcohol use. The issues studied in this project are expected to generalize to other substances.

Measuring Alcohol Use

Over the past half-century, numerous alcohol use measures have been proposed, most of which can be classified into one of three categories: daily drinking, lifetime drinking, and quantity-frequency measures. First, daily drinking measures assess daily alcohol consumption for a specified time period (e.g., Sobell and Sobell's (2000) Alcohol Timeline Followback (TLFB) and Miller and Del Boca's (1994) Form 90). Advantages of these measures are that they provide more precise estimates of drinking than other techniques and they may be used to distinguish specific drinking patterns such as weekend drinking or types of drinkers (e.g., heavy episodic drinkers). Disadvantages include that they can take a great deal of time to complete and it can be difficult for participants to recall their exact alcohol consumption for days in the distant past (Sobell & Sobell, 1995). Second, lifetime drinking measures require that participants recall typical drinking patterns from adolescence through the present, providing a developmental overview of alcohol use (e.g., Skinner and Sheu's (1982) Lifetime Drinking History). These instruments face substantial criticism because they rely heavily on long-term retrospective recall and they lack precision (Skinner and Allen 1982). These measures are also time consuming and can be burdensome to complete. Third, and possibly most widely used, are quantity-frequency (QF) measures, which gather data on typical alcohol consumption. Participants are asked a question about their typical rate of alcohol consumption, frequency (F), and a question about their average quantity per drinking occasion (Q). The responses of two questions are multiplied together (e.g., QxF) to provide an estimate of the total volume consumed.

Researchers often choose to utilize either the quantity per drinking day or frequency of alcohol use independently to test specific research hypotheses. Quantity-frequency measures offer several practical advantages such as short administration time, easy computations, intuitive meaning, and researchers can use quantity or frequency measures separately to test unique effects involving alcohol use. However, these methods have been criticized on a variety of grounds including the underestimation of true alcohol consumption, as well as the omission of important information about variability in alcohol consumption patterns (Dawson & Room, 2000; Ivis, Bondy, & Adlaf, 1997; Sobell & Sobell, 1995).

In this project, I focused on frequency of alcohol use because it is widely used in applied research and the findings likely generalize to similar types of substance use measures. By definition, alcohol frequency data are counts because participants report on the number of days in which they drank alcohol over a given timeframe. Despite this fact, most researchers provide binned ordinal response categories for frequency of alcohol use (e.g., 1-3 times a month, 1 time per week, etc.) rather than leaving responses open-ended because it lessens participant burden and errors in cognitive recall (Ivis et al., 1997). To help standardize these measures, the National Council on Alcohol Abuse and Alcoholism established and recommended sets of alcohol consumption questions, totaling between 3 and 6 questions (NIAAA, 2003). An example of a frequency item modified for a past 30 day timeframe is displayed in the first column of Table 1. Importantly, the committee recommended using binned ordinal response categories.

Previous psychometric research has focused on evaluating alcohol frequency measures from a traditional validity (convergent, divergent, predictive, etc.) and reliability (test-retest, internal consistency, etc.) perspective, but there is a dearth of research investigating how these alcohol frequency measures perform when testing theories of adolescent alcohol use with misspecified statistical models (e.g., fitting linear models to ordinal data). Even if an alcohol measure has strong traditional psychometric support, it can still produce invalid tests of substantive theories in commonly used statistical models.

For example, consider if there was a true effect of age on frequency of alcohol use such that older adolescents, on average, drank more frequently than younger adolescents. Two researchers may measure frequency of alcohol in the same participants using two different measures; one with five point scales the other with a 12 point scale, assuming perfect reliability from the reporter. Although both of these measures show strong traditional psychometrics properties and the same statistical model is fitted to the data, one model may find a significant age effect while the other does not because of a reduction in statistical power moving from 12 categories to five categories (MacCallum, Zhang, Preacher, & Rucker, 2002; Taylor, West, & Aiken, 2006). For another example, assume alcohol frequency is measured with two seven point scales with response categories characterized by different bin sizes. A model fitted to the data derived from one of the seven point scales could produce a significant age effect while the age effect turns out to be non-significant using the other seven point scale. This discrepancy has the potential to occur in practice because different binning methods may produce different relationships between a set of covariates and the outcome. These two inconsistences exemplify invalidity in testing substantive theory using statistical models because the different patterns of effects are due to the alcohol measures.

Current Practices in Modeling Alcohol Frequency Data

The standard practice in adolescent substance use research is to test theoretical models by treating an ordinal alcohol frequency outcome as continuous in a linear statistical method (A few recent examples of this are Dogan, Stockdale, Widaman, and Conger, 2010; Rice, Milburn, & Monro, 2011; Patrick & Schulenberg, 2011). These linear models are not ideal from a statistical standpoint, but researchers use this strategy because alternative statistical models are not well studied or readily available (Curran & Willoughby, 2003). Furthermore, closer examination of the distributional properties and generalized model techniques for discrete alcohol frequency data helps to clarify why standard linear models are so widely utilized by alcohol researchers. Through briefly exploring the Generalized Linear Model (GLM) framework, I will provide insights as to why using ordinal alcohol frequency data in traditional linear models can lead to invalid tests of substantive theory.

Distributions of Alcohol Frequency Data

Alcohol researchers frequently treat ordinal alcohol frequency data as continuous in linear models. In doing so, they inherently assume that the ordinal outcome follows the probability density function (pdf) for the normal distribution

$$P(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \quad (1)$$

The parameters μ and σ^2 are the mean and standard deviation. Figure 1 shows the standard normal distribution with $\mu = 0$ and $\sigma^2 = 1$. This pdf shows that alcohol frequency cannot be normal because it implies that values range from negative infinity to positive infinity on a continuous scale. The underlying goal of alcohol frequency measures is to gather a count of the number of days in a given timeframe that an adolescent consumed alcohol. Alcohol frequency counts are characterized by a large proportion of non-alcohol users, also known as zero-inflation in the statistics literature. For this reason, alcohol frequency counts likely conform to a negative binomial distribution rather than the more familiar Poisson distribution because its added dispersion parameter allows for much more flexibility. The probability mass function (pmf) for the negative binomial distribution can be expressed as

$$g(x) = P(y) = \frac{\theta^{\theta}\mu^{y}\Gamma(\theta + y)}{\Gamma(y+1)\Gamma(\theta)(\mu+\theta)^{\theta+y}}, \qquad y = 0,1,2,... \tag{2}$$

where the distribution has mean of μ and a variance of $\mu(1+(1/\theta)\mu)$. Figure 1 show the negative binomial distribution with $\mu = 2$ and $\theta = .8$. A dual process, zero-inflated negative binomial (ZINB) process is a second viable option for characterizing adolescent alcohol frequency data. The ZINB probability function can be expressed as

$$P(y) = \begin{cases} \vartheta + (1-\vartheta)g(0) & if \ y = 0 \\ (1-\vartheta)g(y) & if \ y > 0 \end{cases} \tag{3}$$

where ϑ is the probability of zero counts and the function g(.) represents counts drawn from the negative binomial distribution as described in Eq. 2. Figure 1 shows the ZINB distribution after adding an excess zero probability $\vartheta = .42$ to negative binomial distribution described above. This probability function for the ZINB implies that zeros are generated from two sources: (1) the inflated zeros probability (e.g., structural zeros) and (2) the expected zeros from a negative binomial distribution (e.g., sampling zeros).

However, most adolescent alcohol studies collect frequency data using ordinal scales with many response categories. This inherent process of binning raw alcohol frequency counts makes it difficult to identify the underlying distribution because it is clearly no longer a count variable and ordinal models with a large number of response categories are often intractable. For example, Figure 2 shows how the ZINB distribution from Figure 1 is affected after binning the counts into ordinal categories based on NIAAA scale from Table 1. This binning process makes the selection and implementation of an appropriate statistical method for validly testing theoretical hypotheses of adolescent alcohol use challenging for applied research.

Clearly, there is a disconnect between the distributions associated with the underlying count of days using alcohol (Equations 2 and 3) and what researchers often assume in statistical models (Equation 1). This incongruity is worsened by the process of binning frequency counts in ordinal categories. By briefly exploring the GLM, I will highlight why adolescent alcohol researchers often use linear models for ordinal alcohol frequency data and form the foundation for describing how linear models fit to ordinal data can lead to invalid statistical inferences.

Generalized Linear Model (GLM) and Alcohol Frequency Data

The GLM offers a unifying modeling framework that subsumes traditional continuous linear models with various models for discrete outcomes. The GLM operates on three components: (1) Stochastic Component - this is commonly thought of as the error structure of response distribution, (2) Systematic Component - this is how the predictors affect the outcome that is transformed through the specified link function (e.g., $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$), and (3) Link Function –this connects the Stochastic Component with the Systematic Component (e.g. $g(\boldsymbol{\mu}) = \boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$). I will next lay out how the GLM encompasses continuous and discrete outcomes.

By fitting standard linear models to alcohol frequency data, adolescent alcohol researchers connect their set of linear predictors (e.g., the systematic component) to the expected value of the specific exponential form through the identity link function,

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \tag{4}$$

where **X** is an *n x p* matrix of covariates, $\boldsymbol{\beta}$ is *p x 1* vector of regression coefficients, and the outcome follows a normal distribution (e.g., multiple regression, ANOVA). However, if researchers were fitting models to the underlying count of the number of data of alcohol use, they would likely use a logarithmic link function such that

$$\log(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} \tag{5}$$

where **X** is an *n x p* matrix of covariates, $\boldsymbol{\beta}$ is *p x 1* vector of regression coefficients, and the outcome follows a negative binomial distribution. The ZINB model may also be useful for modeling adolescent alcohol data. The ZINB model is a two-component mixture model combining a point mass at zero with a negative binomial distribution. Zeros arise from two sources, the probability of excess zeros and the zeros naturally occurring in the NB distribution. The mean models for the ZINB model can be expressed as

$$\boldsymbol{\mu} = \vartheta \cdot 0 + (1 - \vartheta) \cdot \exp(\mathbf{X}\boldsymbol{\beta}) \tag{6}$$

where **X** is an *n x p* matrix of covariates, $\boldsymbol{\beta}$ is *p x 1* vector of regression coefficients for the count process, and the outcome follows a negative binomial distribution. The unobserved probability of excess zeros ϑ is modeled with a binomial GLM with a logit link as

$$\operatorname{logit}(\vartheta) = \mathbf{Z}\boldsymbol{\gamma} \tag{7}$$

Z is an *n x q* matrix of covariates, ϑ is *q x 1* vector of regression coefficients for the zero process, and the outcome follows a binomial distribution. Alcohol researchers rarely collect frequency data as open-ended counts, so we can consider a basic ordinal model: the proportional odds model which is defined as

$$\operatorname{logit}[P(Y \le j \mid \mathbf{x})] = a_j + \boldsymbol{\beta}'\mathbf{x}, \quad j = 1, \ldots, J - 1 \tag{8}$$

The outcome is now modeled as cumulative logits, which are simply logits of cumulative probabilities for the *j* response categories. Each cumulative logit has its own intercept expressed as α_j and the model implies the same effect, $\boldsymbol{\beta}$, for each logit (this is the "proportional odds" assumption). The outcome is assumed to follow a multinomial distribution. These models are quite complex and have strong assumptions, which are often unfeasible for alcohol frequency data. For instance, a frequency scale with 10 response categories would need to have 9 intercepts. This is precisely why applied researchers fit linear models such as described in Equation 4 to ordinal alcohol frequency data.

There are stark differences between the models expressed in Equations 4 through 8. Fitting the negative binomial model is often not an option because alcohol frequency counts are seldom collected by researchers. Although the utilization of linear statistical models to ordinal alcohol frequency data is often defensible from a practicality standpoint, these models are highly susceptible to producing invalid statistical inferences. The factors producing these invalid inferences can again be tied to the GLM framework.

Alcohol Frequency and Validity of Inferences

Two factors are critical in drawing valid inferences from models testing theories of adolescent alcohol use. The first factor consists of the specific characteristics of alcohol frequency measures. These characteristics include the number of response categories and range of drinking occasions represented within each ordinal category (e.g., Does each category represent equal range of drinking occasions or does the range of drinking occasions increase as the category number increases such as the measures in Table 1). The second factor pertains to how scores are assigned to the ordinal data so that they can be fitted in standard linear models. Adolescent alcohol use researchers often use one of two scoring approaches: category numbers and the midpoint value within each ordinal category. In practice, both of these factors vary in adolescent alcohol use applications and there is currently no gold standard alcohol frequency measure or scoring approach. Often times, researchers do not even report the characteristics of a measure or how they scored the ordinal alcohol variable. In my project, I will empirically examine whether or not these two factors are critical for validly testing theories of adolescent alcohol use.

Characteristics of Alcohol Frequency Measures

The quantitative elements of alcohol frequency measures are the number of ordinal response categories and the method of binning the underlying days of alcohol use counts into ordinal response categories. While prior quantitative research suggests that a larger number of response categories should be better suited for testing models of adolescent alcohol use than fewer categories because researchers lose less information (Taylor, West, & Aiken, 2006), it is unclear how the method of binning underlying counts into ordinal categories affects statistical inferences. It is plausible that the number of categories may interact with binning strategies such that the effect of the number of ordinal categories on the validity of results generated from statistical models depends on how the response categories are binned.

Quantitative researchers refer to the process of binning an underlying continuous or count quantitative variable into a smaller number of ordered categories as coarse categorization (Taylor, West, & Aiken, 2006). Several studies have shown that coarse categorization can be problematic from a quantitative standpoint. For instance, MacCallum, Zhang, Preacher, and Rucker (2002) detailed the statistical repercussions caused by dichotomization (e.g., performing a median split) such as loss of power and effect size, reduction in reliability, and the possible introduction of spurious effects. Similarly, Taylor, West, and Aiken (2006) expanded this work to show that coarsely categorized ordinal outcomes lead to a loss of power in logistic, ordinal logistic, and probit regression. Findings from this study suggest that, typically, more categories that have a rectangular distribution are better than fewer categories that are skewed. However, this study assumed that underlying these categories was a normal continuous variable.

Far less information exists about the statistical implications of binning counts into a series of unequally spaced ordinal responses, which is the case for alcohol frequency measures. Note that for the NIAAA alcohol frequency question, moving from the zero category to the one category is not the same as moving from the three category to the four category. As previously stated, alcohol researchers use ordinal measures with unequal categories to minimize errors in cognitive recall. However, the implications of using these measures with standard linear models are currently unclear. It appears that binning alcohol frequency counts into categories may act as pseudo data transformation. Consider the popular logarithmic transformation for use in linear models, which converts multiplicative relationships to additive relationships and consequentially transforms exponential trends to linear trends. Log transformations are popular with negative binomial distributed data (see Equations 2 and 3) because they pull outlying data from a positively skewed distribution closer to the bulk of the data. Examining the NIAAA alcohol frequency item in Table 1, by binning the frequency counts into increasingly larger categories, the larger outlying counts are being pulled in closer to the rest of the data if the categories numbers are used as scores. This idea of logarithmically transforming frequency counts is also consistent with how the underlying count data would be handled in the GLM.

Equation 5 displays the negative binomial model. Notice that the appropriate nonlinear link function is the logarithmic link. The regression coefficients can be interpreted in terms of changes in the transformed mean response in the study population, and their relation to the set of covariates. From a statistical standpoint, it appears that a strategy that bins alcohol frequency counts in increasingly larger categories that mirror the logarithmic transformation should be best for linear modeling (assuming one uses the category number as the alcohol frequency scores). However, in practice, there is no widely accepted method for defining alcohol frequency categories. Most researchers do not report the response scales of their alcohol frequency measures in academic journals so the extent to which measures vary in practice is unknown. Prior research in statistics has shown that applying different data transformations to the same data can indeed lead to different patterns of statistical significance and inaccurate predictions (Adams, 1991; O'Hara & Kotze, 2010). Extending this finding to adolescent alcohol use, it is expected that the more these alcohol measures vary in their binning approaches (e.g., the more the pseudo data transformations differ), the more likely researchers are to observe invalid patterns of effects caused solely by measurement.

Scoring Approaches for Alcohol Frequency Data

In the field of adolescent alcohol use, there are two primary methods for creating alcohol frequency scores based on ordinal measures. The first scoring method uses the category number. For example, using the NIAAA frequency measure in Table 1, the scores to be used in the linear statistical models would be the integers ranging from zero to seven. The second method for scoring is to use the median frequency value within each category. The median approach involves taking mid-value within each category. Using the NIAAA frequency measure, the scores would be 0, 1, 2.5, 4.5, 7.5, and so on. Although the differences between these scoring methods appear trivial, they have the ability to seriously impact the validity of hypothesis tests concerning adolescent alcohol use. These scoring approaches for ordinal data have yet to be systematically studied in the context of adolescent alcohol use.

From a statistical viewpoint, the category number scoring approach appears advantageous to the median approach. As previously described, many ordinal alcohol frequency measures functionally work as a data transformation that helps to linearize the relationship between a set of predictors and the alcohol frequency outcome. This is consistent with the logarithmic link model expressed in Equation 5. Conversely, the median approach takes the ordinal alcohol frequency response and converts it back to a metric similar to the underlying frequency count (e.g., Equations 2 and 3). In doing so, it likely introduces a nonlinear relationship between the set of predictors and the alcohol frequency outcome and applying linear statistics models to these median scores worsens the degree of model misspecification, which can seriously affect the reliability and validity of tests of substantive theory (Long, 1997). Relating this back to the GLM, this is akin to fitting negative binomial distributed data (e.g., Equation 2) with the linear model expressed in Equation 4. This fact about median scores in adolescent alcohol use research goes widely unnoticed or, worse, is misunderstood. For example, a large epidemiological study of adolescent drunkenness by Kuntsche et. al (2011) states that "midpoints of categories were used to create a linear measure". This statement is in direct contrast to what is actually occurring from a statistical standpoint. Clearly, there is strong rationale for using the category numbers for scores in linear models. However, the current recommendation from the field of biostatistics is to use median scores (Agresti, 2002). Prior research has yet to rigorously study how these ordinal scoring approaches impact our ability to test theories of adolescent substance use.

Summary

Adolescent substance use is a significant public health concern. Over the past half century, many measures of adolescent alcohol use have been developed, but perhaps the most widespread are quantity-frequency measures. My project focused on frequency of alcohol use measures. The vast majority of frequency measures assess alcohol use using ordinal response categories. However, prior research has not investigated the impact of using these ordinal measures in standard linear models. Statistical theory suggests that there is a difference between the distributional assumptions of standard linear models and characteristics of alcohol frequency data. This difference has a strong potential to affect the validity of inferences drawn from statistical tests of substantive theory. Two factors that may impact the validity of inferences are the characteristics of alcohol frequency measures (e.g., number of response categories and binning method) and the scoring method (category number or median value). It is currently unclear which combination of measurement characteristics and scoring method is optimal for adolescent alcohol use research.

My project used a simulation study and an empirical demonstration to evaluate the current measurement and modeling practices used in cross-sectional adolescent alcohol research. In my project, three core hypotheses were tested. First, there should be an interactive effect between the alcohol frequency measure and the scoring approach. More precisely, I expected that alcohol measures with few response categories that are defined to be dissimilar to the logarithmic transformation should be more sensitive to scoring approaches, especially category scores, compared to measures with more response categories that are binned similarly to the logarithmic transformation. Second, the validity of statistical inferences should depend on the quantitative characteristics of the alcohol frequency measures. There should be a general tendency for alcohol measures with few response categories that are defined to be different from the logarithmic transformation to perform poorer than alcohol measures with more response categories that closely follow the logarithmic transformation, regardless of the scoring approach. Third, the validity of inferences drawn from statistic tests of substantive hypotheses should depend on the scoring approach used. Generally, I expected that category number scores should outperform the median scoring approach for scales with a reasonable number of categories that are defined similar to a log transformation (this trend should not hold for measures with few poorer defined response categories).

Method

The simulation study was a four-step process starting with generating count data from a ZINB model. Then, these data were binned into a series of ordinal alcohol measures and scored with multiple ordinal scoring approaches. After scoring, I fitted linear regression models to each of the measure-by-scoring combinations and evaluated the results of the simulation in terms of the proportion of significant effects, Type I and II errors, and percentage of different patterns of effects.

A similar analytic process was applied to empirical data from the National Longitudinal Survey of Youth (NLSY97). Again, I binned the open-ended count data into ordinal alcohol frequency scales and scored the ordinal data. The scored ordinal data were then fitted with linear regression models so that each of the measure-by-scoring combinations could be evaluated in terms of proportion of significant effects and percentage of different patterns of effects.

Simulation Study

My simulation study had four steps. First, count data were generated from known population models. Second, the count data from Step 1 were binned in categories according to the prescribed ordinal alcohol frequency measures. Third, the ordinal data from Step 2 were scored according to the prescribed scoring methods. Fourth, standard linear regression models were fitted to the scored data from Step 3.

Step 1: Data Generation

To be consistent with commonly observed distributions of adolescent alcohol frequency data for a past 30 day timeframe, the underlying count data were generated from a ZINB distribution in which the count process was conditioned on two predictors, one binary and one continuous. This was accomplished using Equations 6 and 7 presented earlier. The zero-process of the model (Equation 7) was not conditioned on covariates and did not vary across conditions. The zero-process was generated to have a probability of $\vartheta = .43$. The count process for the four effect size conditions were generated as follows:

$$\textbf{Condition1(Binary} - \textbf{MediumEffect)} : \log(\mu_i) = \begin{pmatrix} 1 & x_{1i} & x_{2i} \end{pmatrix} \begin{pmatrix} .75 \\ .7 \\ 0 \end{pmatrix}$$

$$\textbf{Condition2 (Binary} - \textbf{SmallEffect)} : \log(\mu_i) = \begin{pmatrix} 1 & x_{1i} & x_{2i} \end{pmatrix} \begin{pmatrix} .75 \\ .49 \\ 0 \end{pmatrix}$$

$$\textbf{Condition3(Cont.} - \textbf{MediumEffect)} : \log(\mu_i) = \begin{pmatrix} 1 & x_{1i} & x_{2i} \end{pmatrix} \begin{pmatrix} .75 \\ 0 \\ .34 \end{pmatrix}$$

$$\textbf{Condition4(Cont.} - \textbf{SmallEffect)} : \log(\mu_i) = \begin{pmatrix} 1 & x_{1i} & x_{2i} \end{pmatrix} \begin{pmatrix} .75 \\ 0 \\ .23 \end{pmatrix}$$

In these conditions, I generated x_{1i} as a binary predictor and x_{2i} as a continuous predictor. In all conditions, the dispersion parameter was generated as $\theta = .8$. Combining the zero and count processes, the ZINB mean model can be expressed as in Equation 6. First, the binary predictor x_{1i} had a medium effect size and the continuous predictor x_{2i} had no effect. This effect for x_{1i} was equal to group 1 having a mean of about 2.42 and group 2 having a mean of 1.22. Second, the binary predictor x_{1i} had a small effect size and the continuous predictor x_{2i} had no effect. This effect for x_{1i} was equal to group 1 having a mean of about 1.97 and group 2 having a mean of 1.22. Third, the binary predictor x_{1i} had no effect and the continuous predictor x_{2i} had a medium effect. This effect for x_{2i} was equal to about 1.16 when

x_{2i} was at the mean and 1.86 when x_{2i} was one standard deviation above the mean. Fourth, the binary predictor x_{1i} had no effect and the continuous predictor x_{2i} had a small effect. This effect for x_{2i} was equal to about 1.16 when x_{2i} is at the mean and 1.56 when x_{2i} was one standard deviation above the mean. Medium and small effects were defined as an empirical power of .8 and .5 with an n=250 based on fitting the population generating ZINB model to the count data. These generating values were motivated by the NLSY. The means, proportion of zeros, and shape of the generated data were generally aligned with these empirical data.

The simulation had a single sample size, n=250, because sample size was not expected to be influential beyond what is normally expected (e.g., statistical power). Each of the four data generation conditions were replicated 500 times resulting in a total of 2000 generated datasets. Figure 3 shows the marginal distribution of the simulated outcome for all 500 replications from Condition 1. Because counts had to be between 0-30, I recoded any counts greater than 30 to missing. This affected a very small percentage of the generated data (> 99.5% of the generated data generated across all of the conditions had counts between 0-30).

Step 2: Alcohol Measures

The underlying counts generated in each of the 2000 dataset produced in Step 1 were next binned into ordinal categories to conform to four alcohol frequency measures. The first measure had eight categories and was based on NIAAA recommendations. The second measure had five categories that were defined to be dissimilar to the log transformation. The third measure had eleven categories that were consistent with what is commonly observed in practice. These measures are displayed in Table 1.

Although there are already many different alcohol frequency scales, I also explored whether it was possible to draw on statistics (e.g., GLM theory) to improve ordinal frequency measures for use in standard linear models. This experimental fourth measure was a seven category measure that I created based on a logarithmic transformation. More specifically, I created the 7 categories by dividing the maximum log(day+1) by 7 and creating cutoffs based on increments of that magnitude (See Table 2). For instance, 3.43 divided by 7 is .49, so I created bins using a .49 cutoff for the log(days+1).

Step 3: Scoring Approaches

Three scoring approaches were applied to the ordinal data produced in Step 2. The first two scoring approaches used were the category numbers and median values within each category. The third scoring approach was an experimental approach I propose that draws on the logarithmic transformation. This approach simply takes the log of the median value within each category plus one (e.g., log(median+1)). I derived this experimental scoring method as an attempt to optimize the performance of ordinal scores in standard linear models using statistical theory.

Step 4: Model Fitting

To be consistent with applied research, standard linear regression models were fitted to the ordinal data scored in Step 3. Linear regression models were fitted to data using SAS PROC REG. Using Equation 4, the mean model can be expressed as

$$\mu_i = \begin{pmatrix} 1 & x_{1i} & x_{2i} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

In this equation, x_{1i} was the binary predictor and x_{2i} was the continuous predictor and the ordinal outcome was assumed to follow a normal distribution. I also fitted the population generating ZINB model to the count data using SAS PROC COUNTREG to validate data generation.

Simulation Evaluation

I used a series of meta-models to test for the potential interaction of measure-by-scoring approaches and the measure and score main effects. The meta-models were eight separate GEE models with logit link functions, binomial response distributions, and exchangeable correlation structures (4 effect size conditions by 2 predictors) fitted to binary outcome of Type II error (e.g., 0=No Type II error, 1=Type II error). I used GEE models to account for the correlations among scale-by-scoring combinations fitted the same underlying count data (the same general pattern of results were obtained using random effects models).The eight models had a total of 6,000 observations because for each of the 500 replications there were 12 lines of data representing the various measure-by-scoring approaches. I created reference codes for measures and scoring approaches to formally test for the main effects and their interaction. These meta-models were used only to test for omnibus scale-by-scoring effects or scoring and measure main effects when the interaction was non-significant. This was accomplished with Wald test statistics. I did not use any model implied probabilities or odds ratios to evaluate the simulation.

Because there was not a direct correspondence between the generating ZINB models and the multiple linear regression models fitted to the ordinal data, I was not able to evaluate the simulation with standard methods such as raw and standardized bias, root mean squared error, and effect sizes. I could not examine the parameter estimates across the multiple regression models because the alcohol frequency outcomes were on different metrics due to the different ordinal alcohol measures and scoring approaches. Given these conditions, I evaluated my simulation using three criterion.

First, I examined differences in the proportions of significant effects obtained when using the ordinal data and standard regression models compared to the counts fitted to population generating ZINB model. For additional comparison information, I also calculated the standardized regression coefficients for the multiple linear regression models fitted to the ordinal data. I did not do this for the population generating ZINB model because there were not satisfactory computational methods. Second, I examined Type I and II error rates and odds ratios for the various measure-by-scoring approaches compared to the population generating ZINB models. Third, I calculated the proportion of the generated datasets that produced different patterns of effects due to measures and scoring approaches despite having the same underlying count data. For example, assuming the same underlying data and alcohol frequency measure, if x_1 was significant using the category number scoring approach but not using median scoring approach, I labeled this as a “different pattern of effects”.

Simulation Summary

To summarize, 2,000 datasets containing count data were generated (500 replications per condition; n=250). For each of the 2,000 data sets, four alcohol frequency measures were used to bin the counts into ordinal data. Then, the ordinal data were scored using three approaches and linear regression models were fitted to the data. Thus, for each of the 2,000 simulated data sets, 12 models were fitted to all of the measure-by-scoring approach combinations (4 measures times 3 scoring approaches). Also, for comparison, the 2,000 ZINB models were fitted to the underlying count data.

Empirical Demonstration

About the NLSY

I employed an empirical demonstration to determine if standard linear models fitted to ordinal alcohol frequency data can lead to different substantive results in practice. I used repeated random samples instead of working with a single dataset as is typically done in empirical demonstrations because I wanted to work with a sample size that is reflective of those commonly observed in adolescent alcohol use studies. The study that I used for this demonstration had a very large sample size of almost 9,000 adolescents and young adults. Using repeated random sampling allowed me to specify a much smaller sample size (n=250) so that my findings would generalize to a broader audience and there was not excessive statistical power. Even without knowing the population generating model, repeated random sampling allowed me to describe my empirical demonstration in terms of proportion of significant effects and the proportion of datasets that had inconsistent patterns of effects caused by measures and scoring approaches.

Data for the empirical demonstration came from the first round of data collection for the NLSY (NLSY97). The NLSY was selected because, unlike most studies, they collected count data for alcohol frequency, which is necessary for my evaluation strategy. Briefly, the NLSY collected extensive information on several domains including educational experiences, employment data, delinquent behavior, alcohol and drug use, sexual activity, youth's relationships with parents, contact with absent parents, marital and fertility histories, dating, onset of puberty, training, participation in government assistance programs, expectations, and time use.

The first round of data collection consisted of a nationally representative sample of

8,984 participants ranging in age from 12 to 18 years old. The sample was 51% male and the Race/Ethnicity breakdown was 51.9% Non-black/non-Hispanic, 26% Black, 21.2% Hispanic or Latino, and 0.9% Mixed Race/Ethnicity.

Study Subsample

For the purposes of my project, a subsample was created. The subsample consisted of 4,442 14-16 year old adolescents (51.9% male; 33.2% 14 year olds, 34.1%15 year olds, 32.7% 16 year olds; 35.1% minority). There were very few 12 and 13 year olds (1.3% of the initial sample) and the NLSY did not collect relevant items concerning maternal monitoring for participants that were older than 16 years old so these ages were dropped. Additionally, because this demonstration used standard linear regression models for the analyses, participants that had missing data on any variables used for the analyses were dropped so that there was a common sample size across all models (11% of the 14-16 year old adolescents were dropped because they were missing on at least one covariate).

Measures

Alcohol Frequency

Alcohol frequency was a single open ended item, “During the last 30 days, on how many days did you have one or more drinks of an alcoholic beverage?”. The responses were discrete counts ranging from 0-30.

Maternal Monitoring

Maternal monitoring was a composite consisting of the mean of three items: “How much does she (mother) know about your close friends, that is, who they are?”, “How much does she (mother) know about your close friends' parents, that is, who they are?”, and “How much does she (mother) know about whom you are with when you are not at home?”. The items had a five point likert response scale: 0=”Know Nothing”, 1=”Knows Just A Little”, 2=”Knows Some Things”, 3=”Knows Most Things”, 4=”Knows Everything”. The items had a Cronbach’s alpha of .72.

Analytic Strategy

The analytic strategy for this empirical demonstration was the same as the simulation with the exception of Step 1. For the empirical demonstration Step 1 involved taking 1000 random samples of n=250 with replacement from the total subsample of n=4,442. The 1,000 unique datasets subsequently went through Steps 2-4 from the simulation strategy. Figure 4 shows the marginal distribution of alcohol frequency counts for the whole subsample of n=4,442. The linear regression model based on Equation 4 fitted to the data was

$$\mu_i = \begin{pmatrix} 1 & Age_i & Minority_i & Gender_i & Monitor_i \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

For comparison purposes, I fitted a negative binomial (NB) model with the same predictors to the count data, see Equation 5. The NB model provided what should be a more appropriate standard for evaluating the linear regression models fitted to the ordinal alcohol frequency data. I fitted NB models instead of the ZINB models because many of the 1,000 randomly sampled datasets appeared to be consistent with the NB distribution. In these cases, the fitted ZINB models often led to seemingly unstable estimates and convergence difficulties. Given this, I decided to use the more stable NB models for comparison.

Evaluating the Empirical Demonstration

The central goal of this empirical demonstration was to evaluate whether different alcohol frequency measures and scoring approaches may cause researchers to draw invalid inferences from linear statistical models. First, I calculated the different proportions of significant effects caused by measures and scoring approaches. I also provided the standardized regression coefficients for the predictor effects in the linear regression model. The results produced from the linear regression models fitted to the ordinal data were compared among each other and to the NB model fitted to the counts. Second, I calculated the proportions of the generated datasets that had different patterns of effects caused by measures and scoring approaches despite having the same underlying count data. These proportions only considered patterns of effects from the linear regression models fitted to the ordinal alcohol frequency data (did not consider the patterns of effects produced by the NB models fitted to the counts).

In sum, the goals of this study were to evaluate the impact of ordinal measures and scoring approaches on our ability to draw valid inferences from linear statistical models. The simulation study provided a highly controlled environment with known population generating models so that I could test my proposed research hypotheses. The empirical demonstration extended my simulation study so that I could assess how well the simulation results translated to real data. Taken together, these two components provided a holistic approach for rigorously testing the potential influence of ordinal measures and scores on researchers' tests of theoretical models of adolescent alcohol use.

Results

First, I will present results from the simulation study. I used omnibus Wald tests from the GEE meta-models to assess whether or not there was scale-by-scoring approach interaction effect (and main effects of scale and scoring, if the interaction was non-significant) on the probability of making Type I and II errors. To untangle the interaction effects, I examined various outcomes including the proportions of Type I and II errors, odds ratios for making Type I and II errors compared the population generating ZINB model, standardized regression coefficients from the linear regression models, and percentages of different patterns of effects. Second, I will present results from the empirical demonstration. Again, I used omnibus Wald tests from the GEE meta-models to generally assess whether or not there was scale-by-scoring approach interaction effect (and main effects of scale and scoring, if the interaction was non-significant) on the probability finding significant effects. I examined the proportions of significant effects, odds ratios for finding significant effects compared the NB models, standardized regression coefficients for the linear regression models, and percentages of different patterns of effects.

Simulation Study

Recovery of Population Generating Values

Table 3 shows the recovery of the population generating values by the ZINB models. The ZINB models showed fair recovery of the parameter estimates across all four effect size conditions. For example, the mean dispersion parameter (α, which is equivalent to $\frac{1}{\theta}$), intercept, x_1, and x_2 estimates over the 500 replications each fell within .05 of the population generating values across all conditions. The mean of the parameter estimates for the inflated zero portion of the model was downwardly biased. The parameter was generated to be -.3, but had average estimates of -.44 in condition 1 (binary predictor, x_1, had a medium effect), -.75 in condition 2 (binary predictor, x_1, had a small effect), -.73 in condition 3 (continuous predictor, x_2, had a medium effect), and -1.09 in condition 4 (continuous predictor, x_2, had a medium effect) across the 500 replications. Although the mean point estimate for the inflated zero parameter was biased across conditions, the standard errors were large. This reflected imprecision in the estimates and this is consistent with research that has suggested that parts of these models, including point estimates, can be sensitive to smaller sample sizes (Ghosh, Mukhopadhyay, & Lu, 2006).

Hypothesis 1: Scale-by-Scoring Approach Interaction

First, I evaluated my hypothesis that there would be a scale-by-scoring approach interaction effect on the inferences drawn from linear statistical models. The meta-models described earlier in the Methods section consistently showed a significant scale-by-scoring approach interaction on the probability of Type II errors across the four effect size conditions (for x_1: condition 1 $\chi^2(6)=56.94$, condition 2 $\chi^2(6)=27.18$; x_2: condition 3 $\chi^2(6)=56.56$, condition 4 $\chi^2(6)=35.73$; $p<.0001$ for all tests). To help understand these interactions, I examined outcomes based on the raw data from simulation results, not the parameter estimates from the GEE meta-models. Table 4 through Table 7 display the proportion of significant effects, proportion of Type I and II errors, odds ratios for Type I and II errors for the scale-by-scoring combinations compared to the population generating ZINB models, and the standardized regression coefficients across the four effect size conditions.

Results showed that the five point measure interacted with scoring approaches differently than the other three measures. For example, the eight point NIAAA-motivated, eleven point, and seven point experimental log scales had a clear pattern of median scores (e.g., Condition 1: 8pt NIAAA 29%; 11pt 29% ; 7pt log 31%) producing a reduced percentage of Type II errors compared to category number scores (e.g., Condition 1: 8pt NIAAA 40%; 11pt 36%; 7pt log 41%), which had reduced percentage compared to log median scores (e.g., Condition 1: 8pt NIAAA 44%; 11pt 43%; 7pt log 44%). However, this pattern in the scoring approaches did not hold for the five point measure. The percentage of Type II errors using the log median scoring approach (Condition 1: 48%) was slightly less than those for the category number (Condition 1: 51%).

Another way I conceptualized this interaction was by examining how the scoring approaches depended on measures. Results indicated that category scores were more influenced by measures than median and log median scores. More specifically, category number scores that were used with the five point scale led to more Type II errors relative to the other measures than median and log scores. For instance, the difference in Type II errors for category number scores applied to the five point scale versus the eleven point scale (Condition 1: 51% vs. 36%) was three times larger than the difference between these scales for median (34% vs. 29%) and log median scores (Condition 1: 48% vs. 43%). This general trend holds across all conditions for non-zero effects. The impact of the scale-by-scoring approach interaction was relative to the designated effect size (e.g., the discrepancies in Type II error rates for the "small" effect were reduced by roughly one-half). The proportion of significant effects and standardized regression coefficients from Tables 4 through 7 reiterated these findings except they had an inverse relationship with Type II errors. Higher Type II error rates corresponded with a lower proportion of significant effect and smaller standardized regression coefficients.

The meta-models did not find effects of scoring approaches, measures, or their interaction on the Type I error rate across the conditions (for all models p > .05). Tables 4 through 7 show that there were no clear differences in the proportion of Type I errors across the various measure-by-scoring approach combinations. Type I error rates were consistently around .03-.05 across conditions. I also tested whether the scale-by-scoring combination could potential induce a spurious interaction between x_1 and x_2. Results did not support the existence of this interaction in any conditions.

Hypothesis 2 and 3: General Effects of Scoring and Measures

Beyond the complexities addressed with the interactive effect of scoring approaches and measures, there were three general trends in the simulation results. First, median scores outperformed other scoring approaches with regard to Type II error rates and, by necessity, proportion of significant effects. For example, in condition 1, the percentage of significant effects for x_1 for the four scales were higher using median scores (8pt NIAAA 71%; 5pt scale 66%; 11pt scale 71%; 8pt log scale 69%) than log median and category number scores (8pt NIAAA 56%, 60%; 5pt scale 52%, .49%; 11pt scale 57%, 64%; 7pt log scale 56%, 59%). Second, as stated earlier, category number scores outperformed log median scores on all measures except the five point measure (Condition 1: Category Numbers Type II errors: 8pt NIAAA 40%; 11pt 36%; 7pt log 41%; Log Median Type II Errors: 8pt NIAAA 44%; 11pt 43%; 7pt log 44%). Third, the five point measure was consistently outperformed by the other measures (e.g., Condition 1: 5pt scale with log median scores Type II errors: 48%; other scales with log median scores: 43-44%). In sum, these results indicated that both measures and scoring approaches impacted statistical power.

Measures and Scores Leading to Different Patterns of Effects

I evaluated how scores interacted with measures to impact the validity of inferences by identifying whether different scores led to different patterns of effects using the same measure. For example, consider a case in which the eight point NIAAA measure with median scores was used and there was significant effect of x_1 and no significant effect of x_2. Then, given the same underlying data and eight point NIAAA measure, category scores were applied and there were *no* significant effects of x_1 or x_2. This was identified as a “different pattern of effects”. Table 8 shows that in condition 1 and 3, across the four measures between 19-24% had different patterns of effects caused by scoring approaches. In condition 2 and 4, across the measures 15-20% had different patterns of effect caused by scoring. These results indicated that different patterns of effects were often caused by scoring approaches. In all conditions, the five point scale had a slightly higher proportion of inconsistent patterns of effects compared to the other scales.

Similarly, I evaluated the effect of measures on the validity of inferences by identifying whether different measures led to different patterns of effects using the same scoring technique. For instance, consider a case in which the median scores were applied to the eight point NIAAA measures and there was significant effect of x_1 and a no significant effect of x_2. Then, given the same underlying data and median scores were applied to the five point measure, but there were *no* significant effects of x_1 or x_2. This was considered a “different pattern of effects”. I found that across all conditions, different patterns of effects occurred often when using the category number and median scoring approaches (see Table 9). Results indicated that about 17-18% of the models from conditions 1and 3 and about 18-20% of the models from conditions 2 and 4 led to different patterns of effects due to measures. Different patterns of effects due to measures occurred less frequently using the log transformation scoring approach (8%-9%).

In sum, this simulation study showed that there was an interaction between scores and measures that led to increased Type II errors. The five point scale depended on scoring in a fundamentally different way than the other scales. More specifically, the five point scale performed worst with category scores whereas this was not the case for the other scales. There was a general trend of median scores outperforming category scores and category scores outperforming log median scores. Additionally, the five point measure was consistently outperformed by the other measures across the scoring approaches. My simulation showed that measures and scores often created different patterns of effects. In sum, these results suggested that scales and scoring approaches likely impact adolescent substance researchers' ability to test theoretical models through the reduction of statistical power and changes in patterns of effects.

Empirical Demonstration

I used empirical data from the NLSY to further evaluate my hypothesis that there should be a scale-by-scoring approach interaction effect of the inferences drawn from linear statistical models. The omnibus Wald tests from GEE meta-models described earlier found significant scale-by-scoring approach interactions on the log odds of obtaining a significant effects for each of the four predictors (age: $\chi^2(6)=46.00$, minority: $\chi^2(6)=77.38$, $p<.0001$; maternal monitoring: $\chi^2(6)=20.46$, $p<.001$). The interaction effect for gender was significant, but its magnitude appeared smaller given the high statistical power ($\chi^2(6)=16.71$, $p<.05$). Since all of the interactions were significant, I examined potential differences across the measure-by-scoring approach combinations for all predictors.

Table 10 displays the proportion of significant effects, odds ratios for obtaining a significant effect for the scale-by-scoring combinations compared to the NB model fitted to the counts, and standardized regression coefficients. These outcomes were based on the empirical results, not the estimated GEE meta-models. For the age and minority effects, the scale-by-scoring approach interaction was driven by a dependency between the five point scale and median scores. For instance, across the eight point NIAAA, eleven point, and seven point log scales there was a general trend of median scores resulting in a smaller proportion of significant effects of age than category and log median scores (i.e., for the 11 point scale: log median 47% and category number 48% vs. median scores 39%). However, the five point scales had systematically lower proportion significant and the difference between the proportions for category and log median scores versus median scores was smaller (log median and category number 40% vs. median scores 37%).

For the minority effect, this interaction manifested itself in a slightly different way. Across the eight point NIAAA, five point scale, and seven point log measures, the log median and category scores had comparable proportions of significant effects (around 59% to 62%). However, using the five point scale, the proportion significant effects was higher using median scores compared to the other measures (five point scale 45% vs. other scales 39% to 42%). It was also interesting that many of the scale-by-score combinations had a higher proportion of significant effects compared to the negative binomial model fitted to the counts. For instance, log median scores applied to the eight point NIAAA measures had a substantially larger proportion of significant effects than the negative binomial model (62% vs. 52%). This scenario could not be more rigorously examined to determine how measures and scores impact Type I and II errors because the population generating model was unknown.

I was not able to make conclusive statements about the effects of scoring and scales on the gender predictor because of the small effect size (proportion significant 5-8% across measures-by-score combinations). The maternal monitoring predictor appeared to be more consistent with a main effect of scoring on the proportion significant effects. More specifically, across all measures, median scores had a smaller proportion of significant effects compared to log median and category number scales (39-40% vs. 48-52%). Setting aside the specific complexities addressed with the interactive effect of scoring approaches and measures, there was a general trend of median scores having a lower proportion of significant effects compared to other scoring approaches. Log median scores and category scores performed quite similarly across the predictors.

I examined how frequently different scoring approaches led to a different pattern of effects using the same alcohol frequency measure. This process was similar to that outlined in the simulation study. For instance, if the eight point NIAAA measure with median scores was used and there was significant effect of age and maternal monitoring but with category scores there was only a significant age effect, this was considered a “different pattern of effects”. Different scoring approaches applied to the same measures and data frequently caused different patterns of effects; 48% of the models using the eight point NIAAA measure, 41% of using the five point measure, 51% using the eleven point measure, and 45% of seven point log median measure. I also computed how frequently different measures led to a different pattern of effects using the same scoring approach. Results indicated that measures caused high proportion of different patterns of effects; 46% of category number scores, 48% of median scores, and 35% of log median. Taken together, these results showed that the patterns of significance for the covariates varied substantially depending on what measures and scoring approaches were used.

Summary of Results

In sum, both the simulation study and empirical evaluation showed that scales and measures interacted to impact inferences drawn from linear models. Results suggested that the five point scale were more sensitive to scoring than the other three scales. Also, in both the simulation study and empirical demonstration, the five point scale almost always had the lowest proportion of significant effects compared to the other scales. However, the patterns of Type II errors, proportions of significant effects, and standardized regression coefficients suggested that the general impact of scores were markedly different in the simulation study compared to the empirical demonstration. Most notably, median scores performed best in the in simulation study whereas they appeared to perform the worst in the empirical demonstration (e.g., lower proportion of significant effects, lower standardized regression coefficients). These findings highlighted that even though scores depend on measures, this dependency likely is not the same across research settings.

Discussion

I used a comprehensive simulation study and an empirical demonstration to test my set of theoretically generated research hypotheses concerning the effect of ordinal alcohol measures and ordinal scoring approaches on our ability to draw valid inferences from linear statistical models. The results of my project provided support for my three research hypotheses.

Results from the simulation study and empirical demonstration suggested that measures and scoring approaches interacted to influence inferences drawn from linear statistical models. In the simulation, category number scores performed worse using the five point scale with response categories defined to be dissimilar to a log transformation compared to if category numbers were applied to the other three scales. The five point scale also had a general tendency to produce lower proportions of significant effects and larger Type II error rates compared to the other scales. Results suggested that it is not only the number of response categories, but also how counts are binned into categories, that impacted the performance of alcohol frequency measures. Although median scores had the lowest proportion of Type II errors in all conditions of the simulation study, the empirical demonstration showed that median scores led to the smaller proportion of significant effects compared to the other scoring approaches. In both the simulation and empirical demonstration, applying different scoring approaches to the same underlying data frequently led to different patterns of effects. There was also evidence suggesting that using different scales with same underlying data and scoring approaches can often lead to different effects. I will briefly examine each of my research hypotheses.

Hypothesis 1: Scale-by-Scoring Approach Interaction

I hypothesized that ordinal alcohol measures and scoring approaches would have an interactive effect on the validity of results obtained from linear models. I predicted that the five point scale should be more influenced by scoring approaches than the other measures because there were fewer response categories and the categories were not closely reflective of the log transformation. Results from my project supported this hypothesis. In the simulation, I found that with the five point scale category number scores performed the worst whereas log median scores were the worst with the other scales (e.g., highest Type II error rates). Category number scores applied to the five point measure was by far the worst scale-by-scoring approach combination used in the simulation study. For example, in condition 1, 51% of the models fitted with category scores applied to the five point scale led to Type II errors compared to 29% using median scores applied to the 11 point measure. Results from the empirical demonstration indicated that this scale-by-scoring approach interaction may manifest itself in a slightly different way. For instance, the age effect showed that the discrepancy in the proportion of significant effects between category numbers and log median scores versus median scores was less for five point scale than for any of the other three scales.

Statistical theory explains this scale-by-scoring approach interaction effect. Scoring approaches are inherently dependent on the ordinal measures (e.g., category numbers are based on the ordinal bins of measures) and if an ordinal measure does not effectively transform the underlying alcohol frequency counts, the relationships among a set of covariates and outcome will not be effectively captured in standard linear models. Because the five point measure had fewer response categories that were poorly binned, the relationships among the covariates and the alcohol frequency outcome were likely not linearized as well as with the other three ordinal alcohol scales. Moreover, because category numbers scores are highly dependent on the category bin sizes, I expected that their performance would be more affected by the poorly defined response categories. This dependency between scoring and how alcohol frequency counts are binned into ordinal categories has not been identified previously in the adolescent alcohol use literature. To my knowledge, this is the first study to explicitly examine how ordinal scales and scoring approaches interact in linear statistical models.

Hypothesis 2: Effect of Scoring

Beyond the scale-by-scoring interaction effect, I hypothesized that category and log median scores should generally outperform median scores (except with the five point measure) because they are better suited for linearizing the relationships among the covariates and ordinal alcohol frequency outcome. I found that ordinal scoring approaches applied to binned counts have a large effect on our ability to draw valid inferences from linear statistical models. I was unable to define a clear optimal scoring approach that was robust across the simulation study and empirical demonstration. In the simulation study, the median scoring approach outperformed the category number and log median scoring approaches (e.g., higher proportion of significant effects, lower Type II error rates, and highest standardized regression coefficients) across the four conditions.

Even though I did not hypothesize that the median would outperform the other scoring approaches, there have been similar findings highlighted in the field of biostatistics with linear trend tests. For example, in assessing the relationship between maternal drinking and congenital malformation, Graubard and Korn (1987) showed that by using median scores there was a highly significant association whereas the ordinal category scores led to a non-significant association. Findings like this have caused biostatisticians to recommend assigning scores that are reflective of true distance between categories such as median values (Agresti, 2002). However, supporting my original hypothesis, the results from the empirical demonstration are in direct evidence of the median approach being outperformed by the category number and log median approaches (e.g., lower proportion of significant effects and smaller standardized regression coefficients). This trend cannot be verified because I did not know the population generating model for the empirical data. However, these findings have led me to believe that other characteristics of adolescent alcohol data not considered in this project such as different underlying distributions (e.g., zero-inflated negative binomial, negative binomial, and censored negative binomial) may interact with scoring approaches to affect the results produced in these models.

Another important finding in both the simulation study and empirical demonstration was that, given the same underlying count data and alcohol frequency measure, different scoring approaches frequently produced differing patterns of effects. This finding reaffirms related research on different patterns of statistical significance due to ordinal scoring from other fields such as Graubard and Korn (1987). Taken together, these findings showed that ordinal scoring methods play an integral role in testing theoretical models of adolescent substance use using linear models.

Hypothesis 3: Effect of Measures

I hypothesized that there should be a general tendency for the five point scale to perform worse than the other three scales. Results from this project showed that ordinal alcohol measures can impact the substantive inferences drawn from linear statistical models.

In my simulation study, all four of the alcohol frequency scales showed high Type II error rates across all conditions, indicating that standard alcohol measures failed to find effects that truly exist. However, the ordinal scales did not frequently lead to Type I errors. Although the differences among the scales were not always large, the five category scale consistently performed the poorest (e.g., lowest proportion of significant effects, highest Type II error rate, smallest standardized regression coefficients) while the other three alcohol frequency scales performed comparable across the four effect size conditions.

The poorer performance of the five category scale in relation to the other scales was expected for two reasons based on prior research (e.g., MacCallum, et. al, 2002; Taylor, West, & Aiken, 2006). First, the five point scale had the fewest response categories, which is associated with an assortment of negative statistical consequences such as reduced statistical power and effect sizes. Second, the five categories deviated from a viable transformation (e.g., log transformation) more than the other three measures. The impact of how counts were binned is consistent with what Generalized Linear Model theory suggested (e.g., count outcome are often modeled with a log link function; McCullagh, & Nelder, 1989). The seven point log, eight point NIAAA, and 11 point measures likely performed the same because they each had a moderate to large number of reasonably defined response categories (e.g., bin size increased as the response category got higher).

I also found that applying different scales to the same underlying count data frequently led to different patterns of effects. These findings should not be surprising because, in many ways, binning acts like a data transformation (albeit a poor and unsystematic one). Prior quantitative work has clearly illustrated that performing data transformations on outcome variables can easily impact statistical significance (Adams, 1991). My study was the first to generalize these findings to a process of binning underlying counts into ordinal categories in the context of adolescent alcohol use.

Implications and Recommendations for Applied Research

My findings have several implications for applied research. The results showed that measures and ordinal scoring approaches can sometimes interact in unpredictable ways to influence the researchers' ability to validly test substantive hypotheses. My study was the first to explicitly show that the performance of alcohol frequency measures depends not only on the number of categories, but also how the open ended alcohol frequency counts are binned. My study was also the first to empirically examine the strong influence of scoring on substantive findings. I clearly demonstrated that both measure and scoring approach can substantially lower statistical power. Equally concerning is the idea that a researcher can have one data set and apply multiple scoring approaches, yet come to completely different substantive interpretations. There was no conclusive evidence that measures and scores lead to increased Type I errors, but that does not mean it cannot happen in practice. For example, in the empirical demonstration, the minority effect actually had a higher proportion of significant effects using linear models fitted to ordinal data than the negative binomial model fitted to the counts. This trend suggested that elevated Type I errors may arise in practice.

Given these results, I expect that existing published and unpublished research likely has been affected by these differences in scores and measures. These factors may have caused researchers to fail to uncover or replicate true effects. Currently, few researchers report details on their ordinal alcohol measures or how they scored the ordinal data, which makes it difficult to evaluate the potential impact of these factors. Ordinal scores applied to binned counts presents an additional concern because researchers can collect alcohol frequency data on a single ordinal scale and knowingly, or unknowingly, change their findings solely because of how the alcohol frequency data are scored. This type of risk is typically nonexistent with ordinal alcohol measures because the categories are usually defined on a survey a priori.

Based on the findings of this study, I offer four recommendations for adolescent alcohol use researchers. First, it is vital that methodological work like I have conducted in this project be disseminated so that researchers can become aware that the quantitative characteristics of alcohol measures and ordinal scoring approaches can cause substantively different model results. Second, researchers should strongly consider collecting open-ended count data and modeling the counts with appropriate nonlinear models (e.g., Poisson, Negative Binomial, Zero-inflated regression models). Collecting open-ended count data may result in unreliable measurements because of errors in cognitive recall. However, I believe that benefits of being able to fit the appropriate GLM to the count data and bypassing the issues that I have identified with fitting standard linear models to ordinal data likely outweigh the potential costs. I recognize that this recommendation of collecting count data has yet to be rigorously evaluated and many applied researchers will likely continue to follow the standard practice of fitting linear models to ordinal data. My third recommendation for these researchers is to use an ordinal measure with at least seven categories that are defined to be similar to a log transformation (e.g., increasing bin sizes as categories increase). There is likely no added benefit to having an excessively large number of categories. In fact, having too many categories could make defining the response categories difficult and cause sparseness in the upper categories.

Finally, to minimize the negative impact of scoring approaches on linear models, I recommend doing a two-step sensitivity evaluation. First, researchers should perform a thorough exploration of their data using graphical and descriptive techniques with different ordinal scoring approaches. Descriptive explorations should involve examining the basic properties of the alcohol frequency data independently (e.g., mean, standard deviations, tests of distributions) and conditional on covariates (e.g., correlations, conditional means). Graphical explorations such as scatter plots fitted with smoothed curves and other visualization of functional form are essential to ensuring that ordinal scoring approaches are not inducing nonlinear relationships between the covariates and the outcome (median scores and scores based on poorly defined measures should be of greatest risk for this). Researchers should fit multiple models to different ordinal scores to assess sensitivity. If results are the same across these models, researchers can feel confident in their pattern of results under the assumption that all of the models are not consistent and wrong. If the results differ, researchers should refer back to the data exploration from step 1 and consider what pattern of effects is most consistent with substantive theory so that they can make the best decision possible. Researchers should always inform the reader of their scoring method, define the response categories of their measures, and note if their findings are sensitive to different scoring approaches.

Limitations and Future Directions

Clearly, the results from my study cannot be generalized to all cases. Several factors were not investigated in my simulation such as alternative population distributions (e.g., standard Poisson/Negative Binomial, censored Poisson/Negative Binomial, and Zero-inflated Poisson), alcohol frequency measures, and scoring approaches. My empirical demonstration showed the impact of measures and scores results using existing data, but there was no way to know which scoring method and measure was best because the population generating model was unknown. Findings from my project cannot be generalized directly to the longitudinal setting because scores and measures likely interact in more complicated ways to impact our ability to accurately test theoretical models of adolescent alcohol use over time. I did not include a condition where I fit the ordinal alcohol frequency data with ordinal statistical models in my project. This is not standard practice in applied research and I currently cannot comment on the performance of these types of statistical models for adolescent substance use data.

Future directions include extending this cross sectional work more broadly to other constructs that are captured through ordinal measures representing binned counts. For example, a few logical extensions would be to other drugs (e.g., marijuana, cigarettes, and cocaine), number of delinquent behaviors, number of depressive episodes, and number of stressful life events. Most substance use measures operate exactly the same as alcohol frequency measures by binning counts into a smaller number of ordinal categories. I expect that these findings will generalize well to other alcohol measures (e.g., frequency of binge drinking, quantity of use, frequency of drunkenness) and substances (e.g., marijuana, cocaine, prescription drugs), but this needs to be confirmed. It is possible that these various dimensions of substance use have completely different underlying distributions (e.g., Poisson, Negative Binomial, Zero-inflated), which may impact how ordinal measures and scoring approaches function in standard linear models. The current project should be extended longitudinally to help understand how these measures and scoring approaches operate in statistical tests of more complex theoretical models.

Possibly the most important direction is determining whether or not researchers should collect adolescent substance use data on an ordinal scale. Ordinal measures may help to eliminate some unreliability in participants' recall of the actual count of the number of days they have used alcohol, but it is unclear if this assumed improvement in reliability offsets the statistical consequences of fitting linear models to ordinal data (e.g., increased Type II errors). Future studies should clarify whether or not collecting more unreliable count data that can be modeled in appropriate nonlinear statistical models has added benefits over the current standard practices in measuring and model adolescent substance use data. Currently, the field of adolescent substance use has failed to fully capitalize on several recent advances in nonlinear statistical models for count data such as Poisson, Negative Binomial, and various techniques for Zero-inflated data. These novel nonlinear methods have the potential to improve our ability to draw accurate inferences from statistical models, while at the same time increase the breadth of hypotheses that can be formally tested compared to current practices. However, in order to capitalize on the flexibility of these innovative models, we must first justify that the collection of count data over more commonly collected ordinal data.

Conclusion

In sum, my project has added to the existing quantitative literature in areas such as coarse categorization and ordinal scoring approaches in three ways. First, several researchers have found that coarsely categorizing data leads to statistical consequences such as decreased power (MacCallum et. al, 2002; Taylor, West, & Aiken, 2006), but my study was the first to generalize this research to underlying counts binned into ordinal categories. Second, although prior research has shown that ordinal scoring approaches can make a substantial difference in the results obtained from linear statistical methods (e.g., Graubard and Korn , 1987), my study explicitly compared multiple methods and showed that scores depend on measures and likely other unknown factors (e.g., underlying distributions). Existing advice on ordinal scores from the field of biostatistics recommends using median scores (Agresti, 2002), but my study suggested that median scores may not always be the best choice for all research settings. Third, my study outlined how measures and scoring approaches can interact in complicated ways. For instance, category scores applied to the five point scale with poorly defined ordinal categories led to far more Type II errors than category scores applied to any of the other three measures. Moreover, pairing different score-by-measure combinations on the same underlying data will often lead to researcher drawing substantively different inferences from linear models. Taken together, these findings clearly showed that measures and ordinal scoring approaches have the ability to affect adolescent alcohol researchers' ability to build a cumulative science through rigorous tests of substantive theory.

Table 1.

Four alcohol frequency scales for a past 30 day time frame

How many days in the past 30 days have you had one or more drinks?			
8 Point NIAAA Scale	**5 Point Scale**	**11 Point Scale**	**7 Point Log Scale**
7. 28-30 days	4. 24-30 days	*10. 25-30 days*	*6. 19-30 days*
6. 18-27 days	3. 16-23 days	*9. 20-24 days*	*5. 11-18 days*
5. 10-17 days	2. 6-15 days	*8. 15-19 days*	*4. 7-10 days*
4. 6-9 days	1. 1-5 days	*7. 11-14 days*	*3. 4-6 days*
3. 4-5 days	0. 0 days	*6. 8-10 days*	*2. 2-3 days*
2. 2-3 days		*5. 6-7 days*	*1. 1 days*
1. 1 days		*4. 4-5 days*	*0. 0 days*
0. 0 days		*3. 3 days*	
		2. 2 days	
		1. 1 day	
		0. 0 days	

Table 2.

Experimental seven point measure based on the log transformation.

	C1	C2	C3		C4			C5				C6				
Days	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Log(day+1)	0	.70	1.10	1.39	1.61	1.79	1.95	2.08	2.20	2.30	2.40	2.48	2.56	2.63	2.71	2.77

	C6 (cont.)			C7											
Days	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Log(day+1)	2.83	2.89	2.94	3.00	3.04	3.09	3.14	3.18	3.22	3.26	3.30	3.33	3.37	3.40	3.43

Table 3.

Recovery of population generating values from simulation.

	Condition 1		**Condition 2**		**Condition 3**		**Condition 4**	
	Population Generating	Simulation Est.(se)	Population Generating	Simulation Est.(se)	Population Generating	Simulation Est.(se)	Population Generating	Simulation Est.(se)
α	1.25	1.23(.53)	1.25	1.25(.66)	1.25	1.24(.69)	1.25	1.30(.75)
ϑ	-0.30	-0.44(1.36)	-0.30	-0.75(3.00)	-0.30	-0.73(2.86)	-0.30	-1.09(3.91)
β_0 (intercept)	0.75	0.72(.22)	0.75	0.75(.25)	0.75	0.72(.27)	0.75	0.71(.28)
β_1 (x_1)	0.70	0.67(.25)	0.49	0.47(.25)	0.00	0.03(.26)	0.00	0.01(.25)
β_1 (x_2)	0.00	0.00(.12)	0.00	0.01(.11)	0.34	0.33(.12)	0.23	0.23(.12)

Note: There were 500 replications per condition

Table 4.

Simulation Condition 1: Proportion of significant effects, Type I and II Errors, ORs, and standardized regression coefficients

			x_1				x_2		
Condition	**Scale**	**Scoring**	**Proportion Sig.(sd)**	**Type II Errors (sd)**	**OR Type II Error**	**Std. β(sd)**	**Type I Errors(sd)**	**OR Type I Error**	**Std. β(sd)**
1 x_1: Medium Effect x_2: No Effect 500 Replications	8pt. NIAAA	Category #	0.60(.49)	0.40(.49)	2.22	0.15(.07)	0.05(.21)	0.96	0(.07)
		Median	0.71(.46)	0.29(.46)	1.41	0.16(.06)	0.05(.21)	1.00	0(.07)
		Log Median	0.56(.50)	0.44(.50)	2.70	0.14(.07)	0.05(.21)	1.00	0(.07)
	5pt. Scale	Category #	0.49(.50)	0.51(.50)	3.55	0.13(.07)	0.05(.22)	1.09	0(.07)
		Median	0.66(.47)	0.34(.47)	1.71	0.16(.06)	0.04(.20)	0.87	0(.07)
		Log Median	0.52(.50)	0.48(.50)	3.08	0.13(.07)	0.05(.22)	1.04	0(.07)
	11pt. Scale	Category #	0.64(.48)	0.36(.48)	1.90	0.15(.06)	0.05(.21)	1.00	0(.07)
		Median	0.71(.46)	0.29(.46)	1.40	0.16(.06)	0.05(.22)	1.04	0(.07)
		Log Median	0.57(.50)	0.43(.50)	2.55	0.14(.07)	0.05(.22)	1.09	0(.07)
	7pt. Log Scale	Category #	0.59(.49)	0.41(.49)	2.35	0.14(.07)	0.05(.21)	0.96	0(.07)
		Median	0.69(.46)	0.31(.46)	1.49	0.16(.06)	0.05(.21)	1.00	0(.07)
		Log Median	0.56(.50)	0.44(.50)	2.66	0.14(.07)	0.05(.22)	1.09	0(.07)
	True ZINB		0.77(.42)	0.23(.42)	-	-	0.05(.21)	-	-

Note: Odds ratios for Type I and II errors are in comparison to the true ZINB.

Table 5.

Simulation Condition 2: Proportion of significant effects, Type I and II errors, ORs, and standardized regression coefficients

			x_1				x_2		
Condition	**Scale**	**Scoring**	**Proportion Sig.(sd)**	**Type II Errors (sd)**	**OR Type II Error**	**Std. β(sd)**	**Type I Errors(sd)**	**OR Type I Error**	**Std. β(sd)**
2 x_1: Small Effect x_2: No Effect 500 Replications	8pt. NIAAA	Category #	0.33(.47)	0.67(.47)	1.82	0.10(.07)	0.06(.24)	1.30	0(.07)
		Median	0.39(.49)	0.61(.49)	1.41	0.12(.06)	0.05(.21)	0.96	0(.07)
		Log Median	0.30(.46)	0.70(.46)	2.05	0.10(.07)	0.07(.25)	1.39	0(.07)
	5pt Scale	Category #	0.28(.45)	0.72(.45)	2.26	0.09(.07)	0.07(.26)	1.47	0(.07)
		Median	0.35(.48)	0.65(.48)	1.62	0.11(.06)	0.05(.21)	0.96	0(.07)
		Log Median	0.31(.46)	0.69(.46)	1.96	0.09(.07)	0.07(.26)	1.47	0(.07)
	11pt Scale	Category #	0.35(.48)	0.65(.48)	1.65	0.11(.07)	0.06(.24)	1.30	0(.07)
		Median	0.40(.49)	0.60(.49)	1.35	0.12(.06)	0.05(.22)	1.00	0(.07)
		Log Median	0.30(.46)	0.70(.46)	2.03	0.10(.07)	0.07(.25)	1.39	0(.07)
	7pt Log Scale	Category #	0.34(.47)	0.66(.47)	1.75	0.10(.07)	0.06(.23)	1.13	0(.07)
		Median	0.39(.49)	0.61(.49)	1.41	0.12(.06)	0.05(.21)	0.96	0(.07)
		Log Median	0.31(.46)	0.69(.46)	1.96	0.10(.07)	0.06(.24)	1.26	0(.07)
	True ZINB		0.47(.50)	0.53(.50)	-	-	0.05(.22)	-	-

Note: Odds ratios for Type I and II errors are in comparison to the true ZINB

Table 6.

Simulation Condition 3: Proportion of significant effects, Type I and II errors, and standardized regression coefficients

Condition	Scale	Model	***x_1***			***x_2***			
			Type I Errors(sd)	**OR Type I Error**	**Std. β(sd)**	**Proportion Sig.(sd)**	**Type II Errors(sd)**	**OR Type I Error**	**Std. β(sd)**
3 x_1: No Effect x_2: Medium Effect 500 Replications	8pt. NIAAA	Category #	0.04(.20)	0.62	0(.07)	0.62(.48)	0.38(.48)	2.19	0.15(.07)
		Median	0.05(.21)	0.71	0(.07)	0.72(.45)	0.28(.45)	1.41	0.17(.07)
		Log Median	0.04(.21)	0.65	0(.07)	0.58(.49)	0.42(.49)	2.61	0.14(.07)
	5pt Scale	Category #	0.04(.21)	0.65	0(.07)	0.53(.50)	0.47(.50)	3.24	0.13(.07)
		Median	0.04(.20)	0.62	0(.07)	0.69(.46)	0.31(.46)	1.66	0.16(.07)
		Log Median	0.04(.20)	0.59	0(.07)	0.56(.50)	0.44(.50)	2.83	0.14(.07)
	11pt Scale	Category #	0.04(.20)	0.59	0(.07)	0.66(.47)	0.34(.47)	1.85	0.16(.07)
		Median	0.05(.21)	0.68	0(.07)	0.73(.44)	0.27(.44)	1.32	0.17(.07)
		Log Median	0.04(.21)	0.65	0(.07)	0.58(.49)	0.42(.49)	2.61	0.14(.07)
	7pt Log Scale	Category #	0.04(.20)	0.62	0(.07)	0.61(.49)	0.39(.49)	2.28	0.15(.07)
		Median	0.04(.21)	0.65	0(.07)	0.71(.45)	0.29(.45)	1.48	0.17(.07)
		Log Median	0.04(.20)	0.62	0(.07)	0.58(.49)	0.42(.49)	2.63	0.14(.07)
	True ZINB		0.07(.25)	-	-	0.78(.41)	0.22(.41)	-	. .

Note: Odds ratios for Type I and II errors are in comparison to the true ZINB

Table 7.
Simulation Condition 4: Proportion of significant effects, Type I and II errors, and standardized regression coefficients

			x_1			x_2			
Condition	**Scale**	**Model**	**Type I Errors(sd)**	**OR Type I Error**	**Std. β(sd)**	**Proportion Sig.(sd)**	**Type II Errors(sd)**	**OR Type II Error**	**Std. β(sd)**
4 x_1: No Effect x_2: Small Effect 500 Replications	8pt. NIAAA	Category #	0.05(.22)	1.00	0(.07)	0.38(.49)	0.62(.49)	1.58	0.11(.07)
		Median	0.05(.21)	0.88	0(.07)	0.45(.50)	0.55(.50)	1.21	0.12(.07)
		Log Median	0.05(.22)	1.00	0(.07)	0.35(.48)	0.65(.48)	1.81	0.10(.07)
	5pt Scale	Category #	0.06(.23)	1.08	0(.07)	0.31(.46)	0.69(.46)	2.23	0.10(.07)
		Median	0.05(.23)	1.04	0(.07)	0.40(.49)	0.60(.49)	1.50	0.12(.07)
		Log Median	0.06(.24)	1.16	0(.07)	0.34(.47)	0.66(.47)	1.91	0.10(.07)
	11pt Scale	Category #	0.05(.23)	1.04	0(.07)	0.41(.49)	0.59(.49)	1.43	0.11(.07)
		Median	0.05(.22)	0.96	0(.07)	0.45(.50)	0.55(.50)	1.20	0.12(.07)
		Log Median	0.05(.22)	1.00	0(.07)	0.36(.48)	0.64(.48)	1.75	0.11(.07)
	7pt Log Scale	Category #	0.06(.23)	1.08	0(.07)	0.37(.48)	0.63(.48)	1.66	0.11(.07)
		Median	0.05(.21)	0.92	0(.07)	0.44(.50)	0.56(.50)	1.24	0.12(.07)
		Log Median	0.05(.22)	0.96	0(.07)	0.35(.48)	0.65(.48)	1.83	0.10(.07)
	True ZINB		0.05(.22)	-	-	0.50(.50)	0.50(.50)	-	- -

Note: Odds ratios for Type I and II errors are in comparison to the true ZINB

Table 8.

Simulation: Percent of different patterns effect caused by scoring across measures

Condition	**Scale**	**% with Different Pattern of Effects**
1	8pt. NIAAA	20.4
	5pt.	23.4
	12pt.	19.2
	7pt. Log	19.6
2	8pt. NIAAA	18.4
	5pt.	19.8
	12pt.	19.4
	7pt. Log	16.0
3	8pt. NIAAA	22.8
	5pt.	23.8
	12pt.	21.6
	7pt. Log	21.4
4	8pt. NIAAA	16.8
	5pt.	19.4
	12pt.	16.8
	7pt. Log	15.2

Table 9.

Simulation: Percent of different patterns effect caused by scale across scoring approaches

Condition	Scoring Approach	% with Different Pattern of Effects
1	Category Number	18.2
	Median	17.0
	Log Median	8.8
2	Category Number	13.8
	Median	20.6
	Log Median	8.2
3	Category Number	19.0
	Median	18.2
	Log Median	8.0
4	Category Number	16.4
	Median	19.4
	Log Median	9.2

Table 10.

Empirical Demonstration: Proportion of significant effects, OR compared to NB model, and standardized regression coefficients

		Age Effect			**Minority**			**Male**			**Maternal Monitoring**		
Scale	**Model**	**Prop. Sig.(sd)**	**OR**	**Std. β(sd)**	**Prop. Sig.(sd)**	**OR**	**Std. β(sd)**	**Prop. Sig.(sd)**	**OR**	**Std. β(sd)**	**Prop. Sig.(sd)**	**OR**	**Std. β(sd)**
8pt. NIAAA	Category #	0.47(.50)	0.77	0.12(.06)	0.59(.49)	1.37	-0.13(.06)	0.08(.26)	0.95	-0.02(.06)	0.50(.50)	0.81	-0.13(.06)
	Median	0.40(.49)	0.57	0.11(.06)	0.42(.49)	0.67	-0.11(.06)	0.05(.22)	0.60	-0.01(.06)	0.39(.49)	0.51	-0.11(.06)
	Log Median	0.46(.50)	0.74	0.12(.06)	0.62(.49)	1.50	-0.14(.06)	0.07(.26)	0.90	-0.03(.06)	0.52(.50)	0.85	-0.13(.06)
5pt. Scale	Category #	0.40(.49)	0.57	0.11(.06)	0.59(.49)	1.35	-0.14(.06)	0.08(.27)	1.04	-0.03(.06)	0.50(.50)	0.81	-0.13(.06)
	Median	0.37(.48)	0.51	0.10(.06)	0.45(.50)	0.78	-0.11(.06)	0.06(.23)	0.68	-0.02(.06)	0.40(.49)	0.54	-0.11(.06)
	Log Median	0.40(.49)	0.57	0.11(.06)	0.59(.49)	1.36	-0.14(.06)	0.08(.28)	1.07	-0.03(.06)	0.51(.50)	0.83	-0.13(.06)
11pt. Scale	Category #	0.47(.50)	0.77	0.12(.06)	0.53(.50)	1.05	-0.13(.06)	0.06(.24)	0.78	-0.02(.06)	0.48(.50)	0.75	-0.12(.06)
	Median	0.39(.49)	0.55	0.11(.06)	0.39(.49)	0.59	-0.10(.06)	0.05(.22)	0.60	-0.01(.06)	0.39(.49)	0.51	-0.11(.07)
	Log Median	0.48(.50)	0.79	0.12(.06)	0.60(.49)	1.43	-0.14(.06)	0.08(.27)	0.99	-0.03(.06)	0.52(.50)	0.87	-0.13(.06)
7pt. Log Scale	Category #	0.47(.50)	0.76	0.12(.06)	0.60(.49)	1.41	-0.13(.06)	0.07(.25)	0.85	-0.02(.06)	0.50(.50)	0.81	-0.13(.06)
	Median	0.42(.49)	0.63	0.11(.06)	0.42(.49)	0.67	-0.11(.06)	0.05(.22)	0.64	-0.01(.06)	0.40(.49)	0.54	-0.11(.06)
	Log Median	0.46(.50)	0.73	0.12(.06)	0.62(.49)	1.50	-0.14(.06)	0.08(.27)	0.96	-0.03(.06)	0.52(.50)	0.88	-0.13(.06)
	NB	0.54(.50)	-	-	0.52(.50)	-	-	0.08(.27)	-	-	0.56(.50)	-	-

Note: Results are over 1,000 random sample of n=250. Odds ratios are in comparison to the negative binomial models

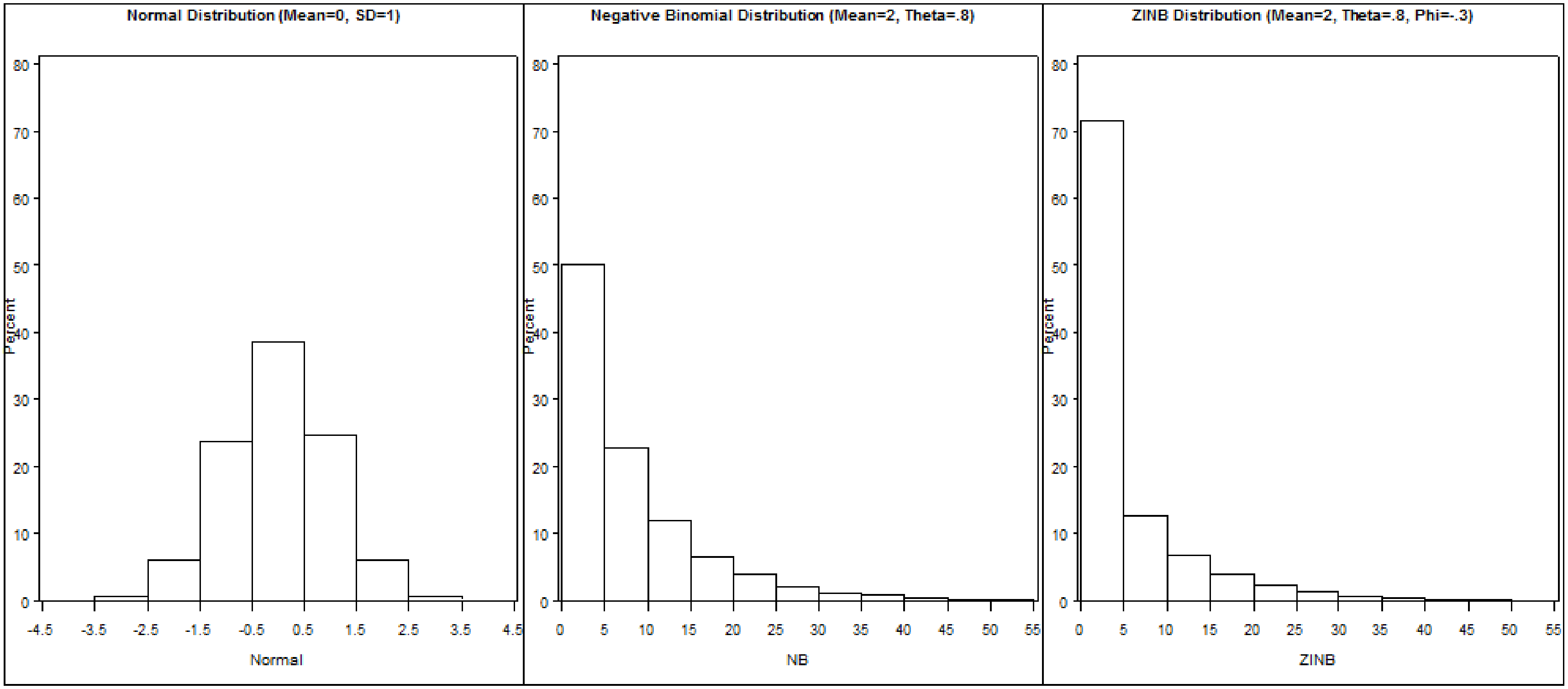

Figure 1. Histograms for normal, negative binomial, and zero-inflated negative binomial distribution

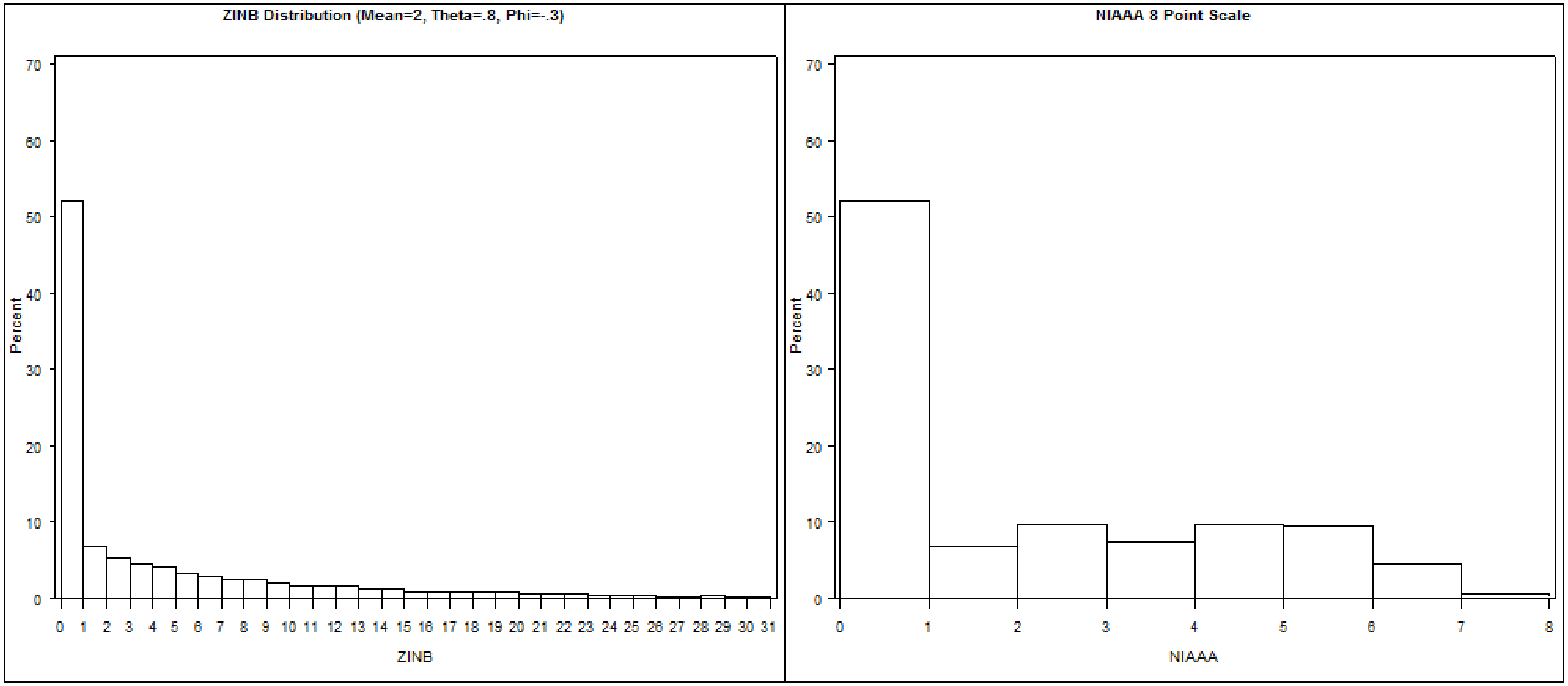

Figure 2. Histograms for ZINB counts and NIAAA ordinal scale. Counts above 30 were truncated for this illustration (~1.5%)

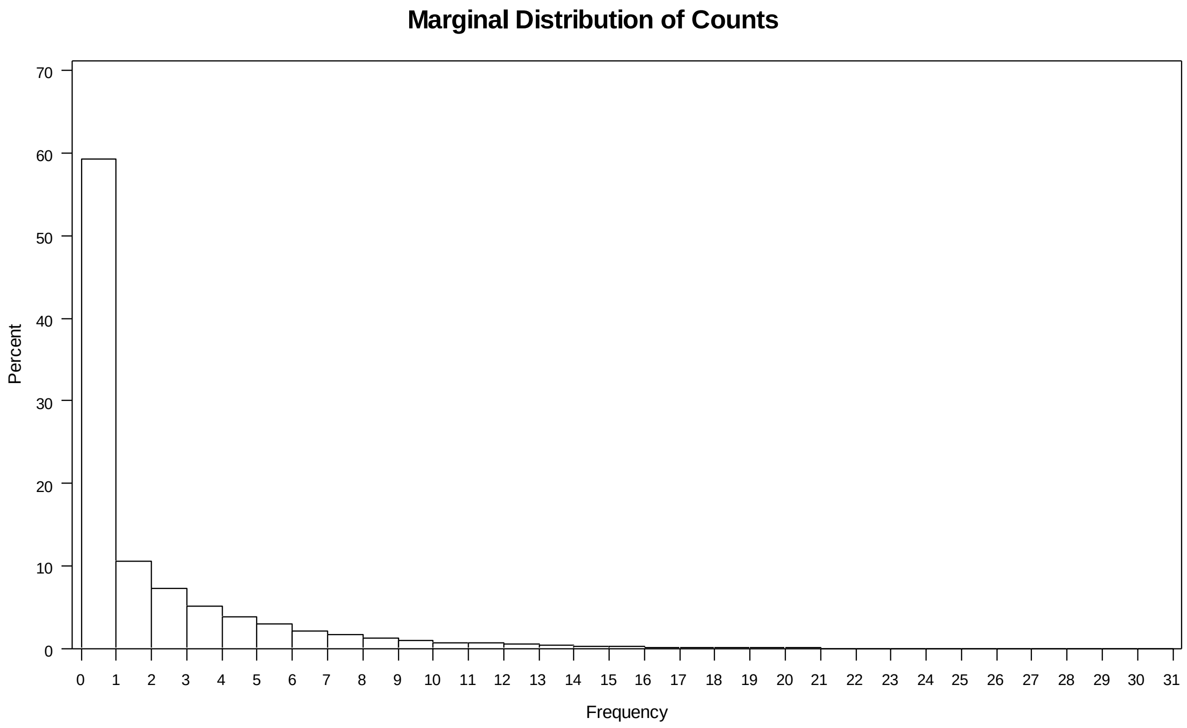

Figure 3. Simulation Study, Condition 1: Marginal distribution of counts.

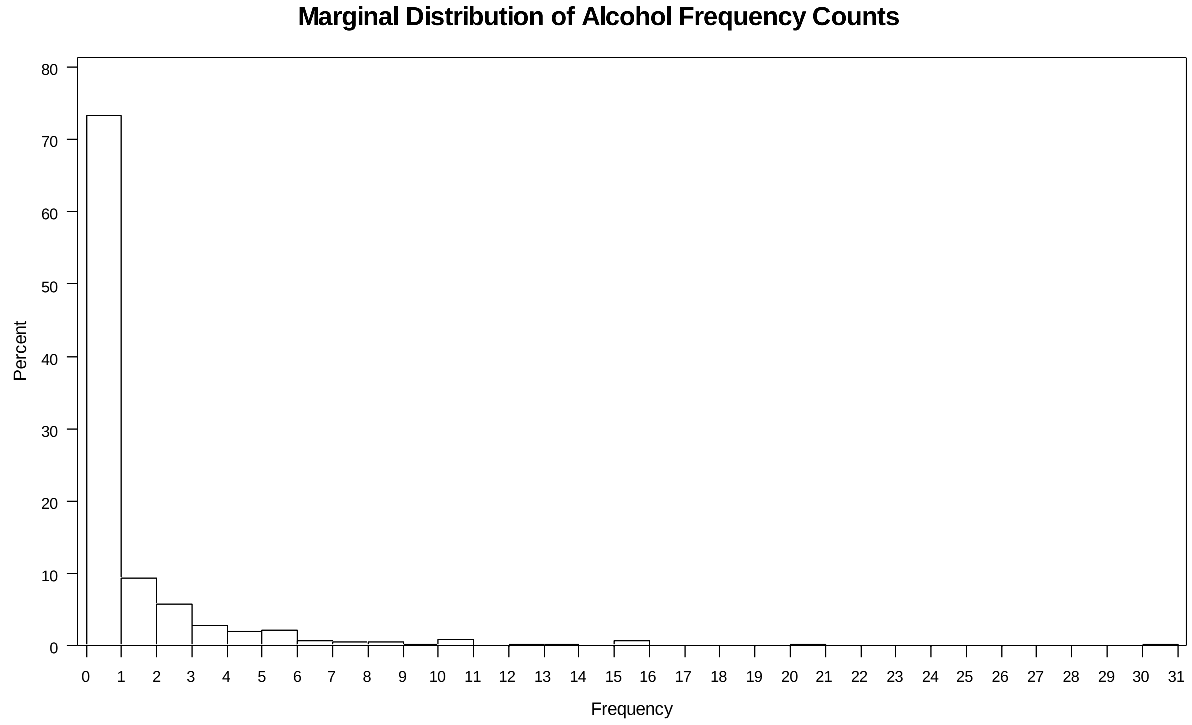

Figure 4. Empirical Demonstration: Marginal distribution of alcohol frequency counts

REFERENCES

Adams, J. L. (1991). A computer experiment to evaluate regression strategies. *Proceedings of the American Statistical Association, Section of Statistical Computing.* 55-62.

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

Chassin, L., Hussong, A., & Beltran, I. (2009). *Adolescent substance use*. In R. Lerner & L. Steinberg (Eds.), Handbook of adolescent psychology (Vol. 1, pp. 723–763). Hoboken, NJ: Wiley.

Chassin, L., Pillow, D. R., Curran, P. J., Molina, B. S. G., & Barrera, M., Jr. (1993). Relation of parental alcoholism to early adolescent substance use: A test of three mediating mechanisms. *Journal of Abnormal Psychology, 102*, 3–19.

Chassin, L., Presson, C. C., Pitts, S., & Sherman, S. J. (2000). The natural history of cigarette smoking from adolescence to adulthood in a midwestern community sample: Multiple trajectories and their psychosocial correlates. *Health Psychology*, *19,* 223–231.

Cooper, M. L., Frone, M. R., Russell, M., & Mudar, P. (1995). Drinking to regulate positive and negative emotions: A motivational model of alcohol use. *Journal of Personality and Social Psychology, 69*, 990–1005.

Curran, P. J., & Willoughby, M. J. (2003). Reconciling theoretical and statistical models of developmental processes. *Development and Psychopathology, 15*, 581-612.

Dawson, D. A., & Room, R. (2000). Towards agreement on ways to measure and report drinking patterns and alcohol–related problems in adult general population surveys: The Skarpo Conference overview. *Journal of Substance Abuse 12*, 1–21.

Dogan, S. J., Stockdale, G. D., Widaman, K. F., & Conger, R. D.(2010). Developmental relation and patterns of change between alcohol use and number of sexual partners from adolescence through adulthood. *Developmental Psychology, 46*(6), 1747-1759.

Emanuele, M. A., Wezeman, F., & Emanuele, N. V. (2002). Alcohol's effects on female reproductive function. *Alcohol Health and Research World, 26*, 274–281.

Graubard, B. I. & Korn, E. L. (1987). Choice of column scores for testing independence in ordered 2 × K contingency tables. *Biometrics, 43*, 471–476.

Hallfors, D. D., Waller , M. W. , Bauer , D. , Ford, C. A. , & Halpern, C. T. (2005). Which comes first in adolescence - sex and drugs or depression? *American Journal of Preventive Medicine, 29*(3), 163 – 170.

Hussong, A., & Chassin, L. (2002). Parent alcoholism and the leaving home transition. *Development and Psychopathology, 14*, 139–157.

Hussong, A. M., Curran, P.J., & Chassin, L. (1998). Pathways of risk for children of alcoholics' accelerated heavy alcohol use. *Journal of Abnormal Child Psychology, 26,* 453-466.

Ivis, F. J. Bondy, S. J. & Adlaf, E. M. (1997). Effect of question structure on self–reports of heavy drinking: Closed–ended versus open–ended questions. *Journal of Studies on Alcohol 58*, 622–624.

Johnston, L. D., O'Malley, P. M., Bachman, J. G., & Schulenberg, J. E. (2009). *Monitoring the future national survey results on drug use, 1975-2008. Volume I: Secondary school students* (NIH Publication No. 09-7402). Bethesda, MD: National Institute on Drug Abuse.

Kuntsche, E., Kuntsche, S., Knibbe, R., Simons-Morton, B., Farhat, T., Hublet, A., Bendtsen, P., Godeau, E., & Demetrovics, Z. (2011). Cultural and gender convergence in adolescent drunkenness. *Archives of Pediatrics & Adolescent Medicine, 165*(2), 152-158.

Leccese, M. and Waldron, H. B. (1994). Assessing adolescent substance use: A critique of current measurement instruments. *Journal of Substance Abuse Treatment 11*, 553–563.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables.* Thousand Oaks, CA: Sage Publications.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*, 19-40.

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models: Second edition*. London: Chapman and Hall.

Miller, W. R., & Del Boca, F. K. (1994). Measurement of drinking behavior using the Form 90 family of instruments. *Journal of Studies on Alcohol 12,* 112–118.

Miller, T. R., Levy, D. T., Spicer, R. S., & Taylor, D. M. (2006). Societal costs of underage drinking. *Journal of Studies on Alcohol, 67*(4), 519–528.

National Institute on Alcohol Abuse and Alcoholism (2003). Task force recommended alcohol questions: National council on alcohol abuse and alcoholism recommended sets of alcohol consumptions questions. Retrieved from Web December 1, 2010. http://www.niaaa.nih.gov/Resources/ResearchResources/TaskForce.htm.

O'Hara, R. B., & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution, 1,* 118–122.

Patrick, M. E., & Schulenberg, J. E. (2011). How trajectories of reasons for alcohol use relate to trajectories of binge drinking: National longitudinal data spanning late adolescence to early adulthood. *Developmental Psychology 47*(2), 311-317.

Rice E., Milburn NG, Monro, W. (2011) Social networking technology, social network composition, and reductions in substance use among homeless adolescents. *Prevention Science*, *12*(1), 80-88.

Scheier, L. M. (2010). *Handbook of drug use etiology: Theory, methods, and empirical findings*. Washington DC: American Psychological Association.

Schulenberg, J. E., & Maslowsky, J. (2009). Taking substance use and development seriously: Developmentally distal and proximal influences on adolescent drug use. *Monographs of the Society for Research in Child Development, 74,* 121–130.

Skinner, H. A. & Allen, B. A. (1982). Alcohol dependence syndrome: Measurement and validation. *Journal of Abnormal Psychology, 91*, 199-209.

Skinner, H. A., & Sheu, W. J. (1982). Reliability of alcohol use indices: The Lifetime Drinking History and the MAST. *Journal of Studies on Alcohol, 43,* 1157–1170.

Sobell, L. C., & Sobell, M. B. (1995). Alcohol consumption measures. In J. P. Allen, & M. Columbus (Eds.), *Assessing Alcohol Problems: A Guide for Clinicians and Researchers.*(pp. 55-73). Bethesda, MD: National Institute on Alcohol Abuse and Alcoholism,

Sobell, L. C., & Sobell, M. B. (2000). Alcohol Timeline Followback (TLFB). In *Handbook of Psychiatric Measures* (pp. 477-479). Washington, DC: American Psychiatric Association.

Tapert, S. F., Caldwell, L., & Burke, C (2005). Alcohol and the adolescent brain: Human studies. *Alcohol Research & Health*, *28*(4), 205–212.

Taylor, A., West, S., & Aiken, L. (2006). Loss of power in logistic, ordinal logistics, and probit regression when an outcome is coarsely categorized. *Educational and Psychological Measurement*, *66*, 228-239.

Trim, R. S., Meehan, B. T., King, K. M., & Chassin, L. (2007). The relation between adolescent substance use and young adult internalizing symptoms: Findings from a high-risk longitudinal sample. *Psychology of Addictive Behaviors, 21*(1), 97–107.

U.S. Department of Health and Human Services [USDHHS] (2007). *The surgeon general's call to action to prevent and reduce underage drinking*. Washington, DC: USDHHS, Office of the Surgeon General.