DOI 10.3758/s13428-013-0402-6

A latent variable model approach to estimating systematic bias in the oversampling method

Katherina K. Hauner · Richard E. Zinbarg · William Revelle

Published online: 19 October 2013

Abstract The method of oversampling data from a preselected range of a variable's distribution is often applied by researchers who wish to study rare outcomes without substantially increasing sample size. Despite frequent use, however, it is not known whether this method introduces statistical bias due to disproportionate representation of a particular range of data. The present study employed simulated data sets to examine how oversampling introduces systematic bias in effect size estimates (of the relationship between oversampled predictor variables and the outcome variable), as compared with estimates based on a random sample. In general, results indicated that increased oversampling was associated with a decrease in the absolute value of effect size estimates. Critically, however, the actual magnitude of this decrease in effect size estimates was nominal. This finding thus provides the first evidence that the use of the oversampling method does not systematically bias results to a degree that would typically impact results in behavioral research. Examining the effect of sample size on oversampling yielded an additional important finding: For smaller samples, the use of oversampling may be necessary to avoid spuriously inflated effect sizes, which can arise when the number of predictor variables and rare outcomes is comparable.

Keywords Sampling · statistical bias · latent variable modeling

Electronic supplementary material The online version of this article (doi:10.3758/s13428-013-0402-6) contains supplementary material, which is available to authorized users.

K. K. Hauner · R. E. Zinbarg · W. Revelle
Department of Psychology, Northwestern University,
Evanston, IL, USA

K. K. Hauner (✉)
Department of Neurology, Northwestern University, Ward Building
10-185 303 E. Chicago Avenue, Chicago, IL 60611, USA
e-mail: hauner@u.northwestern.edu

R. E. Zinbarg
The Family Institute at Northwestern University, Evanston, IL, USA

Sample size remains a ubiquitous hurdle in the study of rare outcomes. Without an adequate number of participants, we cannot perform meaningful analyses from which to draw conclusive inferences. This is especially the case in longitudinal studies that aim to capture significant predictors of outcomes that are rare or difficult to detect. In such studies, garnering a large sample of participants is critical, in order to increase the probability of securing a sufficient number of cases that the study aims to predict. However, obtaining large samples generally presents practical and financial encumbrances (e.g., Allison, Allison, Faith, Paultre, & Pi-Sunyer, 1997), and even more so when the variables of interest are rare in the general population. Examples of research areas for which this issue is particularly relevant include the study of psychiatric and neurological disorders, given that even the most "common" psychiatric and neurological conditions have a low prevalence in the general population (e.g., Hirtz et al., 2007; Kessler et al., 2005). Full-range random sampling, therefore, is not always an efficient methodological choice.

Sampling approaches

Extreme groups approach

One solution to this problem involves restricting study inclusion criteria to participants with extremely high or extremely low scores on a preselected measure of interest. This method, termed the *extreme groups approach* (EGA), diminishes sample size (as compared with a random sampling method), since data are selected only on the basis of extreme values within a sample distribution. Thus, the predictor variable consists of two "extreme groups," while the criterion variable is not directly manipulated.

Some have suggested that EGA may be a more powerful—and thus, more efficient—method for testing hypotheses than the method of random sampling (e.g., Abrahams & Alf, 1978; Alf & Abrahams, 1975; Borich & Godbout, 1974; Preacher, Rucker, MacCallum, & Nicewander, 2005). However, other researchers have exhorted against the use of EGA for numerous reasons (e.g., Preacher et al., 2005). First, discarding all values from the middle of the distribution necessarily distorts the distribution of the data (Cohen, 1983; MacCallum, Zhang, Preacher, & Rucker, 2002). Second, Preacher et al. demonstrated that the use of EGA results in an artifactual inflation of reliability. Third, several researchers have shown that EGA is associated with a bias toward increased effect sizes (Humphreys, 1985; McClelland & Judd, 1993; Preacher et al., 2005). In a recent analysis, Preacher and colleagues estimated the magnitude of standardized effect size inflation in the range of approximately 0.0–0.13, depending on the size of the population correlation and on the proportion of data sampled (i.e., tertile or quartile split). These researchers have agreed that effect sizes associated with EGA are a misrepresentation of the true effect sizes in the population. Interestingly, in the original analysis of EGA (Feldt, 1961), Feldt himself argued that this approach should be applied only to detect the presence of a linear relationship between two variables (both conforming to a bivariate normal distribution), rather than to evaluate the size of their relationship. Feldt specifically cautioned that the use of EGA to evaluate the size of a relationship between two minimally associated variables could lead to spuriously inflated estimates.

Oversampling

An alternative to random sampling that is more conservative than EGA involves the disproportionate gathering of data from a particular range of the predictor variable's distribution. This method, termed *oversampling*, allows for an efficient and practical data-gathering process, allowing researchers to focus on the range of data most likely to be associated with the rare outcome. Like EGA, data within the critical range of interest occur more frequently than they would in a random sample. However, unlike EGA, the remaining values are not limited to any particular slice of the distribution, thus retaining a representation of the entire distribution. Most typically, this range of oversampled data consists of very high or very low values of the predictor variable.

In longitudinal studies, for example, oversampling would increase the sample size of individuals at significant risk for the outcome of interest, providing a greater probability for its development. Importantly, increasing the number of cases of the predicted outcome could provide two methodological benefits. First, it could inspire greater generalizability and confidence in the results, since it would increase the likelihood that the individuals in the sample who developed the predicted outcome were representative of the population of individuals with this outcome. Second, for studies involving more than one predictor variable, it would decrease the probability for the number of predictor variables to equal (or exceed) the number of predicted outcomes. An example of *overfitting*, this substantial statistical issue can lead to spuriously inflated estimates of the relationship between the risk factor(s) and the outcome and will be addressed in further detail below.

Specific aims of the present study

Although there are obvious practical benefits to applying the method of oversampling, potential statistical costs of this approach have not been evaluated. That is, the question of whether there is systematic bias in testing an oversampled predictor has not yet been examined. Given the number of researchers who have reported that the use of EGA may result in biased effect sizes (Humphreys, 1985; McClelland & Judd, 1993; Preacher et al., 2005), it is imperative to evaluate the possibility that oversampling may misrepresent effect sizes as well, due to its disproportionate representation of the original range of data.

It should be noted that, although the issue of disproportional sampling has been addressed by the use of sampling weights, sample weighting is a controversial practice that has been heavily criticized as limited in both applicability and interpretability (Gelman, 2007). Sampling weights, which are intended to correct for disproportionate representation of a range of data, are derived from estimations of the extent of disproportionate representation. However, some researchers have argued that not only do these estimations require arbitrary statistical choices, but also they present problems when applied to particular parameter estimates (such as regression coefficients; DuMouchel & Duncan, 1982) and render standard errors uninterpretable (Gelman, 2007). Thus, the purpose of the present investigation was to determine the extent to which the method of oversampling, without the use of sampling weights, may affect estimates of effect size, as compared with results derived from a random sampling design.

In addition, it was critical to examine not only variables that were *directly* oversampled, but also those that were *indirectly* oversampled. This is because, in any multivariate design containing only one oversampled predictor, additional predictor variables that are correlated with the oversampled predictor will have been "indirectly oversampled" as a by-product of their correlation with the directly oversampled predictor variable. Therefore, an added purpose of the present study was to determine whether the oversampling method also affects estimates of effect size in indirectly oversampled variables, as compared with results derived from a random sampling design.

To address its aims, the present study employed simulated data. The following manipulations were included: (1) the proportion of data oversampled (scores were weighted by a factor that represented the probability of overselection); (2) the zero-order correlations between the criterion variable and the oversampled predictor variables, including both the *directly oversampled variable* (DOV) and the *indirectly oversampled variables* (IOV); and (3) the number of participants in the sample. In order for the simulated data set to be representative of a wider range of experimental data sets (i.e., those with measurement error and those including nonlinear relationships), two additional manipulations included (4) the amount of measurement error ($1 - \rho^2_{X,X}$) in the oversampled predictor variables and the criterion variable (to facilitate comparisons between a model with no measurement error and a model with measurement error) and (5) the shape of the relationship between the oversampled variables and the criterion variable (thus analyzing linear, as well as various curvilinear, relationships).

Method

The present study employed simulated data sets, comprising four oversampled predictor variables (DOV, IOV3, IOV2, IOV1) and a criterion variable (DV), as illustrated in Fig. 1. To account for varying levels of measurement error (as described above), data were generated using one of two latent variable models, one under the condition of no measurement error and one under the condition of measurement error (as

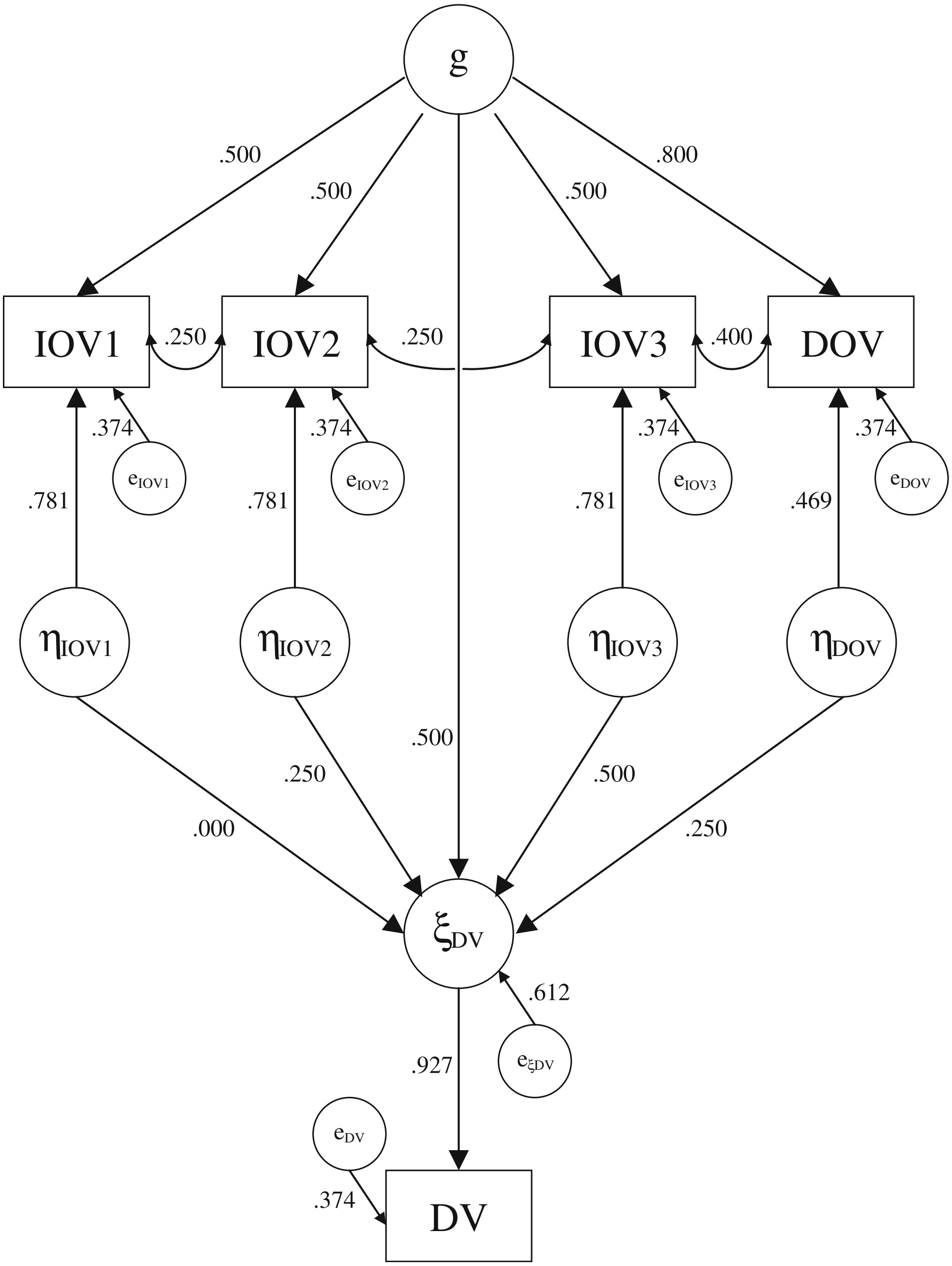

Fig. 1 Latent variable model (model 1) representing relationships between oversampled predictor variables and the categorical dependent variable (DV), in which measurement error was eliminated by setting reliabilities for all variables at 1.000. As in model 2, model 1 incorporated four oversampled predictor variables: a directly oversampled variable (DOV) and three indirectly oversampled variables (IOVs). Correlations between the DV and oversampled predictor variables were as follows: $r_{\mathrm{DOV,DV}} = .558$, $r_{\mathrm{IOV3,DV}} = .691$, $r_{\mathrm{IOV2,DV}} = .480$, and $r_{\mathrm{IOV1,DV}} = .270$

will be described in further detail below). From each of these two models, a constant sample size of 10,000 samples of raw data per cell was generated via R software (R Development Core Team, 2007), thus providing the population from which oversampled data would be selected.

Latent variable models were used to generate the simulated data sets, in order to account for all possible sources of variance (i.e., observable variables, latent variables, error). As is illustrated in Fig. 1, all four oversampled variables (DOV, IOV3, IOV2, IOV1) were influenced by a general factor (g), their unique error (e_{DOV}, e_{IOV3}, e_{IOV2}, e_{IOV1}), and their corresponding, unique latent variable (η_{DOV}, η_{IOV3}, η_{IOV2}, η_{IOV1}). Latent independent variables all influenced the latent dependent variable (ξ_{DV}), which was itself influenced by its own unique error ($e_{\xi\mathrm{DV}}$). Finally, the observable dependent variable (DV) was influenced by its own unique error (e_{DV}), as well as the latent DV (ξ_{DV}). Importantly, in order for the DV to represent a rare, categorical outcome (i.e., the circumstances under which oversampling is typically applied, such as the prediction of disorder onset), the DV was set at “1” for the upper 5 % of the distribution (representing “cases”) and at “0” for the remaining 95 % (representing “no cases”). The cutoff value of 5 % was selected due to its representativeness among reported prevalence rates of psychiatric and neurological disorders (e.g., Hirtz et al., 2007; Kessler et al., 2005)—areas of research for which the use of oversampling is standard.

For each of the two models (one with measurement error and one without), two consecutive sets of conventional regressions were performed. The first set of regressions estimated the effect of each oversampling variable (DOV, IOV3, IOV2, IOV1) on the DV. Importantly, regression coefficients were calculated to include the following conditions (as described in detail further below): two levels of *sample size* (i.e., the number of participants selected for oversampling), three levels of the *proportion of oversampled data* (i.e., the proportion of original data selected for oversampling), and four levels of *relationship shape* (between the oversampled predictor variable and the DV). Oversampled data were thus selected via a 2 (sample size) × 3 (proportion of oversampled data) × 4 (relationship shape) manipulation, with a total of 24 conditions. Regression coefficients for this first regression were calculated separately for each of the 24 conditions for 1,000 simulations, resulting in 24,000 sets of regression coefficients, each of which included four regression coefficients ($b_{\mathrm{DV\sim DOV}}$, $b_{\mathrm{DV\sim IOV3}}$, $b_{\mathrm{DV\sim IOV2}}$, $b_{\mathrm{DV\sim IOV1}}$). Because the present study used a categorical DV rather than a continuous DV (to better approximate a rare, clinical outcome), this first regression was necessarily logistic, thus yielding regression coefficients in units of unstandardized beta (Menard, 2004).

The second set of regressions consisted of four linear multiple regressions, in which the criterion variable was represented by one of the four coefficients from the previous set of logistic regressions ($b_{\mathrm{DV\sim DOV}}$, $b_{\mathrm{DV\sim IOV3}}$, $b_{\mathrm{DV\sim IOV2}}$, $b_{\mathrm{DV\sim IOV1}}$). Predictor variables in this second set of regressions included measures of oversampling (including interactions of oversampling with sample size and relationship shape, as listed in Table 1)—thus providing the final estimates of the effect of oversampling on effect size. For this second set of (linear) regressions, all predictor variables were centered, and resulting standardized regression coefficients were reported.

Description of manipulated variables

Amount of measurement error

The amount of measurement error was manipulated in order to determine whether effects found under the condition of perfect measurement would hold under more generalizable conditions. Thus, two levels of measurement error were chosen, to reflect either (1) a condition of no measurement error in the predictor and outcome variables ($\rho_{\mathrm{X,X}}$ = 1.000, indicating perfect reliability), or (2) a condition with a realistic level of measurement error in the predictor and outcome variables ($\rho_{\mathrm{X,X}}$ = .860). This “realistic” reliability level of .860 was determined by examining recent longitudinal research in which oversampling has been employed and adopting the most representative level of reliability reported across these studies (Zinbarg et al., 2010). The two latent variable models used to test these differences in measurement error are illustrated in model 1 ($\rho_{\mathrm{X,X}}$ = 1.000; Fig. 1) and model 2 ($\rho_{\mathrm{X,X}}$ = .860; Fig. 2).

Proportion of oversampled data

The proportion of data selected for oversampling was calculated as a ratio of the likelihood of being sampled for the oversampled quartile, relative to the likelihood of being sampled for nonoversampled quartiles. Quartiles were chosen for the present analyses, due to the prevalent use of oversampling quartiles in research on psychiatric and neurological disorders (e.g., Alloy et al., 2006; Costello et al., 1996). A range of probability weights (i.e., 1, 4, 8) was chosen to refine the detection of the shape of the relationship between oversampling probability and correlation of oversampled variables with criterion variables.

For the present study, a value from the upper quartile of the original population that had been given a probability weight of 1 had a 1/4 likelihood of being selected for the sample—thus representing a randomly sampled value. Likewise, a value from the upper quartile that had been given a probability weight of 4 had a 4/7 likelihood of being selected for the sample (with values in the other three quartiles having a 1/7 likelihood of being sampled). The probability weights of 4 and 8 were intended to follow traditional splits (Alloy et al., 2006; Costello et al., 1996).

Table 1 Summary of regression analyses for oversampling (predictor variables) and effect size estimates (criterion variable)

Predictor	*DV~DOV* ($\rho_{DOV,DV}$ = *.480)*	
	ß	*SE*
oversampling	-6.72E-03	5.67E-04
oversampling^2	6.73E-04	2.80E-04
(oversampling) * (sample size)	1.66E-05	4.54E-06
(oversampling) * (sample size) * (accelerating curve)	-1.62E-05	6.42E-06
(oversampling) * (sample size) * (logistic curve)	-1.66E-05	6.42E-06
(oversampling) * (sample size) * (decelerating curve)	-1.51E-05	6.42E-06
	DV~IOV3 ($\rho_{IOV3,DV}$ = *.594)*	
	ß	*SE*
oversampling	-2.33E-02	1.84E-03
oversampling^2	5.39E-03	9.06E-04
(oversampling) * (sample size)	7.58E-05	1.47E-05
(oversampling) * (sample size) * (accelerating curve)	-7.69E-05	2.08E-05
(oversampling) * (sample size) * (logistic curve)	-7.64E-05	2.08E-05
(oversampling) * (sample size) * (decelerating curve)	-7.75E-05	2.08E-05
	DV~IOV2 ($\rho_{IOV2,DV}$ = *.413)*	
	ß	*SE*
oversampling	-1.08E-02	7.11E-04
oversampling^2	2.04E-03	3.51E-04
(oversampling) * (sample size)	3.49E-05	5.69E-06
(oversampling) * (sample size) * (accelerating curve)	-3.54E-05	8.05E-06
(oversampling) * (sample size) * (logistic curve)	-3.45E-05	8.05E-06
(oversampling) * (sample size) * (decelerating curve)	-3.73E-05	8.05E-06
	DV~IOV1 ($\rho_{IOV1,DV}$ = *.232)*	
	ß	*SE*
oversampling	-2.58E-03	1.42E-03
oversampling^2	6.77E-04	7.02E-04
(oversampling) * (sample size)	2.62E-05	1.14E-05
(oversampling) * (sample size) * (accelerating curve)	-2.56E-05	1.61E-05
(oversampling) * (sample size) * (logistic curve)	-2.71E-05	1.61E-05
(oversampling) * (sample size) * (decelerating curve)	-2.40E-05	1.61E-05

Note. All predictor variables were centered prior to analysis. The four criterion variables are noted in the right column as DV~OV, and were derived from previous analyses in which oversampled variables predicted the dependent variable, based on reliability set at $\rho_{X,X}$ = .860. Results are reported in units of standardized effect size (*ß*).

Population correlation between the criterion variable and oversampled variables

There were several important considerations regarding the association between the categorical dependent variable and the oversampled predictor variables (i.e., $\rho_{DOV,DV}$, $\rho_{IOV3,DV}$, $\rho_{IOV2,DV}$, $\rho_{IOV1,DV}$). First, although an indirectly oversampled variable (IOV) would technically be defined as any predictor variable exhibiting a nonzero correlation with the DV, values representing typical multivariate analyses were used in the present study (range = .232 [$\rho_{IOV1,DV}$ with reliability at .860] −.691 [$\rho_{IOV3,DV}$ with reliability at 1.000]). Multiple IOVs were considered because including only one IOV would not have provided informative results regarding the differences between the IOV and the DOV, since the use of a single IOV would have resulted in identical standard errors for the IOV and DOV (on the basis of analogies to the equation for standard error in a multiple regression). The number of IOVs was set at three in order to represent weak, moderate, and strong correlations between the IOV and the DV. Note that the four population correlations differed between model 1 (which did not include measurement error) and model 2 (which included measurement error). For model 1 ($\rho_{X,X}$ = 1.000; see Fig. 1), population correlations were set as follows: $\rho_{DV,DOV}$ = .558, $\rho_{DV,IOV3}$ = .691, $\rho_{DV,IOV2}$ = .480, and $\rho_{DV,IOV1}$ = .270. For model 2 ($\rho_{X,X}$ = .860; see Fig. 2), population correlations were $\rho_{DV,DOV}$ = .480, $\rho_{DV,IOV3}$ = .594, $\rho_{DV,IOV2}$ = .413, and $\rho_{DV,IOV1}$ = .232. All correlations were calculated on

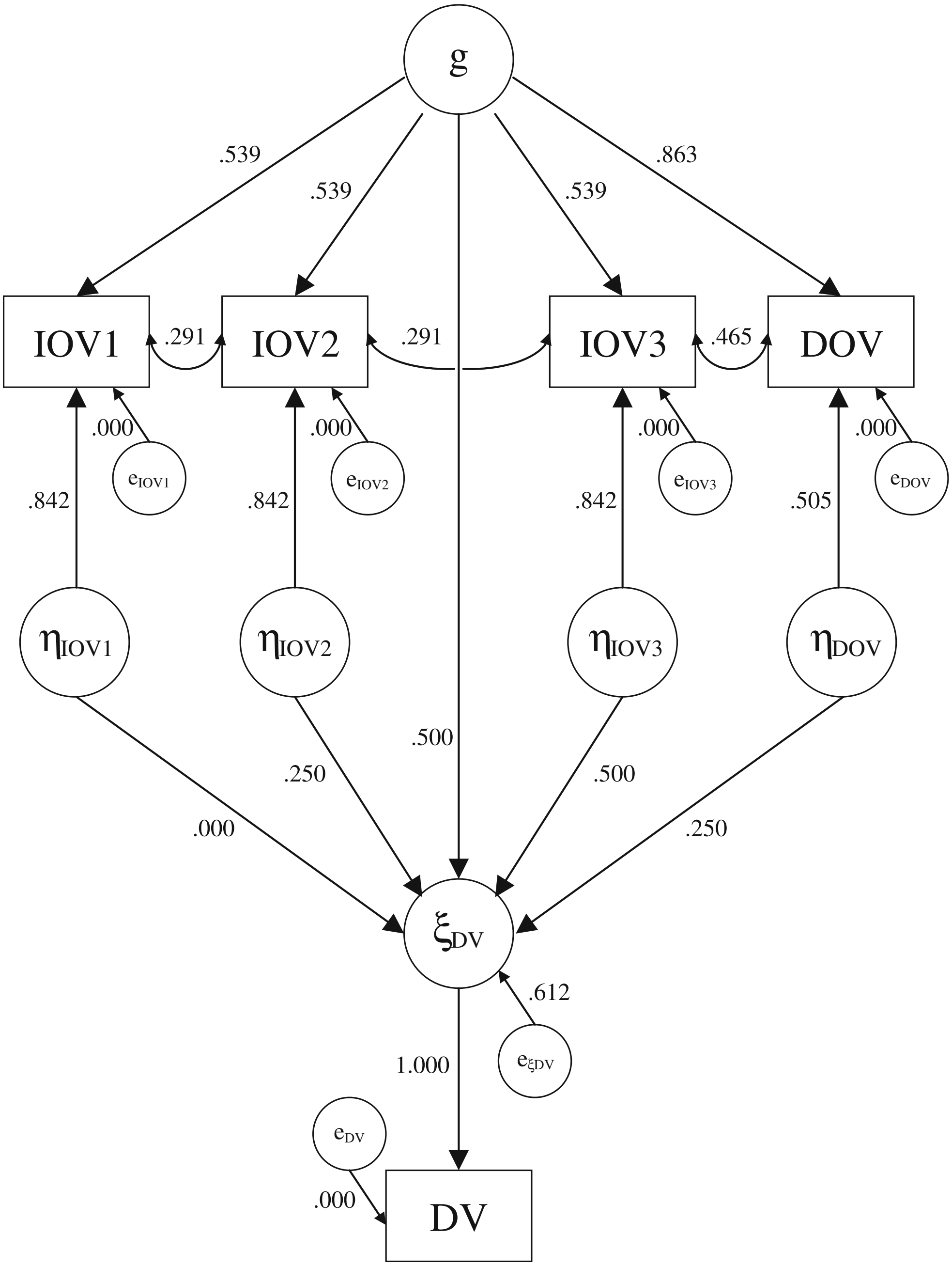

Fig. 2 Latent variable model (model 2) representing relationships between oversampled predictor variables and the categorical dependent variable (DV), in which measurement error was included by setting reliabilities for all variables at .860. Model 2 incorporated four oversampled predictor variables: a directly oversampled variable (DOV) and three indirectly oversampled variables (IOVs). Correlations between the DV and oversampled predictor variables were as follows: $r_{\mathrm{DOV,DV}}$ = .480, $r_{\mathrm{IOV3,DV}}$ = .594, $r_{\mathrm{IOV2,DV}}$ = .413, and $r_{\mathrm{IOV1,DV}}$ = .232

the basis of the relationships with the dichotomized, categorical DV.

Population correlations in model 2 were calculated by decreasing reliabilities from 1.000 to .860, while keeping constant both of the following: (1) all paths between the latent DV and the latent independent variables (η_{IOV1}, η_{IOV2}, η_{IOV3}, g, and η_{DOV}; see Figs. 1 and 2), which were 0, .25, .50, .50, and .25, respectively, and (2) the ratio of the factor loading paths (i.e., any path between an observable variable and its corresponding latent variable) to the paths between g and the observable predictor variables (DOV, IOV3, IOV2, IOV1). In model 1, the calculation of this ratio of factor loadings to $r_{g,\mathrm{OV}}$ was 0.469/0.800 = 0.586 for the DOV and 0.781/0.500 = 1.562 for the IOVs; and in model 2, this ratio was 0.505/0.863 = 0.585 for the DOV and 0.842/0.539 = 1.562 for the IOVs. Factor loadings were chosen to represent typical values in multivariate analyses.

Number of participants

In initial analyses, sample size was examined at two levels: N = 250 and N = 500. Sample sizes of this range were selected due to the frequency at which they appear within the types of longitudinal studies in which oversampling is typically applied, including studies on the development of rare psychological or neurological outcomes (Alloy et al., 2000; Zinbarg et al., 2010). In a set of additional analyses (as will be discussed further below), two additional sample sizes were also examined: N = 100 and N = 1,000. These were included to increase generalizability of results to studies with sample sizes outside of the typical range of oversampled data sets.

Shape of the relationship between the criterion variable and oversampled variables

Relationship shape was manipulated in order to determine whether results would generalize beyond the case of linear relationships between the oversampled predictor variable and the latent DV. Thus, for the simulated data set to be representative of a wide range of experimental data sets, four common relationship shapes were chosen: linear, accelerating curve, decelerating curve, and logistic function. The linear relationship was created by forming the underlying latent model with a multivariate normal distribution of errors and applying the equation $y = f(x)$. Linearity (y) was then transformed into accelerating, decelerating, and logistic shapes, by using the logistic transform with a constant c. As shown below, c could equal −1 (leading to a decelerating curve), 1 (leading to an accelerating curve), or 0 (approximating a logistic curve), and x represented the underlying latent model of oversampled predictor variables and the DV (M = 0.00, SE = 0.01):

$$y = \frac{1}{1 + (e)^{(c-x)}} \qquad (1)$$

A decelerating curve is marked by an initial steep increase, followed by a plateau. An accelerating curve is marked by a slow increase, followed by a sharp incline. The first half of a logistic function consists of an accelerating curve, and the second half consists of a decelerating curve.

Data analysis

As was explained previously, two sets of multiple regressions were performed. In the first set, the predictor variables were made up of all oversampled variables (DOV, IOV3, IOV2, IOV1), and the criterion variable was made up of the categorical DV (i.e., a dichotomous variable consisting of “0” or “1” values only, with a proportion of .05 “1” values). Results of this first regression yielded unstandardized effect sizes (i.e., regression coefficients), which were regressed separately on the linear term for the oversampling factor (1, 4, 8), the quadratic term for the oversampling factor (oversampling^2, to test for curvilinearity), sample size (250, 500), and the categorical variable of relationship shape (linear, accelerating curve, decelerating curve, logistic function). Additionally, two- and three-way interactions of oversampling (both linear and quadratic terms) with sample size and relationship shape were included as predictor variables in the regression equation. In order to obtain results based on data with differing levels of reliability (as discussed above), both sets of regressions were performed twice: once for observable variables with a reliability of 1.000 (model 1; see Fig. 1) and once for those with a reliability of .860 (model 2; see Fig. 2).

Results

Simulations identified relationships between the degree of oversampling and the DV~OV effect sizes (i.e., the amount of variance in the criterion variable that was predicted by the oversampled variables). As will be evidenced in detail below, our results supported a consistently negative relationship between the degree of oversampling and DV~OV effect sizes, regardless of the oversampled variable (i.e., DOV, IOV3, IOV2, IOV1) or measurement error ($\rho_{X,X}$ = 1.000 or $\rho_{X,X}$ = .860). The linear oversampling term yielded larger associations with DV~OV effect sizes than did the quadratic oversampling term (see Table 1), again, regardless of the oversampled variable or measurement error. Finally, although there was a main effect of oversampling on DV~OV associations, it is critical to note that the effect sizes for these relationships were small in absolute value (see Table 1).

Oversampling and measurement error

When effect sizes for all 24 conditions were examined, 20 were larger (in absolute value) for the condition of perfect measurement ($\rho_{X,X}$ = 1.000), as compared with the condition of imperfect measurement ($\rho_{X,X}$ = .860; see Table 1). When effect sizes were examined across all 24 conditions, these effect sizes were small in both perfect measurement (−.051–.013 β) and imperfect measurement (−.023–0.005 β) conditions.

Proportion of oversampled data

The effect of oversampling probability was relatively consistent across oversampled variables (DOV, IOVs); however, the size of these effects was small. As is shown in Table 1 ($\rho_{X,X}$ = .860), negative associations were found for relationships between the main effect of oversampling and DV~OV effect sizes for all OVs excepting the IOV1 (as will be discussed below). The same pattern of results was also revealed when no measurement error was included in the model ($\rho_{X,X}$ = 1.000), but with slightly stronger (i.e., more negative) effects. Data with no measurement error showed consistently stronger relationships between the main effect of oversampling and DV~OV effect size ($M = -.020\ ß$, $SE = .012$) than did data with measurement error ($M = -.011\ ß$, $SE = .004$); but these effects were small in both conditions. Effects of the quadratic oversampling term also followed the same pattern of results: Data with no measurement error showed stronger relationships between the quadratic effect of oversampling and DV~OV effect size ($M = .005\ ß$, $SE = .003$) than did data with measurement error ($M = .002\ ß$, $SE = .001$); but these relationships were again small in both conditions.

Examining additional levels of oversampling probability

In the above analyses, oversampling probability was set at the three levels of 1, 4, and 8. In order to view a more detailed illustration of the relationship between DV~OV effect size and a larger range of oversampling probabilities, a new simulated data set was created, using eight levels of oversampling (probability of oversampling = 1, 2, 3, 4, 5, 6, 7, 8). These analyses were based on data characterized by a linear relationship only (since results did not differ between relationship shapes, as is reported in detail below). As with the original set of regressions (described in the Method section), a set of logistic regressions was performed, in which the DV (again a dichotomous variable of "0" or "1" values, with a cutoff of .05) was regressed on all four OVs. Figure 3 illustrates the negative relationship between the degree of oversampling and the DV~OV effect size, while demonstrating that the magnitude of this negative relationship is small. It is important to note that Fig. 3 includes only data with a sample size of 250 and reliability set at $\rho_{X,X} = .860$.[1] Unlike relationship shape and measurement error, which did not affect the pattern of results seen in Fig. 3, sample size did affect results (see Figs. 4 and 5) and will be addressed below.

Oversampling and population correlation

For data generated from model 2 ($\rho_{X,X} = .860$), associations between the linear oversampling term and effect sizes for all four DV~OV relationships are indicated in Table 1. As is listed in Table 1, the population correlation showed a small effect on the relationship between oversampling and DV~OV effect size. The size of the population correlation was commensurate with the size of the association between oversampling and DV~OV effect size. For example, the largest population correlation ($\rho_{DV,IOV3} = .606$) was associated with the largest (absolute value) association between oversampling and DV~OV relationship ($ß = -.023$, $p < .001$; see Table 1). Likewise, the smallest population correlation ($\rho_{DV,IOV1} = .328$) was associated with the smallest (absolute value) association between oversampling and DV~OV relationship ($ß = -.003$, $p < .001$; see Table 1). It is important to note that oversampling had generally small effects on each of the DV~OV associations, since even the "largest" effect of oversampling (i.e., $ß = -.023$, DV~IOV3) was small in magnitude. This pattern was maintained in model 1, for which reliability was set at 1.000 (these results are available upon request[1]).

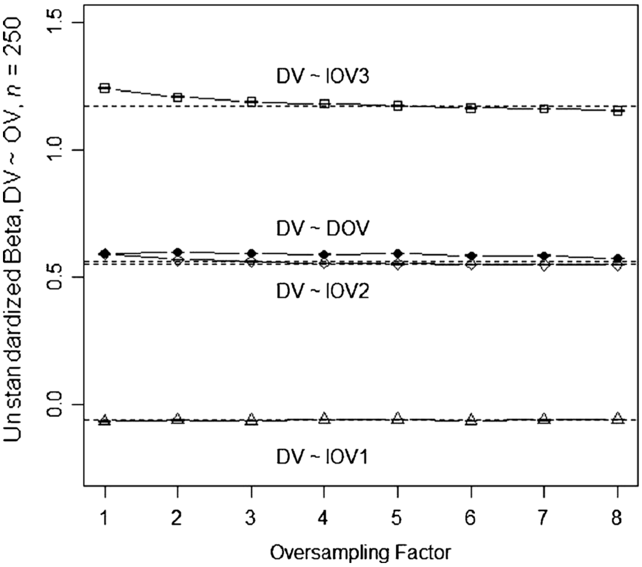

Fig. 3 Unstandardized effect size estimates for DV~OV (oversampled variables) associations at eight levels of oversampling (in which a factor of 1 represented random selection). Note that all values in this graph were based on the following: categorical DV cutoff set at .05, reliability ($\rho_{X,X}$) set at .860, $n = 250$, and linear shape. Error bars represent *SE*s. Dashed lines represent the true score of this estimate for a linear relationship within the total population (with no oversampling)

Oversampling and sample size

For the regression analyses reported in Table 1, sample size was examined at two levels: $n = 250$ and $n = 500$. Although not noted in Table 1 (to conserve space), there was a main effect of sample size, with a decrease in sample size predicting an increase in DV~OV effect sizes. However, effect sizes for this association were so small (i.e., $ß < .0002$) in absolute value that their practical significance is highly questionable and will not be discussed further.

Effects for the interaction of sample size with the degree of oversampling (as well as a three-way interaction with oversampling and shape) were also found. Again, however, these effects were nominal in size and render their practical impact questionable. The largest effects involving two- and three-way interactions with sample size are reported in Table 1;

[1] Additional results not shown in Table 1 or in the figures are available upon request. Supplementary Table 1 presents a summary of regression analyses for oversampling (predictor variables) and effect size estimates (criterion variable) based on reliability set at $\rho_{X,X} = 1.000$.

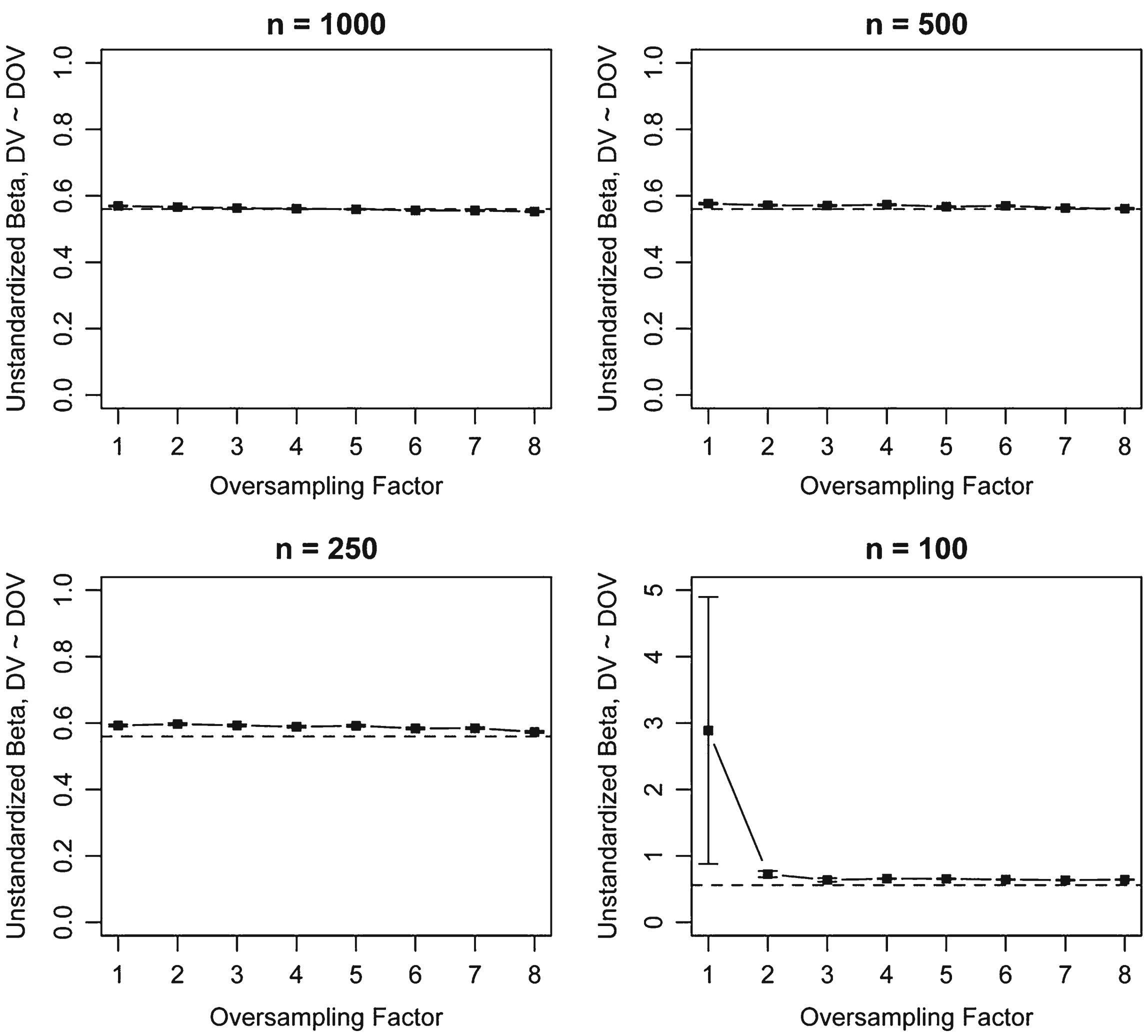

Fig. 4 The association between oversampling and unstandardized effect size estimates of the relationship between the DV and the DOV (directly oversampled variable), with linear shape and reliability ($\rho_{X,X}$) set at .860. This figure illustrates the effect of decreasing sample size. Error bars represent *SE*s. Please note the change in the *y*-axis scale for the last panel (*n* = 100). Dashed lines represent the true score of this estimate (*b* = 0.56) for a linear relationship within the total population (with no oversampling)

however, even these effect sizes (*ß*) were in the range of .00002–.00008.

Examining additional levels of sample size

In the above analyses, sample size was examined at the levels of *n* = 250 and *n* = 500. In order to view a more detailed illustration of the role of sample size, an additional simulated data set was also examined, using four levels of sample size: *n* = 100, *n* = 250, *n* = 500, and *n* = 1,000. These analyses were based on data characterized by a linear relationship (since results did not differ between relationship shapes, as detailed below). As with the original set of regressions (described in the Method section), a set of logistic regressions was performed, in which the DV (again dichotomous, with a cutoff of .05) was regressed on all four OVs.

Although the relationship between degree of oversampling and DV~OV effect size was consistent for sample sizes of 250, 500, and 1,000, the pattern of findings for *n* = 100 was markedly distinct. As can be seen in the last panel of Figs. 4 and 5, effect sizes and standard errors among samples of *n* = 100 were disproportionately large, with some effect sizes exceeding *b* = 3.5 (Fig. 4, showing effect size estimates of the DV~DOV relationship) and *b* = 4.5 (Fig. 5, showing effect size estimates of the DV~IOV3 relationship). Critically, however, this was the case only when oversampling was set at 1—indicating *no* oversampling for these data (i.e., a randomly sampled design). Oversampling served to prevent such overfitting. As in Fig. 3, Figs. 4 and 5 also include data with linear shape and reliability of .860 only; the same effects were present for analyses using data with no error ($\rho_{X,X}$ = 1.000) and nonlinear shapes.[1]

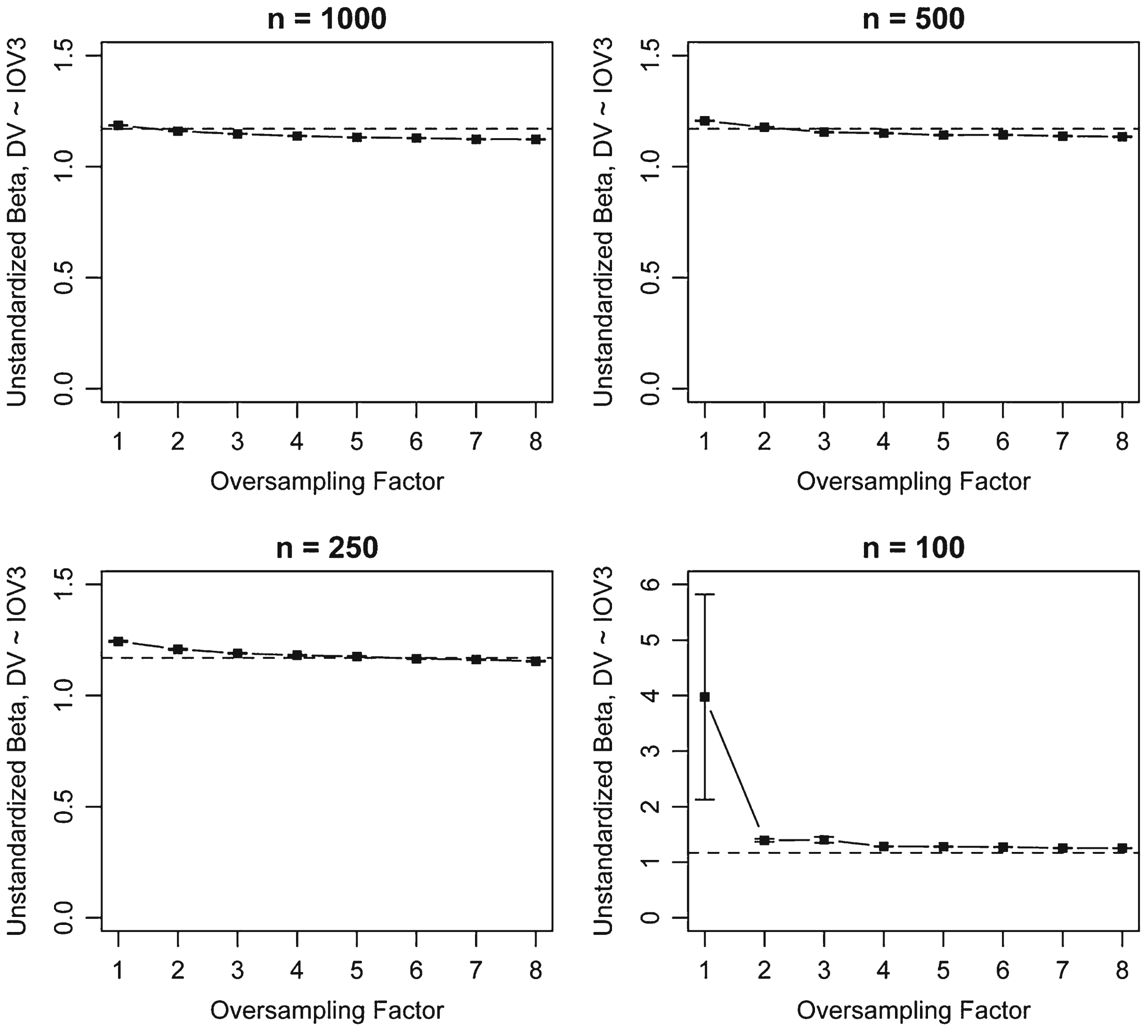

Fig. 5 The association between oversampling and unstandardized effect size estimates of the relationship between the DV and the IOV3 (indirectly oversampled variable), with linear shape and reliability ($\rho_{\mathrm{X,X}}$) set at .860. This figure illustrates the effect of decreasing sample size. Error bars represent *SE*s. Please note the change in the y-axis scale for the last panel (n = 100). Dashed lines represent the true score of this estimate (b = 1.17) for a linear relationship within the total population (with no oversampling)

The disproportionately large effect sizes and standard errors in the case of no oversampling for n = 100 were likely the result of two factors: the categorical cutoff and the number of predictor variables in the original models. For the set of simulated analyses used in the present study, the categorical cutoff was set at .05, to represent a rare outcome. Therefore, for a sample size of 100, five cases were being predicted on average. However, because data sets were randomly created (on the basis of a normal distribution), some analyses yielded fewer than five cases when oversampling was not applied. This was problematic, since analyses included four predictor variables. For example, if a smaller sample set (n = 100) included four cases, four predictors would likely predict these four cases with very large effect sizes (Peduzzi, Concato, Feinstein, & Holford, 1995; Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996; Vittinghoff & McCulloch, 2007). In the present study, with 100 observations and outcome probability set at .05, simulations yielded four cases (rather than the expected five) with a probability of .265. However, because simulations could also yield more than four cases, less exaggerated effect sizes were also included, thus creating large standard errors in the case of n = 100 (when no cases were oversampled). Notably, however, when the sample size was increased to 250, the probability of yielding four cases in an analysis (with no oversampling) plummeted to .00000265. Even with 10,000 repetitions, the probability of finding at least one data set with four cases increased only to .0265 for a sample size of 250. Thus, the same pattern of large effect sizes and standard errors was not observed in analyses with larger sample sizes.

Oversampling and shape

Shape was manipulated in order to represent four patterns between the OV (DOV, IOV1, IOV2, IOV3) and the latent

DV: linear, accelerating curve, decelerating curve, and logistic function. The three nonlinear shapes yielded similar effect sizes, either as main effects (not listed in Table 1, to conserve space[1]) or as interactions with other predictor variables (as listed in Table 1). Regardless of whether measurement error was included in the model, shape as a main effect was significant in predicting all relationships between the DV and the OVs. Since the relationship between the DV and the OV tended to follow a nonlinear shape, the effect size decreased for the associations between the DV and the DOV, IOV3, and IOV2. Interactions between oversampling * shape followed the same pattern (see Table 1) for both levels of measurement error, with the largest effects occurring for the DV~IOV3 relationship (i.e., the DV~OV relationship for which the population correlation was strongest). However, even the largest of these effects was less than β = .00008 (in absolute value). Thus, not all interactions are listed in Table 1.

Discussion

The present study is the first to report that the method of oversampling does not bias effect sizes to a degree that would typically impact results in behavioral research. Across all results for sample sizes of 250 and higher, the effect of oversampling was relatively small (with the greatest single change in effect size from 1.17 *b* with no oversampling to 1.21 *b* with oversampling). Moreover, we found that the effect of oversampling on effect sizes did not appear to differ substantially for directly oversampled predictors, as compared with indirectly oversampled predictors. Thus, oversampling does not appear to be associated with an appreciable bias favoring directly oversampled predictors over indirectly oversampled predictors (or vice versa).

The present study also determined that refraining from using the oversampling method (i.e., using random sampling) could lead to drastically biased effect size estimates for smaller sample sizes. Although sample sizes of 250 and 500 yielded small effect sizes in the original DV~OV regression analyses (see Table 1), lowering the sample size to 100 produced striking results, yielding outliers with extreme values of the associations between the predictor variables and the criterion variable (see Figs. 4 and 5). This finding underscores the danger of using a relatively small sample size to predict a rare, categorical outcome, particularly when the number of predictors (four, in the present study) may be similar to the number of expected cases of the criterion variable (five, in the present study). This finding has critical implications for researchers studying rare outcomes. If the number of predictor variables is expected to be similar to the number of cases being predicted at a particular sample size, it is advisable either to use a larger sample size or to oversample the predictor variables in order to increase the number of expected rare outcomes.

The remaining manipulated variables in the present analyses included population correlation, relationship shape, and measurement error of all observable variables. Regarding population correlation, the present study found that as population correlation increased, the effect of oversampling on DV~OV effect sizes became more negative; however, the size of this effect (β) was less than .02 (see Table 1). Regarding the shapes of the relationship between the OV and the DV, all three curve functions shared strong inverse relationships with DV~OV effect sizes (excepting the DV~IOV1). Interactions of oversampling with shape yielded results that were significant but very small, implying that the lack of sizable effects associated with oversampling was largely robust with respect to relationship shapes. Regarding measurement error, data with perfect measurement ($\rho_{X,X}$ = 1.000) generally proved to be the most sensitive to detecting effects. In comparison with data not measured perfectly ($\rho_{X,X}$ = .860), perfectly measured data yielded stronger effect sizes for oversampling as a main effect and as an interaction with other predictor variables. However, the effects size (β) values in both conditions were small (−.051–.013 β).

The present findings bear two central implications for research on the prediction of rare outcomes. First, when the number of predictor variables is similar to the number of expected cases of the criterion variable, it is advisable to increase the number of expected rare cases—either via oversampling the predictor variables or by increasing the overall sample size. If one of these two strategies is not employed, estimates of the relationship between the predictor and criterion variables may be inordinately biased. Second, if the number of predictor variables is considerably less than the number of expected cases, oversampling may be unnecessary. That is, in the present study, the only substantial effect of oversampling was limited to the case in which the number of predictors and expected cases was comparable. Thus, first determining the expected number of cases, relative to the number of predictor variables, would be a highly valuable step in considering sampling technique.

Generalization of our findings and implications is certainly not without potential limitations, however. Although we have attempted to include as many conditions in our models as methodologically feasible, there are several key untested conditions that could potentially affect our results. For example, the dependent variable in the present study was based on a purely categorical variable with no noise, which may be an overly optimistic assumption for typical data sets. It is thus unclear how a probabilistic (rather than exact) model of a categorical dependent variable would influence the effect of oversampling, and we acknowledge this as a limitation in the generalizability of our results. Similarly, our results may not extend to data sets with overall reliability substantially less than .860. We therefore suggest that future research on this topic not only examine a broader range of reliability values (e.g., $\rho_{X,X}$ = .70), but also include an additional condition in which the categorical dependent variable is based on a probabilistic model.

These limitations notwithstanding, one should be prudent when drawing conclusions regarding characteristics of a population in the use of any sampling method that is not entirely random. In general, our findings suggest that oversampling is not a relevant source of bias in behavioral research and does not appear to have an appreciably different effect on results for directly versus indirectly oversampled variables. Furthermore, this sampling technique can be advantageous and appropriate when implemented in the proper circumstances, such as when the number of expected cases is similar to or less than the number of predictor variables.

References

Abrahams, N. M., & Alf, E. F. (1978). Relative costs and statistical power in the extreme groups approach. *Psychometrika, 43*(1), 11–17.

Alf, E. F., & Abrahams, N. M. (1975). The use of extreme groups in assessing relationships. *Psychometrika, 40*(4), 563–572.

Allison, D. B., Allison, R. L., Faith, M. S., Paultre, F., & Pi-Sunyer, F. X. (1997). Power and money: Designing statistically powerful studies while minimizing financial costs. *Psychological Methods, 2*(1), 20–33.

Alloy, L. B., Abramson, L. Y., Hogan, M. E., Whitehouse, W. G., Rose, D. T., Robinson, M. S., & Lapkin, J. B. (2000). The temple-wisconsin cognitive vulnerability to depression project: Lifetime history of axis I psychopathology in individuals at high and low cognitive risk for depression. *Journal of Abnormal Psychology, 109*(3), 403–418.

Alloy, L. B., Abramson, L. Y., Whitehouse, W. G., Hogan, M. E., Panzarella, C., & Rose, D. T. (2006). Prospective incidence of first onsets and recurrences of depression in individuals at high and low cognitive risk for depression. *Journal of Abnormal Psychology, 115*(1), 145–156.

Borich, G. D., & Godbout, R. C. (1974). Extreme groups designs and the calculation of statistical power. *Educational and Psychological Measurement, 34*(3), 663–675.

Cohen, J. (1983). The cost of dichotomization. *Psychological Measurement, 7,* 249–253.

Costello, E. J., Angold, A., Burns, B. J., Stangl, D. K., Tweed, D. L., Erkanli, A., & Worthman, C. M. (1996). The Great Smoky Mountains Study of youth: Goals, design, methods, and the prevalence of DSM-III-R disorders. *Archives of General Psychiatry, 53*(12), 1129–1136.

DuMouchel, W. H., & Duncan, G. J. (1982). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association, 78*(383), 535–543.

Feldt, L. S. (1961). The use of extreme groups to test for the presence of a relationship. *Psychometrika, 26,* 307–316.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science, 22*(2), 153–164.

Hirtz, D., Thurman, D. J., Gwinn-Hardy, K., Mohamed, M., Chaudhuri, A. R., & Zalutsky, R. (2007). How common are the "common" neurologic disorders? *Neurology, 68*(5), 326–337.

Humphreys, L. G. (1985). Correlations in psychological research. In D. K. Detterman (Ed.), *Current topics in human intelligence* (Research methodology, Vol. 1, pp. 3–24). Norwood, NJ: Ablex Publishing.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry, 62*(6), 593–602.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19–40.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114*(2), 376–390.

Menard, S. (2004). Six approaches to calculating standardized logistic regression coefficients. *The American Statistician, 58*(3), 218–226.

Peduzzi, P., Concato, J., Feinstein, A. R., & Holford, T. R. (1995). Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates. *Journal of Clinical Epidemiology, 48*(12), 1503–1510.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology, 49*(12), 1373–1379.

Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the Extreme Groups Approach: A critical reexamination and new recommendations. *Psychological Methods, 10*(2), 178–192.

Development Core Team, R. (2007). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Vittinghoff, E., & McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology, 165*(6), 710–718. doi:10.1093/aje/kwk052

Zinbarg, R. E., Mineka, S., Craske, M. G., Griffith, J. W., Sutton, J., Rose, R. D., & Waters, A. M. (2010). The Northwestern-UCLA youth emotion project: Associations of cognitive vulnerabilities, neuroticism and gender with past diagnoses of emotional disorders in adolescents. *Behaviour Research and Therapy, 48*(5), 347–358. doi:10.1016/j.brat.2009.12.008