Educ Psychol Rev (2011) 23:245–261
DOI 10.1007/s10648-011-9158-7

REVIEW ARTICLE

Cause and Event: Supporting Causal Claims through Logistic Models

Ann A. O'Connell · DeLeon L. Gray

Published online: 15 May 2011

Abstract Efforts to identify and support credible causal claims have received intense interest in the research community, particularly over the past few decades. In this paper, we focus on the use of statistical procedures designed to support causal claims for a treatment or intervention when the response variable of interest is dichotomous. We identify seven key features of logistic regression studies that should play a critical role in estimating a causal effect and discuss their implications for causal inference. These include elaboration of research design, clarification of link function, model specification, challenges and limitations of sample size, interpretation of treatment effect through odds ratios, statistical tests and examination of model fit, and the potential for multilevel logistic models in pursuit of causal claims. Our recommendations are intended to guide researchers in the critical evaluation of logistic regression models for analyses culminating in causal claims and to promote stronger design and modeling strategies for reliable causal inference.

Keywords Causality · Logistic regression

The quest to understand what works in educational and psychological research—indeed, in all research that seeks to identify ways to improve the human condition—remains a dominant theme in continuing debates on scientifically based research, causality, and evidence-based practice (Riehl 2006; Schneider *et al*. 2007; Towne *et al*. 2004). The most rigorous research designs are perceived as those involving random assignment or that otherwise minimize the effect of selection bias and potential for rival plausible alternatives when establishing a treatment effect. Support for valid estimates of treatment effects from

A. A. O'Connell (✉)
College of Education and Human Ecology, The Ohio State University, 29 West Woodruff Road, Room 211A Ramseyer, Columbus, OH 43210, USA
e-mail: aoconnell@ehe.osu.edu

D. L. Gray
College of Education and Human Ecology, Program in Educational Psychology and Philosophy, The Ohio State University, 29 West Woodruff Road, Columbus, OH 43210, USA
e-mail: gray.561@osu.edu

well-designed observational studies has been documented in the literature for some time (e.g., Benson and Hartz 2000; Concato *et al*. 2000; Rosenbaum 1989; Rosenbaum and Rubin 1983; Rubin 2001; Schneider *et al*. 2007). The past decade has also seen a burgeoning commitment within the research community for transparency in design and responsible reporting of research results for both experimental and non-experimental designs (e.g., Moher *et al*. 2001; Des Jarlais *et al*. 2004; National Research Council (NRC) 2002; Towne *et al*. 2004; Schneider *et al*. 2007). What these efforts have in common is the understanding that credible claims to causality of effect are rooted in quality of *design* and are not the privilege of any particular statistical techniques. However, each statistical technique used in the pursuit of causal inference has unique limitations and challenges that can affect the reasonableness of causal claims.

In this paper, we focus on the use of logistic regression in studies designed to measure a causal effect of a treatment or intervention when the response variable of interest is dichotomous. As one of the initial steps in the process of "prudent inquiry" (Shadish 2005), we emphasize the need for presentation of responsible research conclusions from studies involving logistic regression, and we make recommendations for when prescriptive statements regarding causal effects can be supported.

Before turning to the primary goal of our paper, we wish to review some terminology about data, models, and inference. We refer to data as the observations obtained on a sample of persons or units from a targeted population. The sample constitutes a subset of the population, and the data are used to describe patterns and relationships among variables in the sample. Generally, the variables are chosen to investigate specific research problems, such as whether a prevention program decreases engagement in risky sexual behavior among HIV-positive men, whether a community-level intervention for pregnant women decreases the rate of low-birth-weight infants, or whether a school-based intervention reduces the incidence of dropout among adolescents. The quality of the sample, including how it was collected and the measurement properties of the sample variables, affects the quality of evidence regarding the nature of observed data patterns and variable relationships (Fowler 2009; Kish 1965/1995; O'Connell 2000).

A statistical model is a mathematical representation of a supposition, belief, or theory regarding the patterns and relationships in the data and is often principally concerned with characterizing and simplifying the complex relationships among a constellation of predictor or explanatory variables and one or more outcomes, responses, or dependent variables. A famous quote by Box captures the challenges in using statistical models to represent complex processes: "All models are wrong, but some are useful" (Box 1979, p. 202). The legitimacy of a given statistical model to accurately capture and simplify complex relationships and the legitimacy of inferences made about those relationships are tightly intertwined. Both rely heavily on the adequacy and quality of the sample, the research problem, and the research design (Dannels 2011; Fox 2008; Shadish *et al*. 2002).

Statistical models may be useful, but they are not determinant. They are capable of describing and elucidating the structure of complex educational and social phenomena (Fox 2008), yet the situations in which we can use statistical models to move towards causal inference require an appreciation for research design in general as well as an understanding of the strengths and limitations of a selected statistical modeling method. Regardless of its elegance or appeal, there is no single statistical method capable of resolving problems in theory or research design (McCoach *et al*. 2007). Our focus here is on logistic regression for modeling occurrence of events and its applicability to causal inference. Our hope is that this work will encourage readers to more critically evaluate the use of logistic regression models for analyses culminating in causal claims and promote stronger design and modeling strategies for reliable causal inference from logistic models.

We begin by describing the statistical model for logistic regression of event occurrence. We then discuss the estimation of causal effects when the response of interest is dichotomous and review the validity of inferring causal effects from experimental and non-experimental studies. We conclude with our recommendations for when prescriptive statements are warranted for studies utilizing logistic regression, emphasizing conditions necessary for defensible prescriptive claims. We offer suggestions for the improvement of interpretation of results from logistic regression analyses, clarifying relevant terminology for probability, odds, odds ratios, relative risk, and effect sizes.

Modeling Event Occurrence through Logistic Regression

Logistic regression (LR) is used to model event occurrence when the event under study is dichotomous, such as whether an adolescent drops out of school. Given the limited range of the response variable (i.e., 0=event does not occur; 1=event occurs), LR falls into the class of generalized linear models (GLM; Fox 2008; Long 1997; McCullagh and Nelder 1989) for which the familiar normal distribution is a special case.

GLMs are formally identified through three specific features: (1) a random component that is based on the exponential family (e.g., normal, binomial, Poisson); (2) a linear component that describes how a transformation of the expected value of the response variable corresponds to a set of covariates or explanatory variables; and (3) a link function that specifies the connection between the original and the transformed responses. In this paper, we focus on outcomes that can be modeled through LR, for which extensions exist for ordinal responses, counts, and time-to-event data.

LR models the probability for one of the outcomes, typically referred to as the "success" event, while conditioning on a set of covariates. LR captures the relationship between the probability of success and the linear predictor, which typically resembles an S-shaped curve. The Bernoulli distribution (a binomial distribution where the number of trials equals one) forms the random component of the LR model, and the distribution of the collection of outcomes is expressed as $Y_i \sim B(1, \pi_i)$, where $Y_i=1$ for the ith person if the success outcome is observed and 0 otherwise, B indicates the binomial distribution, 1 indicates the number of trials (each individual forms a single trial), and π_i represents the probability of a successful outcome for the ith case. Distributional assumptions give the mean and variance of Y_i as π_i and $\pi_i*(1-\pi_i)$, respectively. Consequently, the probability of success is heteroscedastic across cases, in that the variance for each case is different and depends on the expected value.

The linear component of the LR model describes how a transformation of the expected values η_i is related to a linear combination of p predictors (covariates), one of which represents treatment assignment, with $T_i=1$ for the treatment group and 0 for the control group.

$$\eta_i = \beta_0 + \beta_1 T_i + \beta_2 X_{i2} + ...\beta_p X_{ip} \tag{1}$$

In logistic regression, the logit link function, or natural log of the odds of success, serves as the link connecting the expected values and the collection of predictors. For any given collection of predictors (including the treatment), $\underline{x}_i$, the odds is a quotient that compares the probability of success, $\pi(\underline{x}_i)$, to the probability of failure, $1-\pi(\underline{x}_i)$. Thus, the LR model can be written as:

$$\eta_i = \text{logit}(\pi(\underline{x}_i)) = \ln\frac{\pi(\underline{x}_i)}{1-\pi(\underline{x}_i)} = \beta_0 + \beta_1 T_i + \beta_2 X_{i2} + ...\beta_p X_{ip} \tag{2}$$

Although probability is constrained to be between 0 and 1, the logit maps the expected values onto the real line from $-\infty$ to $+\infty$. Each slope (holding all other effects constant) represents the change in the logit that is expected to occur given a one-unit change in its respective predictor.

The logistic regression model describes how the log of the odds of success varies by a given set of predictors including the treatment variable but is ultimately used to estimate probability of success conditioning on the predictors in the model. The antilog (inverse) of an estimated logit provides a prediction in terms of the odds of success, given the set of predictors:

$$\exp(\widehat{\eta}_i) = \text{Odds}(\text{success}|\underline{x}_i) = \frac{\widehat{\pi}(\underline{x}_i)}{1-\widehat{\pi}(\underline{x}_i)} = \exp(\widehat{\beta}_0 + \widehat{\beta}_1 T_i + ...\widehat{\beta}_p X_{ip}) \tag{3}$$

This expression is then used to estimate the conditional probability of success:

$$\widehat{\pi}(\underline{x}_i) = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 T_i + ...\widehat{\beta}_p X_{ip})}{1+\exp(\widehat{\beta}_0 + \widehat{\beta}_1 T_i + ...\widehat{\beta}_p X_{ip})} = \frac{1}{1+\exp(-(\widehat{\beta}_0 + \widehat{\beta}_1 T_i + ...\widehat{\beta}_p X_{ip}))} \tag{4}$$

For any single predictor, an odds ratio (OR) can be formed by exponentiating its respective regression coefficient (i.e., e^{β}). The OR for a variable yields a measure of effect size that describes the association, in terms of odds of success, between that variable and the success outcome. In Eq. 1 above, β_1 represents the expected change in the logit for a one-unit change in T, holding all else constant. Thus, for the treatment effect, we obtain an odds ratio that compares the odds of success within the treatment group to the odds of success within the control group:

$$\text{OR}(\text{treatment}) = \exp(\widehat{\beta}_1) = \frac{\text{odds}(\text{success}|T_i = 1, \underline{x}')}{\text{odds}(\text{success}|T_i = 0, \underline{x}')} \tag{5}$$

In Eq. 5, $\underline{x}'$ represents all remaining covariates in the model. The odds and ORs are always non-negative and range from 0 to infinity. In general, an OR of 1.0 implies that the predictor has no associative effect on the odds of success while smaller values of an OR ($0<\text{OR}<1$) indicate that the odds of success tend to decrease as the predictor increases by one unit. Conversely, ORs larger than 1 imply that the odds of success tend to increase as the predictor increases by one unit. The relationship between the odds ratio and a causal effect for treatment is described below.

Defining a Causal Effect for Binary Outcomes

The framework commonly referred to as Rubin's Causal Model (Holland 1986; Little and Rubin 2000) characterizes the problem of estimation of a causal effect as one of missing data. All measures of effect are comparative, but, within most treatment studies, participants only experience and provide a response under one of the possible treatments. This situation forms what is known as the counterfactual: What would a person's response be *if* they had been assigned to the competing treatment? Essential to Rubin's model is the assumption of independence among participants (or units) in the study, known as the stable unit-treatment value assumption, or SUTVA. This assumption implies that one participant's response or experience has no effect on another person's response. More specifically, SUTVA is an "exclusivity restriction," which states that a participant's potential outcomes under either

treatment or control are stable and are not affected by another participant's assignment to treatment or control (Guo and Fraser 2010; Rubin 2004).

To estimate a causal effect according to Rubin's framework, we need to approximate the responses of all participants (i.e., same background characteristics, same context, etc.) that would have occurred if they had received the other condition. The quality of this approximation is largely, but not completely, resolved through randomization of assignment, which eliminates the possibility of selection bias and ensures that effects of unobserved variables are balanced between treatment groups (Shadish *et al*. 2002).

In general, the causal effect of a treatment on a single case or individual can be conceptualized as the difference between the potential observed response and the counterfactual, although only one response is actually observed. For observed responses, R, in the treatment (T_1) and comparison (T_0) groups, this difference is estimated by the *average treatment effect*, $E(R(T_1))$–$E(R(T_0))$, which, for a binary outcome, reduces to a comparison of two probabilities: $[P(R(T_1)=1)]-[P(R(T_0)=1)]=\widehat{\pi}_{\text{treatment}}-\widehat{\pi}_{\text{control}}$ (Rosenbaum and Rubin 1983). These two probabilities are conditional probabilities for success within the treatment and control groups, respectively, and are adjusted for the presence of the other covariates in the model. They are precisely those probabilities estimated through the logistic regression model as expressed in Eq. 4, and they each contribute, respectively, to the determination of the odds shown in the numerator and denominator of the odds ratio in Eq. 5 above. Thus, the odds ratio for the treatment assignment variable serves as the most appropriate measure of the causal effect for binary outcomes.

Randomization to treatment is an important component in the experimental design of treatment or intervention studies, but, at times, researchers who are interested in the causes of an event may be unable to assign participants to a condition. In such instances, quasi- or non-experimental research designs can yield strong causal inferences when methods to approximate the randomization process and adjust for selection bias are incorporated into the design and alternative explanations for the results are shown to be implausible (Little and Rubin 2000; Shadish *et al*. 2002). The goal of these approximation strategies is to balance preexisting differences in the data between treatment and control groups so that the estimated treatment effect is free of any potentially confounding effects of observed covariates.

Case–control studies (Breslow 1996) can be useful designs from which to infer causality when outcomes are dichotomous and randomization to treatment is not an option. In these studies, cases of participants who have the event of interest (e.g., high school dropouts) are matched on covariates to similar control participants who do not have the event of interest (e.g., non-high school dropouts). Important (and often difficult) aspects of case–control studies are the selection of control cases and the quality of matching.

Several matching methods exist (Rubin 1973, 1979, 1980), but it is not always possible to find an adequate match on key variables for individual participants in both conditions. Propensity scores can be used to overcome this problem and, when designed appropriately, are an effective matching tool in case–control as well as within observational or quasi-experimental studies (see Guo and Fraser 2010, for an extensive discussion on development, adequacy, use, and limitations of propensity scores). Propensity scores are typically estimated through a logistic regression analysis where the outcome is the assignment to treatment or comparison group. When included in the analysis, they act to mimic some of the desirable properties of randomization by balancing the effect of observed covariates and thus reducing overt bias in the estimation of treatment effects. In addition, other model-based adjustments including the use of instrumental variables

(Angrist *et al*. 1996) or covariance adjustments (Rosenbaum 2002) are enhanced when propensity score methods are used (Rubin 2001). Thus, when appropriately constructed, interpretations and prescriptive statements inferred from results of studies utilizing propensity score adjustments may be more confidently supported.

Common techniques that incorporate propensity scores include matching, stratification, and weighting; the propensity score has been used in some studies as a covariate, but this is prone to extrapolation issues (Little and Rubin 2000). Furthermore, when key variables are omitted from the propensity model, it is impossible to adjust for potential confounds and subsequently, to infer causality, regardless of how propensity scores are used in an analysis. Sensitivity analyses can and should be used to investigate the impact of hidden bias due to variables excluded from the propensity model or the analysis and to rule out the impact of unmeasured variables that may be related to treatment group or response (Rosenbaum 1995).

Reliable propensity scores balance individuals across treatment and control conditions though the use of selection predictors—thus adjusting for potential confounds and estimation bias resulting from non-randomization. Since propensity scores are designed to mimic the randomization process, they reduce reliance on covariates in non-randomized designs. In constructing good propensity scores, it is important for researchers to include predictors that are likely to influence group selection in the propensity score, even when such variables are not significant predictors of the event of interest (Rubin and Thomas 2000). Residual bias may still exist even in the presence of well-constructed propensity scores when covariates hold prognostic value. Rubin and Thomas recommend that these special covariates be included in the construction of the propensity score as well as adjusted for when predicting the event of interest. Such covariance adjustment reduces bias in the estimation of the treatment effect (Hill 2008). The overall goal is to compare outcomes on individuals *after* treatment that were similar *before* treatment. However, further research is needed on approaches to controlling for covariates that may be differentially related to assignment and to the outcome(s) (Guo and Fraser 2010).

The desire to randomize is compelling, but, even with randomization or a strong approximation thereof, other aspects of the research design can impinge on the validity of researcher recommendations regarding causal effects. These design elements include, as a minimum, the use of a correctly specified statistical model, including any necessary interactions among predictors; a sample that sufficiently models the desired population; the ability to truly manipulate the purported cause through random assignment to treatment or intervention groups or the quality of efforts to mimic this randomization; fidelity to implementation of the treatments or interventions employed; measurements of the dependent variables and all predictors that are valid for the population employed and that yield reliable scores for that population; information on the presence of missing data, particularly the impact of attrition in treatment intervention studies; and quality of model fit in terms of absence of extreme or unusual cases or outliers in the data. Attention to these integral aspects of a research design enhances the possibility that competing or rival hypotheses for the observed effect can be refuted and motivates support for prescriptive statements regarding causal effects. It should be noted, however, that basing study rigor on any single characteristic can itself be flawed; all studies will have their flaws, the question is, "where and how much" (Amico 2008, p. 5).

When Are Prescriptive Statements Warranted?

A poorly designed research study will yield uninformative results, regardless of the sophistication of the statistical methods employed by the researchers. Light *et al*. (1990)

succinctly state, "you can't fix by analysis what you bungled by design" (p. viii). In Table 1, we present our recommendations for interpreting and reporting on logistic regression results when support for a causal effect is the primary goal of the research. We emphasize, however, that the use of responsible research design is necessary before evaluating the reasonableness of prescriptive statements regarding effectiveness of a treatment or intervention, regardless of the analysis technique employed to address the primary research question. In fact, much of what we have discussed above cuts across all inferential statistical techniques. Thus, our first recommendation argues for use and presentation of solid research design principles to support causal claims. Our remaining recommendations focus on the specific estimation issues, interpretation of effects, statistical tests, and design criteria specifically relevant to logistic regression and that can impact the quality of causal claims based on a particular evaluation or study.

Table 1 Recommendations on research design and logistic regression features for supporting causal claims

Recommendations	
Design	Research design should be presented in sufficient detail to allow readers to understand its limitations and strengths for the phenomena under study.
Specification of link function	The validity of the selected link function and its appropriateness given the distribution of the data and the nature of the response variable (dichotomous, ordinal, nominal) should be examined and justified by the researcher.
Inclusion of all relevant predictors	Strong theoretical support for all model variables and interactions scientifically relevant to the outcome must be framed around existing literature and conceptually tied to previous research in the area being investigated.
	The use of propensity scores or an alternative strategy when randomization to treatment is not feasible must be detailed and justified. When propensity scores are employed, the researcher must justify the inclusion method used (matching, subclassification, weighting). The impact of potential hidden bias through sensitivity analyses must also be reviewed and discussed.
Sample size	Researchers must provide detailed information on the total sample size and the distribution of variables within each response level; the degree of sparseness within their data; and the impact that the presence of this sparseness has on estimation quality, particularly for the treatment effect variable, and ensure that the sample size is sufficient for reliable estimation of the causal effect.
Correct interpretation of effects through odds ratios	Researchers should use the treatment variable odds ratio as the primary estimate of causal effect and justify its strength in terms of the context of the research study and the phenomena being examined.
Model fit	Statistical assessment of model fit must be supplemented by substantive measures of model quality including measures of predictive efficiency for classification and pseudo-R^2 measures for model comparisons.
Multilevel models for variability of treatment effects	Research designs examining the impact of interventions or treatments across multiple settings should employ multilevel analyses in order to assess variation in size of treatment effects across settings and, where the number of settings allows, identify factors associated with treatment size variability.

Recommendations

Research design

To contribute to the progression of scientific inquiry, researchers must be prepared to elaborate on the quality of their research design. It takes great skill and patience to reflect on an executed design process and consider the many confounders that might weaken a desired effect. Causal inferences must be coupled with the knowledge that research methods and subsequent results are reported with integrity and in such a way as to invite responsible professional scrutiny, which is seen as essential to continued scientific progress (Dannels 2011; NRC 2002). This principle is relevant not only for studies employing logistic regression but for research in general.

Specification of link function

Logistic regression is not impervious to the effects of model misspecification (Begg and Lagakos 1992). In particular, recent research has shown that even in unconfounded studies examining treatment effects based on logistic regression, the logistic model is not robust under misspecification of the link function (Cangul *et al*. 2009). Failure to examine the validity of the link function for one's data can lead to invalid conclusions regarding a treatment effect. This is disconcerting, because the logit link is the most commonly applied link function for binary data. However, it is not the only choice available to researchers. For example, another popular choice is the probit link, which relates the probability of the success outcome to the cumulative density function of the unit normal distribution. Another alternative, often applied when response data are ordinal, is the complementary log–log (clog–log) link, which relates the probability of response to the extreme value distribution (Agresti 2007; O'Connell 2006). In comparison, probability under the logit link is related to the odds of the success event, as shown in Eq. 4. Odds and log-odds are substantively easier to understand and interpret than values obtained through the probit or clog–log links, which is often seen as an advantage to use of logistic regression over other link options (Fox 2008). When the response data are dichotomous, probit and logit models perform similarly for a probability that lies in the range between 0.1 and 0.9 (McCullagh and Nelder 1989), but for extremely rare (or extremely common) events, researchers should compare results using alternative link functions.

Inclusion of all relevant predictors

Claims for causal effect, even within a well-executed randomized design, are compromised by missing covariates, exclusion of substantively meaningful and relevant interactions, and weak theoretical support for the expected behavior of predictors on the response. These omissions or misspecifications will affect the quality of the causal effect for the treatment variable in the logistic regression model. Support for model theory—operationalized by identification and inclusion of all relevant predictors, potential confounders, or suppressor variables—must be rigorously established by the researcher in order to justify the theoretical underpinning of the model relative to the phenomena under investigation and to clarify how the constellation of variables included in the logistic regression model relate to the outcome of interest (O'Connell and Amico 2010).

Careful fitting of a statistical model in terms of main effects and interactions helps researchers to accurately represent the phenomenon under investigation (Harrell *et al*.

1996). A model predicting school dropout among adolescents, for example, may include academic performance factors such as grade point average (GPA) and psychological factors such as perceived school belonging. Perhaps the odds of school dropout in this model tend to increase with lower GPA. Like GPA, perhaps lower school belonging is associated with increased odds of school dropout. An interaction between GPA and school belonging may reveal that the odds of school dropout tend to increase with lower GPA only among adolescents who do not feel that they are personally accepted, respected, and included in the school environment (low school belonging). The interaction may also reveal that, among adolescents high in perceived school belonging, the odds of school dropout generally tend to decrease—regardless of GPA. The inclusion of the GPA-by-school belonging interaction is important here because it more accurately reflects the effect of GPA and school belonging on school dropout. An intervention to reduce school dropout among adolescents may be differentially effective for those with high- versus low-levels of school belonging, or high- versus low-levels of GPA. The processes through which an intervention is predicted to succeed must be appropriately represented in the design and analysis in order to establish and support a causal effect of the treatment.

The inclusion of all salient interaction effects is extremely important in analyses designed to recognize and promote a causal effect. Interactions are used to *qualify* causation: They contain information on how specific background conditions or characteristics—and the treatment or program under study—might work together to modify the probability of occurrence of a targeted event, such as dropping out of school, or having a low-birth-weight baby. Given the complexity of educational and social contexts in which the cause for an event is often embedded, the need for credible estimates of a treatment effect requires careful consideration of all relevant interactions among covariates as well as between covariates and the treatment variable. Interactions, however, add complexity to the model, and the capacity to reliably test and interpret these interactions is affected by sample size, which we discuss next.

Sample size

In logistic regression, sample size is a tricky issue. Researchers generally think of sample size as the number of elements or cases in the overall study or analysis, but, in logistic regression, the sample size per outcome (i.e., the number of events and the number of non-events) plays an important role in the validity of causal inferences, as does sample size per covariate pattern. Additionally, odds ratios tend to vary based on sample size and data stratification, and smaller sample sizes result in systematic bias in LR estimates such that odds ratios are falsely inflated (Nemes *et al*. 2009).

Estimation in LR is through maximum likelihood, a large-sample methodology. Several interrelated factors can impact on optimal sample size for reliable estimation of a causal effect. These include the treatment variable effect size (odds ratio) deemed substantively important for support of a causal claim; the base rate of success within the population of interest, sometimes referred to as the rareness of the event; sample size differences between the two response categories; the types of variables included in the model (i.e., continuous or categorical); the sample size for each replicated covariate pattern, which is an indication of sparseness within the data; and the expected number of successes per covariate pattern (O'Connell and Amico 2010). A covariate pattern refers to specific combinations of levels of the predictor variables. For example, if gender (two levels) and GPA (operationalized as five levels) are included in a model as predictors of school dropout, there are at least ten possible covariate patterns based on these two variables alone: males and then females coupled with each of the five GPA levels. The individual cells defined by these

combinations can differ dramatically in size and become sparser as additional predictors are added. In particular, data become thin when continuous predictors are present because similar or shared patterns of predictor variable values are not likely to exist, making reliable tests of variable effects and model goodness-of-fit problematic to establish. Since interactions represent an attempt to capture differential effects due to combinations of levels of the variables they consist of (possibly including the treatment variable), this same limitation in terms of sample size becomes even more important. Hosmer and Lemeshow (2000) recommend that the sample size for the smallest response group be $10(p+1)$, at the very least; others recommend at least ten events per variable (Steyerberg *et al*. 2001; Peduzzi *et al*. 1996). These recommendations, however, should be adjusted for sparse covariate patterns or interactions within the data.

Sparseness is a distinctive and problematic result that frequently occurs during logistic regression modeling and negatively affects the estimation of standard error for a predictor. Data sparseness can lead to non-convergence of the maximum likelihood estimation process, and the problem is exacerbated with smaller sample sizes. In the absence of sufficient preliminary descriptive analyses, sparseness is generally detected by the presence of severely inflated odds ratios, stemming from severely underestimated standard errors for a predictor.

Related to the presence of sparse data are the conditions of separation (sometimes called quasi-separation) and complete-separation. Separation occurs when there are near or perfect predictions for particular covariate patterns and tends to occur within smaller samples or when the event of interest is rare. Interestingly, separation is a condition that does not typically occur in the presence of continuous predictors, but the risk does increase as the number of predictors becomes closer to the sample size. Perfect predictors are rarely informative in a statistical model because they function simply as a duplicate of the response variable on the predictor side, and their perfectly collinear association with the outcome will mask the relationship between any of the other predictors and the response. Thus, researchers must be guarded against variables that may contribute to quasi- or complete-separation, especially within small samples.

Recommendations for minimum sample sizes for logistic regression exist in the literature, and some are based on sample size needed in the smallest response group relative to the number of predictors (e.g., Aldrich and Nelson 1984; Hosmer and Lemeshow 2000; Peduzzi *et al*. 1996). However, we believe that decisions on sample size in a treatment effectiveness study utilizing logistic regression should not rest on established "rules of thumb." In order to ensure that an estimated sample size is sufficient for detection of a substantively meaningful effect size (odds ratio), researchers need to balance expectations regarding the distribution of covariates across response groups, the degree of sparseness anticipated within their data, and the impact the presence of this sparseness has on estimation quality, particularly for the treatment effect variable. Power analysis programs likely will not capture these nuances of logistic regression modeling; consequently, power calculations for minimal sample size should be adjusted upwards. However, this remains an area for continued research.

Correct interpretation of effects through odds ratios

When predictors are categorical or dichotomous, odds ratios and their respective confidence intervals are the most widely used measures of effect in LR, and they can also be used for interpretation of effects for continuous predictors. Some alternatives exist for continuous predictors, such as full or partially standardized logit regression coefficients, but their use is

controversial, and agreement is lacking among researchers as to their utility and validity (Menard 2004a, b). Menard (2000) as well as Hosmer and Lemeshow (2000) emphasize effect interpretation on clinical and theoretical importance of predictors rather than relative to other model predictors.

As a rule of thumb, Haddock *et al*. (1998) state that odds ratios greater than 3.00 (or less than 0.33) indicate strong relationships. However, effect sizes are most meaningful when discussed in context; it is rarely appropriate to discuss effect sizes in terms of strict, decontextualized cutoff criteria. Researchers should base their interpretation of the strength of the odds ratios on the literature about the phenomenon of interest.

An unfortunate common practice in reports of substantive research using LR is to interpret the odds ratio as a relative risk. Here, we present an example interpreting the odds ratio for a causal effect to clarify the similarities and distinctions between odds ratio and relative risk. We refer readers to Menard (2002) and O'Connell (2006) for interpretation of ORs for continuous predictors. Assume we have data from a hypothesized well-designed quasi-experimental study to ascertain the causal effect of an intervention designed to prevent school dropout. Propensity scores were used to adjust for selection bias, and strong design principles were used throughout the study to eliminate plausible alternatives for the findings and strengthen support for causal inferences. Logistic regression was performed, and after controlling for the effects of relevant covariates, the estimated probability of dropping out for students in the treatment group was $\widehat{\pi}_1 = .17$; for the control group, this was $\widehat{\pi}_2 = .21$. Thus, the odds of dropping out for the treatment group is $.17/(1-.17) = .205$; the odds of dropping out for the control group is $.21/(1-.21) = .266$. The OR is $\text{odds(treatment)}/\text{odds(control)} = .205/.266 = .77$. The *correct* interpretation: The odds of dropping out for a student in the treatment group is 0.77 times the odds for a similar student in the control group. While this finding indeed states that dropping out is less likely to occur within the treated group, it would be *incorrect* to interpret this OR as a relative risk, i.e., claim that treated students are 0.77 times as likely to drop out relative to students in the control group. The difference seems subtle, and the confusion stems from a reliance on treating odds ratios as if they represent relative risk (Davies *et al*. 1998; Hosmer and Lemeshow 2000).

For a given level of a predictor (e.g., the treatment group in an intervention study), the "risk" is the proportion of cases for which the event occurred. Relative risk (RR) is a ratio of risks between groups, such as treatment versus control. For our hypothetical example above, $\text{RR} = .17/.21 = .81$. Both the OR and the RR are in agreement in that they are both less than 1.0, and thus, dropping out is *less likely* among the treated. Specifically, the RR shows that the risk of dropping out for a student in the treatment group is 0.81 times the risk for a similar student in the control group. Interpretations of the OR and the RR generally should support each other, but they are not the same, and the OR tends to overestimate the RR, which we have seen here (i.e., 0.77 is *stronger* (farther from 1.0) than 0.81). An OR below 1.0 is often challenging to interpret, but, if our hypothetical study was modeling the probability of retention (rather than dropping out), the same interpretations would obtain. For example, $\widehat{\pi}_1 = 1 - .17 = .83$, and for the control group, $\widehat{\pi}_2 = 1 - .21 = .79$. Thus, the odds of staying in school for the treatment group is 4.88, and the odds of staying in school for the control group is 3.76. The OR is $\text{odds(treatment)}/\text{odds(control)} = 4.88/3.76 = 1.30$, which is also the inverse of the original OR, 1/0.77. For a student in the intervention group, the odds of staying in school is 1.30 times the odds for a similar student in the control group. Relative risk for modeling the event of staying in school is 0.83/0.79=1.05. Again, the OR overestimates RR.

Only for rare events with a base rate in the population of 0.10 or less would the OR and RR coincide. Zhang and Yu (1998) provide a correction method that compensates for this

overestimation and that should be applied if the researcher chooses to use language reflective of a relative risk interpretation of the OR.

Overall, researchers have a responsibility to correctly interpret their statistical results. Statistical tests of the treatment effect in LR are based on the odds ratio, not relative risk. Misleading interpretations should call to question the validity of reported findings, even in a well-executed research study! We urge researchers to interpret the OR as a true odds ratio rather than a comparison of risks, although determinations of substantive importance would likely benefit from reporting of causal effects in terms of the OR, its confidence interval, and the RR.

Statistical tests and model fit

Valid interpretation of a causal effect hinges on the establishment of a model for the data that truly captures the scientifically relevant relationships between the predictors—including the treatment variable—and the outcome. To build support for a given logistic regression model, two kinds of statistical tests are relevant. The first, an improvement-in-fit test, is a Chi-square likelihood ratio test that compares the likelihood of a "constrained" model to the likelihood of a model without the constraints. A statistically significant result suggests improvement in fit relative to the reduced model. The likelihood ratio test can be used to assess the contribution of a single predictor or a block of predictors. Other types of tests for examining contribution of a single predictor within a multivariable logistic regression model include the Wald and score tests. However, the likelihood ratio test is considered the most reliable test, particularly for small samples, samples with sparse data, or models containing continuous covariates.

The second type of test is goodness-of-fit, which compares the fit of the specified model to a saturated or perfect model that exactly reproduces the original data. With continuous covariates present in a data set, or the presence of sparse cells, or when the number of distinct patterns of covariate combinations is large, the likelihood ratio test to compare the fitted to the saturated model will not follow a Chi-square distribution (Hosmer and Lemeshow 2000). Approximations to the goodness-of-fit test are then used; the most familiar one being the Hosmer–Lemshow test which is based on categorization of the estimated probabilities into deciles of risk. As with all goodness-of-fit tests, a large p value from this test is desired to support quality of fit between observed and predicted values.

Neither the improvement-in-fit nor the goodness-of-fit tests are confirmatory but must be considered under the assumption that the model is correctly specified. A non-statistically significant result for goodness-of-fit is not enough to claim that the fitted model is the "best" model for the data; likewise, the addition of another predictor may be found to "improve" model fit even more. Thus, additional strategies should be put in place to build confidence for the final model that the researcher wishes to promote as causally sound. In addition to a theoretically appropriate model (see "Research design", "Specification of link function", and "Inclusion of all relevant predictors"), these include review of Akaike's or Schwarz's information criteria and calculation of pseudo-R^2 or adjusted pseudo-R^2 statistics (Hosmer *et al*. 1997; Liao and McGee 2003; Menard 2000; Tjur 2009). However, consensus has not been reached among statisticians as to which of the many different forms of these coefficients of determination are most reasonable for logistic regression models.

An additional assessment of model quality for logistic regression can be gauged through accuracy of classification, or predictive efficiency, which is based on the probabilities estimated from the model. Typically, if probability for a case is greater than or equal to 0.50, that case is classified into the "success" event. Cutoffs can be established within most

statistical programs to adjust the classification criteria. As is widely known, reporting percent correct is the least effective way to describe quality of classification and should be avoided in favor of measures that adjust for chance or base rate. Stronger measures of classification accuracy exist and have been reviewed elsewhere (Long 1997; Menard 2000, 2002; O'Connell 2006). It is also possible that predictive efficiency can be poor for a given model, while showing adequate improvement-of-fit and goodness-of-fit. Thus, researchers building support for causal claims should supplement statistical tests with additional criteria for model quality.

Multilevel models for variability of treatment effects

Replication of treatment effects and generalization of research findings are critical elements in building and maintaining support for the existence of a causal effect (NRC 2002). Replication of effect is the "gold-standard" for a causal claim; this is the essence of efforts to identify what works in educational and psychological research, which interventions or programs are worthy of state or federal support on a wide scale for dissemination and which interventions can reliably move from established efficacy to effectiveness in the field. Multilevel designs and analyses can contribute to the science of causal effects through their ability to examine how differences in contexts or settings of an intervention relate to variability in an intervention or treatments' effectiveness (Bingenheimer and Raudenbush 2004; Murray 1998; Raudenbush and Bryk 2002).

The logic of causal inference is the focus of almost all methodological research (Oakes 2004), despite, or perhaps because of, the challenges that exist in supporting causal claims. Multilevel modeling has many advantages, but none that outweigh the importance of research design when building a case for legitimacy of a causal inference. We have discussed the challenges and benefits of matching in terms of balancing covariates between intervention and control groups in individual-level quasi- and non-experimental designs. For multilevel designs, creating plausible matches for clusters such as classrooms, schools, neighborhoods, or clinics becomes more difficult as the cluster itself becomes more complex. Matching typically requires agreement or consensus as to which covariates are the most relevant matching factors, and the collection of potential covariates may be huge (Donner and Klar 2000). Propensity score matching methods can be applied, but the adequacy of selection of variables for inclusion in multilevel propensity methods is further complicated by relationships between variables at either level (Diez Roux 2004). Thus, despite the advantages of multilevel modeling, researchers need to be very clear on the limitations of their multilevel designs and wary of drawing causal inferences when possible alternative explanations for the effects of interest still remain.

Multilevel logistic models for dichotomous and ordinal data are fairly straightforward extensions of their individual-level counterparts (Gelman and Hill 2007; Raudenbush and Bryk 2002; O'Connell *et al*. 2008; Snijders and Bosker 1999). However, estimation of multilevel models for non-normal response data, such as dichotomies, ordinal progressions, or counts, is an active and ongoing area of research, and in particular, the size of variance components can be affected by the estimation procedure utilized. Similarly, residual and model diagnostics for multilevel logistic models are not, as of yet, well conceptualized. Thus, the application of these models to dichotomous event data for the purpose of causal inference must be entered into cautiously. The problems that were noted for individual-level designs regarding sparseness of the data, impact of interactions, and covariate pattern cell size become more acute in the multilevel setting for logistic models.

When a treatment or intervention has been identified and determined as causally responsible for an event (i.e., my school dropout program reduces the incidence of dropout

among adolescents), the external validity of that claim can be bolstered through the expectation that this result will replicate in different settings, schools, or other relevant contexts. Multilevel designs can lend credibility to a causal claim when replication of the intended effect is obtained across different settings. Yet, given the estimation problems for multilevel logistic models noted above, researchers are cautioned against drawing causal inferences based on presence or absence of context-level treatment effect variability (Diez Roux 2004). In fact, replication of an effect may occur when the model is well-specified for the data, but the causal theory is not well represented by the statistical model. Thus, replication of an intervention effect is supportive of causal inference but is not sufficient, and this limitation is true regardless of the sophistication of the particular statistical model employed.

Long-term effects of treatments, policies, or programs have also been examined through multilevel models (e.g., Hser *et al*. 2001), as well as in contexts where randomization was precluded for ethical or practical reasons and where selection bias could be addressed through design and application of propensity scores (Hong and Raudenbush 2005). Thus, multilevel designs hold great promise, particularly for replication of effect, but clear understanding of the limits of these models is warranted.

Summary

We have outlined here seven recommendations for responsible reporting of results when causal effects are assessed through application of logistic regression. As with any research, the quality of the research design used in investigations to assess the existence and measure of causal effects is the single most important criteria in establishing credibility of a research finding. Assuming a solid research design, the consequences of failing to follow the recommendations presented here are variable, but the underlying result of potentially overstating the existence of a causal effect remains the same. Invalid claims about the presence and meaningfulness of a causal effect can set the field back more than if the same study had honestly identified its limitations and considered strategies for building on them.

In order to enhance the validity of research findings based on application of logistic regression models, we urge researchers to pay particular attention to the specific analysis issues we have outlined above and to do so in concert with those of design, thus building stronger support for causal claims. In summary, we hope these recommendations will encourage researchers to consider the consequences as well as the benefits of making prescriptive statements when interpreting statistical results from logistic regression and to continue to work towards a design of research that warrants responsible causal claims. To enhance the plausibility of these claims, multiple design and analysis features must be in place in order to move the field from what may be an important conditional association between cause and event to a reliable and credible causal inference.

References

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). New York: Wiley.

Aldrich, J. H., & Nelson, F. D. (1984). *Linear probability, logit, and probit models*. Newbury Park: Sage.

Amico, K. R. (2008). Percent total attrition: A poor metric for study rigor in hosted intervention sites. *American Journal of Public Health, 99*(9), 1–8.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with commentary). *Journal of the American Statistical Association, 91*, 444–472.
Begg, M. D., & Lagakos, S. (1992). On the consequences of model misspecification in logistic regression. *Environmental Health Perspectives, 87*, 69–75.
Benson, K., & Hartz, A. J. (2000). A comparison of observational studies and randomized, controlled trials. *The New England Journal of Medicine, 342*(25), 1878–1886.
Bingenheimer, J. B., & Raudenbush, S. W. (2004). Statistical and substantive inferences in public health: Issues in the application of multilevel models. *Annual Reviews in Public Health, 25*, 53–77.
Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Lauer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). New York: Academic.
Breslow, N. E. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association, 91*, 14–28.
Cangul, M. Z., Chretien, Y. R., Gutman, R., & Rubin, D. B. (2009). Testing treatment effects in unconfounded studies under model misspecification: Logistic regression, discretization, and their combination. *Statistics in Medicine, 28*, 2531–2551.
Concato, J., Shah, N., & Horwitz, R. I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *The New England Journal of Medicine, 342*(25), 1887–1892.
Dannels, S. A. (2011). Research design. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewers guide to quantitative methods in the social sciences* (pp. 343–356). New York: Routledge.
Davies, H. T. O., Crombie, I. K., & Tavakoli, M. (1998). When can odds ratios mislead? *British Medical Journal, 316*, 989–991.
Des Jarlais, D. C., Lyles, C., Crepaz, N., & The TREND Group. (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement. *American Journal of Public Health, 94*(3)), 361–366.
Diez Roux, A. V. (2004). Estimating health effects: The challenges of causal inference in a complex world. *Social Science & Medicine, 58*, 1953–1960.
Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. London: Arnold.
Fowler, F. J. (2009). *Survey research methods*. Thousand Oaks: Sage.
Fox, J. (2008). *Applied regression analysis and generalized linear models*. Thousand Oaks: Sage.
Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. NY: Cambridge University Press.
Guo, S., & Fraser, M. W. (2010). *Propensity score analysis: Statistical methods and applications*. Thousand Oaks: Sage.
Haddock, C., Rindskopf, D., & Shadish, W. (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. *Psychological Methods, 3*(3), 339–353.
Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine, 15*, 361–387.
Hill, J. (2008). Discussion of research using propensity-score matching: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003'. *Statistics in Medicine, 27*, 2055–2061.
Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–990.
Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis, 27*(3), 205–224.
Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.
Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine, 16*, 965–980.
Hser, Y.-I., Shen, H., Chou, C.-P., Messer, S. C., & Anglin, M. D. (2001). Analytic approaches for assessing long-term treatment effects: Examples of empirical applications and findings. *Evaluation Review, 25*(2), 233–262.
Kish, L. (1995). *Survey sampling*. New York: Wiley (Original work published 1965).
Liao, J. G., & McGee, D. (2003). Adjusted coefficients of determination for logistic regression. *American Statistician, 57*(3), 161–165.
Light, R. J., Singer, J. D., & Willet, J. B. (1990). *By design: Planning research on higher education*. Cambridge: Harvard University Press.
Little, R. J., & Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. *Annual Reviews in Public Health, 21*, 121–145.
Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks: Sage.
McCoach, D. B., Black, A. C., & O'Connell, A. A. (2007). Errors of inference in structural equation modeling. *Psychology in the Schools, 44*(5), 461–470.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models. Monographs on statistics and applied probability, 37*. Boca Raton: Chapman & Hall/CRC.
Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *American Statistician, 54*(1), 17–24.
Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Thousand Oaks: Sage.
Menard, S. (2004a). Six approaches to the calculating standardized logistic regression coefficients. *American Statistician, 58*(3), 218–223.
Menard, S. (2004b). Correction. *The American Statistician, 58*(4), 364.
Moher, D., Schulz, K., & Altman, D. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel group randomized trials. *BMC Medical Research Methodology, 1*, 2.
Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.
National Research Council. (2002). *Scientific research in education*. Washington: Author.
Nemes, S., Jonasson, J., Genell, A., & Steineck, G. (2009). Bias in odds ratios by logistic regression modelling and sample size. *BMC Medical Research Methodology, 9*(1), 56.
Oakes, J. M. (2004). The (mis)estimation of neighborhood effects: Causal inference for a practicable social epidemiology. Social Science & Medicine, *58*, 1929–1952.
O'Connell, A. A. (2000). Sampling for evaluation: Issues and strategies for community-based HIV prevention programs. *Evaluation & the Health Professions, 23*(2), 212–234.
O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables*. Thousand Oaks: Sage.
O'Connell, A. A., & Amico, K. R. (2010). Logistic regression. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewers guide to quantitative methods in the social sciences* (pp. 221–239). New York: Routledge.
O'Connell, A. A., Goldstein, J., Rogers, J., & Peng, C. J. (2008). Multilevel logistic models for dichotomous and ordinal data. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data*. Charlotte: Information Age Publishing.
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.
Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models* (2nd ed.). Newbury Park: Sage.
Riehl, C. (2006). Feeling better: A comparison of medical research and education research. *Educational Researcher, 35*, 24–29.
Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association, 84*(408), 1024–2032.
Rosenbaum, P. R. (1995). *Observational studies*. New York: Springer.
Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science, 17*, 286–327.
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.
Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics, 29*(1), 185–203.
Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association, 74*, 318–328.
Rubin, D. B. (1980). Bias reduction using Mahalanobis-metric matching. *Biometrics, 36*(2), 293–298.
Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*, 169–188.
Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. New Jersey: Wiley.
Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association, 95*, 572–585.
Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects: Using experimental and observational designs*. Washington: AERA.
Shadish, W. R. (2005). Prudent inquiry: Conceptual complexity versus practical simplicity in knowing what works. In J. S. Carlson & J. R. Levin (Eds.), *The no child left behind legislation: Educational research and federal funding* (pp. 129–134). Greenwich: Information Age Publishing.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.
Snijders, T., & Bosker, R. (1999). *Multilevel analysis*. Thousand Oaks: Sage.
Steyerberg, E. W., Harrell, F. E., Jr., Borsboom, G. J. J. M., Eijkemans, M. J. C., Vergouwe, Y., & Habbema, D. F. (2001). Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology, 54*, 774–781.

Tjur, T. (2009). Coefficients of determination in logistic regression models—A new proposal: The coefficient of discrimination. *American Statistician, 63*(4), 366–372.

Towne, L., Wise, L. L., & Winters, T. M. (Eds.). (2004). *Advancing scientific research in education*. Washington: Committee on Research in Education, National Research Council.

Zhang, J., & Yu, K. F. (1998). What's the relative risk?: A method of correcting the odds ratio in cohort studies of common outcomes. *Journal of the American Medical Association, 280*(19), 1690–1691.