

Logistic or Linear? Estimating Causal Effects of Experimental Treatments on Binary Outcomes Using Regression Analysis

Robin Gomila
Princeton University

When the outcome is binary, psychologists often use nonlinear modeling strategies such as logit or probit. These strategies are often neither optimal nor justified when the objective is to estimate causal effects of experimental treatments. Researchers need to take extra steps to convert logit and probit coefficients into interpretable quantities, and when they do, these quantities often remain difficult to understand. Odds ratios, for instance, are described as obscure in many textbooks (e.g., Gelman & Hill, 2006, p. 83). I draw on econometric theory and established statistical findings to demonstrate that linear regression is generally the best strategy to estimate causal effects of treatments on binary outcomes. Linear regression coefficients are directly interpretable in terms of probabilities and, when interaction terms or fixed effects are included, linear regression is safer. I review the Neyman-Rubin causal model, which I use to prove analytically that linear regression yields unbiased estimates of treatment effects on binary outcomes. Then, I run simulations and analyze existing data on 24,191 students from 56 middle schools (Paluck, Shepherd, & Aronow, 2013) to illustrate the effectiveness of linear regression. Based on these grounds, I recommend that psychologists use linear regression to estimate treatment effects on binary outcomes.

Keywords: binary outcomes, logistic regression, linear regression, average treatment effects, causal effects

Psychology research often targets binary outcomes, commonly defined as dependent variables that can take two possible values: 0 and 1. For instance, psychologists have explored the validity of people's intuitive judgments (Kahneman & Frederick, 2002), the circumstances that lead them to violate or conform to social norms (Cialdini, Reno, & Kallgren, 1990; Gomila & Paluck, 2020), the determinants of dropping out of therapy (Wierzbicki & Pekarik, 1993), the predictors of college attendance (Brumley, Russell, & Jaffee, 2019), or the influence of defendants' and victims' race on the likelihood that the defendant receives a death sentence (Eberhardt, Davies, Purdie-Vaughns, & Johnson, 2006).

Psychology researchers rely extensively on nonlinear models such as logit and probit when the outcome is binary, although advances in statistics and methods have established that this is often not optimal, justified or appropriate (Angrist & Pischke, 2009; Freedman, 2008; Hellevik, 2009; Woolridge, 2002). In this


article, I summarize the relevant literature on regression analysis of binary outcomes, and I use analytical, simulation, and empirical approaches to demonstrate the effectiveness of simple and multiple linear regression to estimate treatment effects on binary outcomes.¹ This is true independently of the sample size and distribution of the binary outcome variable (Judkins & Porter, 2016), and in the context of experimental and quasi-experimental designs.²

There are several reasons to prefer linear regression to nonlinear models such as logit and probit when the outcome is binary. Linear regression allows for direct interpretation of the coefficients as probabilities, and is safe when the model includes fixed effects or interaction terms. On the contrary, logit and probit coefficients are not immediately interpretable. Converting them into probabilities requires the additional complexity of methods such as marginal standardization, prediction at the means, or prediction at the modes (Angrist & Pischke, 2009; Freedman, 2008; Muller & MacLehose, 2014). Furthermore, nonlinear models such as logit and probit become unsuitable in the presence of interaction terms or fixed effects (i.e., nested models) (Beck, 2018; Freedman, 2008).

First, I review the arguments commonly used to support the use of nonlinear modeling approaches to analyze binary outcomes, and discuss the relevance of these arguments for psychologists. Second, I examine the advantages of using linear regression to esti-

A preprint of this article was posted on PsyArXiv: <https://psyarxiv.com/4gmbv/>. A draft of this article was discussed in Betsy Levy Paluck's lab meeting. I thank Betsy Levy Paluck, Alexander Coppock, Jasper Cooper, Brandon Stewart, Cyrus Samii, Andi Zhou, Yang-Yang Zhou, Chelsey Clark, Roni Porat, Hannah Alarian, Alex Hayes, and John-Henry Pezzuto for their generous feedback on this article.

Materials and codes: Simulations and analyses reported in this article were computed in R. The R codes can be found on the Open Science Framework (OSF): <https://osf.io/ugsnm/>.

Correspondence concerning this article should be addressed to  Robin Gomila, Department of Psychology, Princeton University, 325 Peretsman-Scully Hall, Princeton, NJ 08544. E-mail: rgomila@princeton.edu

¹ When the outcome is binary, linear regression is commonly referred to as the linear probability model.

² For multiple linear regression models with binary outcomes, it is best to use covariates that are categorical and sparse (Woolridge, 2002). This is generally the case of the typical covariates used by psychologists (e.g., gender, religiosity, or income 5-point scales, etc.).

mate causal effects of any variables with any distribution on binary outcomes. Third, I introduce the framework of potential outcomes through the Neyman-Rubin causal model to prove analytically that linear regression yields unbiased estimates of causal effects of treatments, even when the outcome is binary. Finally, I conduct analyses on simulated data as well as existing data to establish linear regression as a simple, flexible and powerful analysis strategy when the outcome is binary. The advantages of using linear instead of logistic regression to analyze experimental data with binary outcomes are summarized in Table 1.

Arguments Supporting Nonlinear Approaches

Nonlinear Models Are Necessary in the Context of Prediction

The main argument in favor of nonlinear models such as logit or probit to analyze binary outcomes is that these models constrain predictions between 0 and 1. That is, contrary to ordinary least squares (OLS), these methods prevent the analyst from making impossible forecasts, such as predicting that the probability to observe an event is less than 0 or greater than 1. For instance, unlike logit or probit, linear regression could lead one to predict that the probability that an individual commits a crime in their lifetime is 1.2, or $-.04$. As a result, nonlinear modeling strategies often constitute the only valid option for researchers interested in modeling the data to predict the probability of occurrence of binary variables. This is often the case in other disciplines such as biomedical research or finance.

The Presence of Predictions Outside of the Interval Unit Leads to Biased and Inconsistent Estimates of Parameters

A different but related argument in favor of nonlinear models is that, specifically because of predictions outside of the interval unit of the outcome variable, OLS coefficients may be biased and inconsistent in the case of binary outcomes. Hoxby and Oaxaca (2006) demonstrated that bias and inconsistency of the estimator increase with the proportion of predicted probabilities that fall outside of the support, and recommend the use of logit or probit models.

Binary Outcomes Impose Heteroskedasticity in Violation of OLS Assumptions

One of the assumptions of OLS regression is homoskedasticity, which holds if the variance of the error term is the same for all values of X . Mathematically, $\text{Var}(\epsilon|X) = \sigma^2$, in which σ is a constant. When this OLS assumption is violated, errors are considered heteroskedastic, which biases the standard errors of the OLS estimates because too much weight is given to some portion of the data.

The presence of heteroskedasticity in the context of binary outcomes can be understood by looking at the variance formula. If we regress a binary variable Y on k variables $X_1, X_2, X_3, \dots, X_k$, the conditional mean and variance of Y are expressed below, and Equation 2 implies that we observe heteroskedasticity unless the coefficients $\beta_0, \beta_1, \dots, \beta_k$ are all equal to 0.

$$E[Y|X] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (1)$$

$$\text{Var}(Y|X) = X\beta(1 - X\beta) \quad (2)$$

Relevance of These Arguments for Psychology Research

Prediction Versus Explanation in Psychology Research

For the most part, psychology research has focused on explaining causal mechanisms that give rise to psychological and behavioral outcomes, not on making predictions about these outcomes (Yarkoni & Westfall, 2017). As a result, for most psychologists, the question of whether using linear regression is appropriate when outcomes are binary becomes: to what extent do out-of-bound predictions from linear regression bias estimates of causal effect? Broadly speaking, the answer to this question is that psychologists do not need to worry about out-of-bound predictions. To understand why, I now turn to examine different study designs.

The Case of Simple Regression in Experimental Designs

To estimate causal effects of treatments on outcomes, psychology researchers mostly use experimental designs, in which participants are randomly assigned to a treatment versus control condi-

Table 1

Comparison of the Attributes of Linear and Logistic Regression for the Analysis of Experimental Data

Desirable attributes of the analytic strategy	Linear regression	Logistic regression
Overall interpretability of coefficients or commonly reported estimates	Yes	No
Provides immediate estimate of the average treatment effect (ATE) in terms of probability of change	Yes	No
Interpretability of interaction terms	Yes	No
Appropriate for models including fixed effects	Yes	No
Predictions constrained to the interval unit [0, 1]	Yes ^a	Yes
Unbiasedness	Yes	Yes
Consistency	Yes	Yes
Robustness to heteroskedasticity	Yes ^b	Yes

Note. The content of this table applies to both simple and multiple regression analysis.

^a As discussed in more details in this article, predictions outside the interval unit sometimes occur in special cases of multiple linear regression: when models are not saturated. This is the case of models that include continuous covariates (e.g., age). ^b Using robust standard errors.

tion. In this framework, researchers generally use simple linear regression of a binary outcome Y_i on a treatment D_i .

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon \quad (3)$$

in which i denotes individuals. In the context of Equation 3, the average treatment effect (ATE)³ is equal to β_1 , directly expressed in terms of probabilities. In the context of Equation 3, which regresses a binary outcome on a binary treatment, linear regression cannot possibly yield out-of-bound predictions, and always provides an unbiased estimate of the causal effect D_i on Y_i (Angrist & Pischke, 2009). The analytical proof of this argument is provided later in this article.

Linear Regression Analysis Including Discrete Covariates

Psychologists typically include covariates (e.g., gender, religiosity, ethnicity) in their regression for two possible reasons. In experimental studies, the presence of covariates increase statistical power and precision (Green & Aronow, 2011; Maxwell & Delaney, 2004). In the context of quasi-experimental designs (i.e., in the absence of random assignment to the treatment condition) such as natural experiments, covariates may be included in the linear regression analysis to account for preexisting differences between groups (Shadish, Cook, & Campbell, 2002, p. 166–170), allowing to obtain approximately unbiased estimates of causal effects.

Multiple linear regression analysis aiming to estimate causal effects of binary treatments on binary outcomes are perfectly suited, as long as covariates are discrete and take on only few values (Woolridge, 2002, p. 456). It is noteworthy that when models are saturated,⁴ which is the case of the simple linear regression model described in Equation 3 as well as fully interacted models with binary covariates, their underlying structure is inherently linear⁵ (Woolridge, 2002). This implies that for saturated models, the linear regression estimator is unbiased and consistent, and predicted values are never out-of-bounds.

Linear Regression Analysis Including Continuous Covariates

The underlying structure of models that include continuous covariates, commonly referred to as the conditional expectation function (CEF), is often nonlinear. This raises the question of the pertinence of linear modeling strategies in these circumstances. As pointed out by Woolridge (2002), linear regression analysis usually provides a good approximation of the effects of any variable X on the binary outcome of interest near the center of the distribution of X .

Undeniably, OLS will not approximate the CEF of an unknown nonlinear CEF. However, there is no reason to believe that logit or probit models constitute the correct approximation of a given CEF. Analysts willing to model nonlinear CEFs in the context of binary outcomes may use clustering methods such as Bernoulli mixture models. These unsupervised learning techniques can be powerful, but are beyond the scope of this article.

Violation of the Homoskedasticity Assumption

As already mentioned, binary outcomes impose heteroskedasticity, which constitute a violation of one of the OLS assumptions.

First, as pointed out by Angrist and Pischke (2009), the OLS assumption of homoskedasticity is generally violated in the real world, even in the case of nonbinary outcomes (Angrist & Pischke, 2009, p. 46). The implications of the homoskedasticity assumption are related to the calculation of standard errors. If the variance of the error term ϵ differs for different values of X , regular standard errors overweight some portions of the data. In order to resolve this common issue (Angrist & Pischke, 2009), researchers have increasingly used heteroskedasticity-robust standard errors, which are valid even in the context of arbitrary heteroskedasticity (Woolridge, 2002, p. 56).

Reasons to Prefer Linear Regression Analysis

Target Estimands and Interpretability

When the target quantity of interest, also referred to as *estimand* (Rubin, 1974, 1977), is the average causal effect of a treatment variable on a binary outcome, linear regression is the optimal strategy. OLS coefficients allow for a direct interpretation of the treatment effect in terms of the percentage point change in the probability to observe $Y_i = 1$. For instance, if Equation 3 yields $\beta_1 = .01$, we immediately understand that the treatment caused an increase of 1 percentage point in the probability to observe $Y_i = 1$.

The coefficients of nonlinear models are never directly interpretable. Logistic regression coefficients, for instance, are on the log-odds scale,⁶ which implies that they are interpretable in terms of signs and statistical significance, but not effect size. As a result, they are often expressed in terms of odds ratios (ORs), which are also difficult to decipher. As described by Hellevik (2009) “an odds is the ratio between the probability of having a certain value on a variable, and the probability of not having this value . . . what the odds ratio shows, is the ratio between odds, not between proportions” (Hellevik, 2009, p. 66). This definition makes it clear that communicating study results in terms of ORs makes their interpretation complex. As pointed out by King and Zeng (2002), “the disadvantage of odds ratios is understanding what it means . . . we have found no author who claims to be more comfortable communicating with the general public using an odds ratio” (p. 1411). Researchers can take more advanced analytical steps to convert logistic regression coefficients into probabilities. These extra analyses, commonly termed “first differences,” include marginal standardization, prediction at the means, or prediction at the modes, and imply additional assumptions (for a detailed treatment of the question, see Gelman & Hill, 2006; Muller & MacLehose, 2014).

Making the decision to use logistic regression instead of linear regression, both with and without these additional steps, comes with costs. When researchers omit these additional steps, they

³ The average treatment effect is also referred to as the average causal effect.

⁴ A model in which the dependent variable is regressed on a set of binary variables is saturated. Researchers can “saturate their model” by turning any categorical variables that have more than two values (e.g., ethnicity, education) into a set of dummy variables.

⁵ In the case of saturated models, the conditional expectation function of the parameter is linear.

⁶ Probit regression coefficients are on the probit scale.

inevitably place the focus on statistical significance and neglect the importance of effect sizes. This limits one's ability to understand the magnitude of the effect, the practical and theoretical relevance of the results, and use the study for power analyses or meta-analyses in the future (Cohen, 1990; Fritz, Morris, & Richler, 2012). When researchers take these additional steps they impose more restrictions and assumption on the results than with linear regression.

(Mis)conception of Interaction Effects in Nonlinear Models

In nonlinear models such as logit and probit, interaction effects are often misinterpreted because they are conditional on other independent variables (Ai & Norton, 2003; Simonsohn, 2017). This implies that the size and the sign of a given interaction effect from a logistic regression generally varies with the values of other independent variables (Ai & Norton, 2003; Simonsohn, 2017). As a result, the sign of interaction coefficients from nonlinear models does not necessarily indicate the sign of the actual interaction effect of interest and their statistical significance cannot be tested with a simple t test (as is the case in linear regression). In fact, their statistical significance often depends on whether the interaction is conceptualized in terms of probabilities, log odds, or *ORs* (Ai & Norton, 2003). The methods to deal with these issues are debated and remain unclear (Ai & Norton, 2003; Greene, 2010).

Nonlinear Models Do Not Perform Well in the Presence Fixed Effects

Researchers usually include fixed effects in regression analyses to account for the nested nature of the data, such as when students are nested within schools, citizens are nested within community, or workers are nested within companies (e.g., Blair, Littman, & Paluck, 2019; McNeish & Kelley, 2019; Paluck, Shepherd, & Aronow, 2016). Nonlinear models such as logit perform poorly in the presence of fixed effects.⁷ A major issue is that logit models drop all the observations that do not vary in the outcome variable (for an analytical treatment of the question, see Andersen, 1973; Hsiao, 1992). Rodriguez and Goldman (1995) used large numbers of simulations to demonstrate that estimates from logit models with fixed effects are sometimes as biased as estimates from models that ignore the hierarchical structure of the data, which is not the case of linear regression. The greater the number of fixed effects included in the model, the better linear regression fares compared to nonlinear models (Beck, 2018).

It is noteworthy that using logistic regression in the context of nested models with binary outcomes may, under very specific circumstances, be effective. That is, researchers have to use the right (nontraditional) model specifications, such as the Chamberlain's conditional logit (CLOGIT) estimator, which is consistent under some conditions (Beck, 2018). However, using these alternative specifications comes with downsides. The CLOGIT estimator targets a different quantity of interest (estimand), is commonly biased for the ATE,⁸ and because it does not allow to express outcomes in terms of probabilities, results are even more difficult to interpret than traditional logit models. For these reasons, it is clear that using linear regression should generally be chosen in the context of nested models.

Analytical Evidence of the Unbiasedness and Consistency of the OLS Estimator

Does the Fact That an Outcome Is Binary Have any Implications for Causal Analysis of Experimental Data?

I now describe the causal relationship between a variable Y and a treatment D using the Neyman-Rubin causal model (Neyman, 1923; Rubin, 1974, 1977). The Neyman-Rubin causal model is a powerful framework, often used to describe estimation strategies of causal effects in terms of counterfactuals or potential outcomes. Let Y_{1i} denote the outcome if individual i is treated and Y_{0i} the outcome if the same individual i is not treated. The average effect τ_i of a treatment D_i on an outcome Y_i may be different from an individual to another, and can be expressed as:

$$\tau_i = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] \quad (4)$$

Note that in Equation 4 the term $E[Y_{1i}|D_i = 1]$ is observed, but the term $E[Y_{0i}|D_i = 1]$ is unobserved when $D_i = 1$. In this framework, $E[Y_{0i}|D_i = 1]$ is considered an unobserved counterfactual average, assumed to be meaningful. The ATE τ_i is often expressed as the effect of the treatment on the treated in the following way:

$$\tau_i = E[Y_{1i} - Y_{0i}|D_i = 1] \quad (5)$$

Since $E[Y_{0i}|D_i = 1]$ is unobserved, the effect of the treatment on the treated can generally not be identified by comparing the outcomes of D_i . Instead, τ_i is derived by comparing treated and untreated individuals, plus a bias term:

$$\begin{aligned} E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\ &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\ &= E[Y_{1i} - Y_{0i}|D_i = 1] + \underbrace{\{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]\}}_{\text{Bias}} \end{aligned} \quad (6)$$

The Average Causal Effect of the Treatment in Experiments Is Unbiased and Consistent

In experimental designs, D_i and Y_{0i} are independent. Therefore, the ATE τ_i can be expressed as:

$$\tau_i = E[Y_{1i} - Y_{0i}|D_i = 1] = E[Y_{1i} - Y_{0i}] \quad (7)$$

In this case, $E[Y_{1i} - Y_{0i}]$ is usually referred to as the unconditional causal effect of the treatment. This demonstrates that in the experimental framework, the fact that Y_i is binary has no implications, that is, the average causal effect of treatment on a binary outcome is unbiased and consistent. When Y_i is binary, the difference in means $E[Y_{1i} - Y_{0i}]$ corresponds to a difference in probabilities.

⁷ For a recent discussion of the advantages and limitations of fixed effects versus mixed effects models, see McNeish and Kelley (2019). Researchers interested in learning more about using logistic regression with random effects may consult Gibbons and Hedeker (1997); Conaway (1990) or Larsen, Petersen, Budtz-Jorgensen, and Endahl (2000) as a starting point.

⁸ The CLOGIT estimator is biased for the average treatment effect (or any other marginal effect) when effects vary by stratum.

Now in nonexperimental settings, in which identification is based on a selection on observables rather than random assignment, causal inference is based on the assumption that Y_{0i} and D_i are independent, conditional on X_i . In this context, effects must be estimated by conditioning on X_i . We express the effect of treatment as:

$$\begin{aligned}
 & E[Y_{1i} - Y_{0i}|D_i] \\
 &= E\{E[Y_{1i}|X_i, D_i = 1] - E[Y_{0i}|X_i, D_i = 1]|D_i = 1\} \\
 &= \int \{E[Y_{1i}|X_i, D_i = 1] - E[Y_{0i}|X_i, D_i = 0]\} \\
 &\quad \times P(X_i = x|D_i = 1)dx
 \end{aligned}
 \tag{8}$$

In this case, linear regression can be used as a smoothing tool because the population regression coefficients are the best linear approximation to $E[Y_i|X_i, D_i]$, independently of the distribution of $Y_i x$ (Angrist, 2001; Goldberg, 1991). As pointed out by Angrist (2001), “with discrete covariates and a saturated model for X_i , the

additive model can be thought of as implicitly producing a weighted average of covariate-specific contrasts” (Angrist, 2001, p. 7).

Comparison of Linear and Logistic Regression Results Using Simulation

Simulation of Population Data

I generate potential outcomes for six different binary variables ($Y_{01i}, Y_{02i}, \dots, Y_{06i}$) with different baseline probabilities, that is, different probabilities of success $P(Y_0 = 1)$. These variables constitute the binary outcomes of interest for the control condition, and have baseline probabilities varying from 0 to .90 (Figures 1 and 2). I also generate the potential outcomes for the treatment condition: $Y_{11i}, Y_{21i}, \dots, Y_{61i}$. Finally, I generate two covariates X_1 and X_2 . X_1

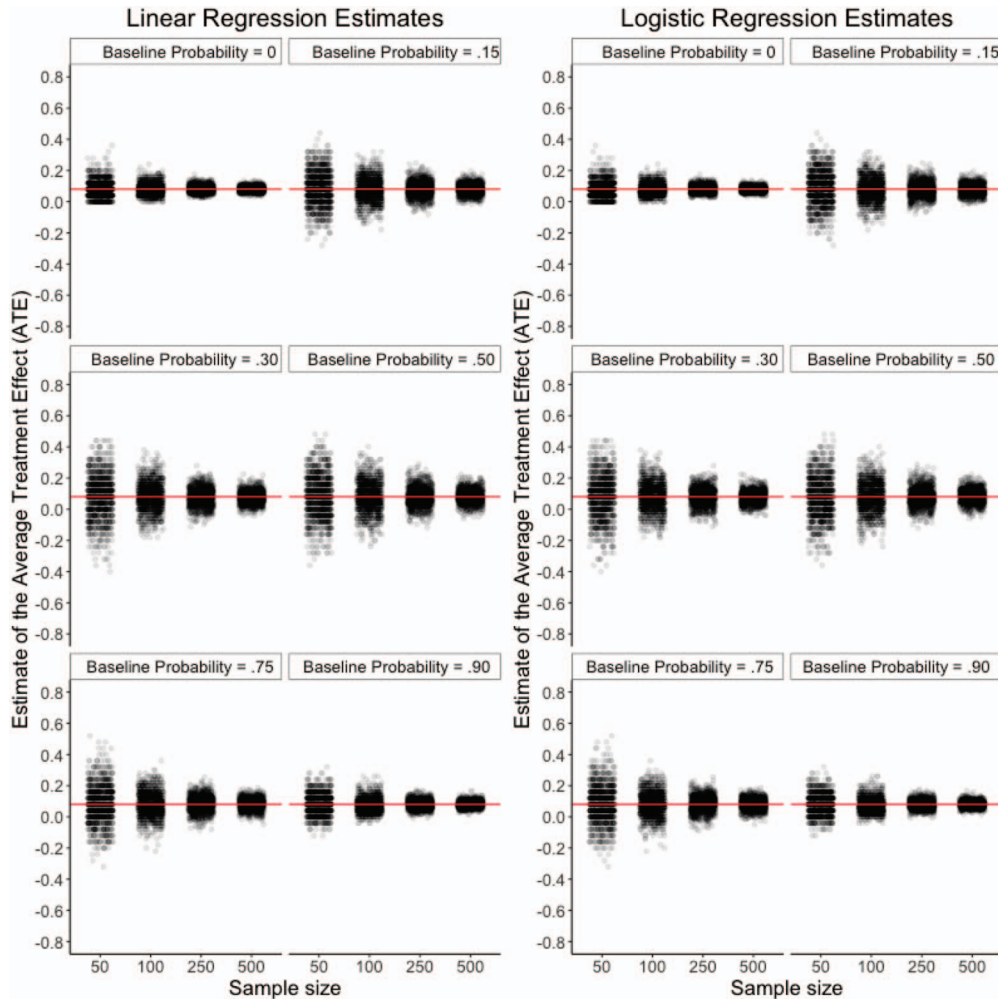


Figure 1. Illustration of unbiasedness and consistency of the simple linear and logistic regression estimators of the average treatment effect (ATE). The red line indicates the ATE ($\tau = .08$) in the simulated population. The left-side panel displays the regression estimates for four different sample sizes ($N_1 = 50, N_2 = 100, N_3 = 250,$ and $N_4 = 500$) and 6 different baseline probabilities ($P_1(Y_i = 1) = 0, P_2(Y_i = 1) = .15, P_3(Y_i = 1) = .30, P_4(Y_i = 1) = .50, P_5(Y_i = 1) = .75, P_6(Y_i = 1) = .90$). For each sample size and baseline probability, the figure displays estimates from 1,000 randomly drawn samples. See the online article for the color version of this figure.

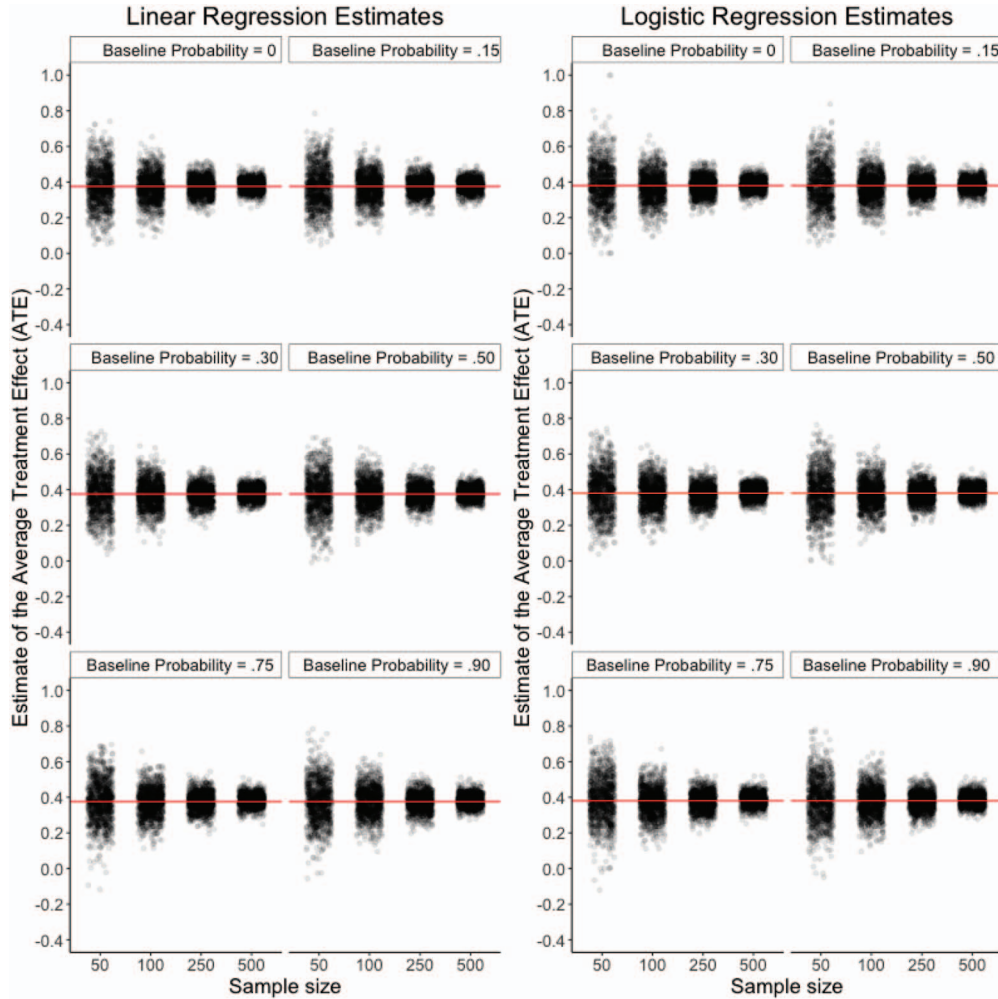


Figure 2. Illustration of unbiasedness and consistency of the multiple linear and logistic regression estimators of the average treatment effect (ATE). The red line indicates the ATE ($\tau = .38$) in the simulated population. The left-side panel displays the regression estimates for four different sample sizes ($N_1 = 50$, $N_2 = 100$, $N_3 = 250$, and $N_4 = 500$) and six different baseline probabilities ($P_1(Y_i = 1) = 0$, $P_2(Y_i = 1) = .15$, $P_3(Y_i = 1) = .30$, $P_4(Y_i = 1) = .50$, $P_5(Y_i = 1) = .75$, $P_6(Y_i = 1) = .90$). For each sample size and baseline probability, the figure displays estimates from 1,000 randomly drawn samples. See the online article for the color version of this figure.

is a binary variable that indicates whether participants have a college degree, and X_2 is a discrete variable on a 5-point scale that indicates the level of religiosity of participants. Simulations are conducted using Declare Design (Blair, Cooper, Coppock, & Humphreys, 2019), and the data generation process for the binary outcomes is described for each simulation below.

Estimation of the ATE in Experiments

In this simulation study, the randomized treatment has an average effect of .08 (i.e., 8 percentage points) on the outcome. I randomly select samples of size $N_1 = 50$ to $N_4 = 500$, randomly assign individuals to the treatment or control condition, put together a variable \bar{Y} that includes the observed outcomes of each individual, and estimate the average causal effect of the treatment using the following simple linear and logistic regression models:

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon$$

$$P(Y_i = 1) = \text{logit}^{-1}(\beta_0 + \beta_1 D_i + \epsilon)$$

The distribution of the ATE, illustrated in Figure 1, demonstrates that both strategies perform equally. That is, both the linear and logistic regression estimators are unbiased and consistent for the ATE.

Estimation of the ATE in Quasi-Experiments

I now consider the multiple regression framework, in which the analyst includes two covariates: the binary variable college degree (X_1) and the discrete variable religiosity (X_2). For a conservative approach, I generate the different potential outcomes Y_0 and Y_1 from a logit model using both covariates, which yields an average treatment effect of .38. Then, I draw random samples from this

dataset, and randomly assign individuals from the original dataset to the treatment versus control conditions. I then estimate the ATE using the following multiple regression models. For the logit model, covariates are held at their mean level.

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_1 + \beta_3 X_2 + \epsilon$$

$$P(Y_i = 1) = \text{logit}^{-1}(\beta_0 + \beta_1 D_i + \beta_2 X_1 + \beta_3 X_2 + \epsilon)$$

The results, illustrated in Figure 2, demonstrate that both estimators are unbiased and consistent for the ATE.

Comparison of Linear and Logistic Regression Results Using Existing Data

Description of the Dataset

I use data from a field experiment examining the impact of an anticonflict intervention in 56 New Jersey middle schools (Paluck, Shepherd, & Aronow, 2013). The dataset includes survey and administrative data on the entire student population of each school, which includes a total of 24,191 students. The 56 schools involved in the study were assigned to blocks of two before being randomly assigned (within blocks) to the control or treatment condition. The researchers deployed a year-long intervention aiming to reduce conflict between students in all 28 treatment schools assigned to the treatment condition. Paluck et al. (2013) collected survey data on all 24,191 students during the school year. In addition, each school provided the researchers with data describing, for each student, instances of disciplinary events such as bullying or other types of peer conflict.

Selection of the Variables and Analytic Strategy

I select a total of 12 different binary variables with varying distributions from this dataset. For each outcome variable, I report the results of linear and logistic regression. Specifically, in addition to reporting the raw regression coefficients and associated p values, I report the estimates that researcher most commonly report for each analytic strategy. That is, I report ORs for logistic regression and the ATE in probabilities for linear regression. These different statistics are derived for the following two models:

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon \tag{9}$$

$$P(Y_i = 1) = \text{logit}^{-1}(\beta_0 + \beta_1 D_i + \epsilon) \tag{10}$$

in which i denotes students. Y_i is the outcome variable, and ϵ_i is the student error term.

The analyses described in Equations 9 and 10 pool all observations, ignoring the nested structure of the data into blocks. For this reason, I subsequently include fixed effects of schools into these linear and logistic regression analyses in the following way:

$$Y_i = \beta_0 + \beta_1 D_i + \gamma_{s[i]} + \epsilon \tag{11}$$

$$P(Y_i = 1) = \text{logit}^{-1}(\beta_0 + \beta_1 D_i + \gamma_{s[i]} + \epsilon) \tag{12}$$

in which i denotes students and $s[i]$ denotes schools. Y_i is the outcome variable, $\gamma_{s[i]}$ is the school fixed effects, and ϵ_i is the student error term.

The comparison between the models from Equations 11 and 12 constitutes an empirical test of the simulation-based findings of

Rodriguez and Goldman (1995), which suggest that in the presence of fixed effects, logistic regression performs poorly in the presence of fixed effects.

Linear and Logistic Regression Results

The results of these two separate comparisons between linear and logistic regression analysis are displayed in Table 2 (models without fixed effects) and Table 3 (models with fixed effects). In both tables, logistic and linear regression analyses produce comparable p values, which indicate that the statistical significance of the findings is not impacted by the selected analytic strategy. The main difference between logistic and linear regression lies in the interpretability of the coefficients or the estimate that researchers typically report.

In Tables 2 and 3, the coefficients and estimates from the linear regression analyses are the exact same and correspond to the average effect of the treatment on each variable, expressed in terms of probability of change. This illustrates how linear regression outputs immediately provide interpretable estimates of the ATE.

In contrast, the logistic regression coefficients are expressed in log of ORs and the estimates, calculated by exponentiating these coefficients, is expressed in ORs . As previously discussed, log of ORs and ORs are difficult to interpret. For instance, the estimated OR of 1.13 for Variable 1 indicates that the odds (which is itself a ratio) for participants exposed to treatment are 1.13 times higher than the odds (another ratio) for participants assigned to the control condition. The cumbersome character of ORs is one reason why ORs make it difficult to assess the theoretical and practical relevance of effects. Another

Table 2
Linear and Logistic Regression Results for Models With Fixed Effects, for 12 Variables From a Large Experimental Dataset

Variable no.	Coefficient		Estimate		p value	
	OLS	Logit	OLS (ATE)	Logit (OR)	OLS	Logit
Var1	0.03	0.17	0.03	1.19	0.00	0.00
Var2	-0.01	-0.03	-0.01	0.96	0.45	0.45
Var3	0.01	0.08	0.01	1.08	0.18	0.18
Var4	0.00	0.02	0.00	1.02	0.64	0.64
Var5	-0.01	-0.06	-0.01	0.95	0.11	0.11
Var6	-0.01	-0.06	-0.01	0.94	0.11	0.11
Var7	-0.00	-0.00	-0.00	1.00	1.00	1.00
Var8	0.01	0.11	0.01	1.11	0.02	0.02
Var9	0.02	0.10	0.02	1.11	0.01	0.01
Var10	0.01	0.04	0.01	1.04	0.17	0.17
Var11	-0.03	-0.14	-0.03	0.87	0.00	0.00
Var12	0.00	0.02	0.00	1.02	0.46	0.46

Note. $N = 24,191$. The two columns under Coefficient display the raw regression coefficients from the linear (ordinary least squares [OLS]) and logistic (logit) regression analyses. Whereas linear regression analysis yields the average treatment effect size in terms of probabilities, logistic regression analysis yields the log of (ORs). The two columns under Estimate display the estimates that researchers typically report after using linear and logistic regression. For linear regression, these estimates correspond to the average treatment effect (ATE), whereas for logistic regression, these estimates are generally ORs . The two columns under p value displays the p values generated by linear and logistic regression for each variable.

Table 3
Linear and Logistic Regression Results for Models With Fixed Effects (14 Clusters), for 12 Variables From a Large Experimental Dataset

Variable no.	Coefficient		Estimate		<i>p</i> value	
	OLS	Logit	OLS (ATE)	Logit (OR)	OLS	Logit
Var1	0.02	0.12	0.02	1.13	0.00	0.00
Var2	-0.01	-0.03	-0.01	0.96	0.44	0.44
Var3	0.00	0.04	0.00	1.04	0.49	0.49
Var4	0.00	0.03	0.00	1.03	0.51	0.50
Var5	-0.01	-0.05	-0.01	0.95	0.19	0.19
Var6	-0.01	-0.04	-0.01	0.96	0.35	0.35
Var7	-0.00	-0.01	-0.00	0.99	0.88	0.88
Var8	0.01	0.10	0.01	1.10	0.04	0.04
Var9	0.02	0.10	0.02	1.11	0.01	0.01
Var10	0.02	0.07	0.02	1.07	0.02	0.02
Var11	-0.03	-0.12	-0.03	0.88	0.00	0.00
Var12	0.01	0.06	0.01	1.05	0.08	0.08

Note. $N = 24,191$. The two columns under Coefficient display the raw regression coefficients from the linear (ordinary least squares [OLS]) and logistic (logit) regression analyses. Whereas linear regression analysis yields the average treatment effect size in terms of probabilities, logistic regression analysis yields the log of (ORs). The two columns under Estimate display the estimates that researchers typically report after using linear and logistic regression. For linear regression, these estimates correspond to the average treatment effect (ATE) whereas for logistic regression, these estimates are generally ORs. The two columns under *p* value display the *p* values generated by linear and logistic regression for each variable.

reason is that the relevance of effects expressed in OR is conditional on the mean (i.e., probability of 1's) of the dependent variable for the control group. Specifically, an OR of 1.13 can correspond to a very small effect if the mean of the dependent variable for the control group is tiny, or to a larger effect if the mean of the dependent variable for the control group is larger. This implies that a same OR can translate into different Cohen's *d* values, depending on the probability of 1's of the dependent variable for control group participants (Chen, Cohen, & Chen, 2010).

As displayed in Figure 1 and 2, it is possible to derive the ATE from logistic regression. And doing so in the context of the analyses displayed in Tables 2 and 3 does yield the exact same estimate of the ATE as linear regression. Researchers who would like to take the extra steps to calculate the ATE in terms of probability of change may use the `predict` or `invlogit` functions in R.⁹ The result that the logistic and linear regression estimators perform equally in this particular setting should not necessarily be generalized to other settings. This dataset includes a small number clusters (i.e., 14), and a large number of observations (i.e., 24, 191 students). The large number of observations in each cluster explains the effectiveness of the logistic regression estimator. Analysts should expect that the performance of logit decreases as the ratio of observations to clusters decreases.

Summary and Conclusions

Psychologists have a long history of using experimental designs, in the lab and in the field, to explain causal effects of treatments.

In the presence of binary outcomes, linear regression analysis is the most powerful, flexible, and the simplest strategy. This is the case for models with and without covariates, and in the presence of adjustments such as interactions or fixed effects. Furthermore, past research suggests that nonlinear models sometimes perform poorly in the presence of fixed effects (though it is not the case in the present analyses of empirical data), and that researchers are often misled by interaction terms from logit and probit regression analyses.

In the presence of binary outcomes, the predominance of nonlinear modeling analysis strategies such as logit and probit in the literature may have negative implications for the field. First, researchers sometimes only report logit or probit regression coefficients. These coefficients are not interpretable, which shifts the focus of interest toward statistical significance, and away from actual effect sizes. Second, researchers often interpret the results of logistic regression in terms of odds-ratios, which are undeniably difficult to interpret. This, once again, focuses the attention onto *p* values, and denies the practical and theoretical importance of effect sizes. On the contrary, linear regression yields results that are immediately interpretable in terms of probability of change, which is the most desirable way to communicate effect sizes.

The analyses of empirical data reported in this article, using logistic and linear regression, illustrate the effectiveness of linear regression to examine causal effects of treatments on binary outcomes. Average causal effects as well as *p* values derived from both methods were the exact same (up to two decimal places). This result is aligned with past research on the correspondence between the *p* values of logistic and linear regressions analyses for sample sizes varying between 200 and 2,500. Hellevik (2009) demonstrated that the correlation between the two sets of *p* values was .9998, and that 90% of time, the difference between the *p* values was less than .005. All in all, choosing linear regression over logit or probit does not involve any tradeoff in terms of statistical significance.

Based on these grounds, I recommend that psychology researchers use linear regression to estimate causal effects of treatments on binary outcomes. This should become the default practice, since there is no apparent reason to use more complex, nonlinear modeling strategies. In the specific case of multiple regression analysis in which the model is not saturated (e.g., when the model includes continuous covariates), different analytic strategies may produce different results. In these circumstances, I recommend that researchers supplement their linear regression analysis with a sensitivity analysis (for a tutorial, see Thabane et al., 2013) to assess the robustness of the findings to other analytic strategies. In their sensitivity analysis, researchers may analyze the data using, for instance, logit or probit models, or clustering methods such as Bernoulli mixture models.

⁹ All of the analyses reported in this article were computed in R. The R codes can be found on the Open Science Framework (OSF): <https://osf.io/ugsnm/>.

References

- Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters*, *80*, 123–129. [http://dx.doi.org/10.1016/S0165-1765\(03\)00032-6](http://dx.doi.org/10.1016/S0165-1765(03)00032-6)
- Andersen, E. (1973). *Conditional inference and models for measuring*. Copenhagen, Denmark: Mentalhygiejnisk forlag.
- Angrist, J. D. (2001). Estimation of limited dependent variable models with dummy endogenous regressors: Simple strategies for empirical practice. *Journal of Business & Economic Statistics*, *19*, 2–16.
- Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.
- Beck, N. (2018). Estimating grouped data models with a binary dependent variable and fixed effects: What are the issues. Retrieved from <https://arxiv.org/abs/1809.06505>
- Blair, G., Cooper, J., Coppock, A., & Humphreys, M. (2019). Declaring and diagnosing research designs. *American Political Science Review*, 1–22. <http://dx.doi.org/10.1017/S0003055419000194>
- Blair, G., Littman, R., & Paluck, E. L. (2019). Motivating the adoption of new community-minded behaviors: An empirical test in Nigeria. *Science Advances*, *5*, eaau5175. <http://dx.doi.org/10.1126/sciadv.aau5175>
- Brumley, L. D., Russell, M. A., & Jaffee, S. R. (2019). College expectations promote college attendance: Evidence from a quasi-experimental sibling study. *Psychological Science*, *30*, 1186–1194. <http://dx.doi.org/10.1177/0956797619855385>
- Chen, H., Cohen, P., & Chen, S. (2010, March). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics - Simulation and Computation*, *39*, 860–864.
- Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*, 1015–1026. <http://dx.doi.org/10.1037/0022-3514.58.6.1015>
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304–1312. <http://dx.doi.org/10.1037/0003-066X.45.12.1304>
- Conaway, M. R. (1990). A random effects model for binary data. *Biometrics*, *46*, 317–328. Retrieved from <https://www.jstor.org/stable/2531437>
- Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy: Perceived stereotypicality of black defendants predicts capital-sentencing outcomes. *Psychological Science*, *17*, 383–386. <http://dx.doi.org/10.1111/j.1467-9280.2006.01716.x>
- Freedman, D. A. (2008). On regression adjustments in experiments with several treatments. *The Annals of Applied Statistics*, *2*, 176–196. <http://dx.doi.org/10.1214/07-AOAS143>
- Fritz, C. O., Morris, P. E., & Richler, J. J. (2012, February). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*, 2–18. <http://dx.doi.org/10.1037/a0024338>
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel-hierarchical models*. New York, NY: Cambridge University Press.
- Gibbons, R. D., & Hedeker, D. (1997). Random effects probit and logistic regression models for three-level data. *Biometrics*, *53*, 1527–1537. <http://dx.doi.org/10.2307/2533520>
- Goldberg, A. S. (1991). *A course in econometrics*. Cambridge, MA: Harvard University Press.
- Gomila, R., & Paluck, E. L. (2020). The social and psychological characteristics of norm deviants: A field study in a small cohesive university campus. *Journal of Social and Political Psychology*, *8*, 220–245.
- Green, D. P., & Aronow, P. M. (2011). Analyzing experimental data using regression: When is bias a practical concern? *SSRN Electronic Journal*. Advance online publication. <http://dx.doi.org/10.2139/ssrn.1466886>
- Greene, W. (2010). Testing hypotheses about interaction terms in nonlinear models. *Economics Letters*, *107*, 291–296. <http://dx.doi.org/10.1016/j.econlet.2010.02.014>
- Hellevik, O. (2009). Linear versus logistic regression when the dependent variable is a dichotomy. *Quality & Quantity*, *43*, 59–74. <http://dx.doi.org/10.1007/s11135-007-9077-3>
- Horrace, W. C., & Oaxaca, R. L. (2006). Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters*, *90*, 321–327. <http://dx.doi.org/10.1016/j.econlet.2005.08.024>
- Hsiao, C. (1992). Logit and probit models. In L. Matyas & P. Sevestre (Eds.), *The econometrics of panel data: Handbook of theory and applications*. Norwell, MA: Kluwer Academic.
- Judkins, D. R., & Porter, K. E. (2016). Robustness of ordinary least squares in randomized clinical trials. *Statistics in Medicine*, *35*, 1763–1773. <http://dx.doi.org/10.1002/sim.6839>
- Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). New York, NY: Cambridge University Press.
- King, G., & Zeng, L. (2002). Estimating risk and rate levels, ratios and differences in case-control studies. *Statistics in Medicine*, *21*, 1409–1427. <http://dx.doi.org/10.1002/sim.1032>
- Larsen, K., Petersen, J. H., Budtz-Jorgensen, E., & Endahl, L. (2000). Interpreting parameters in the logistic regression model with random effects. *Biometrics*, *56*, 909–914.
- Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective, 2nd ed.* Mahwah, NJ: Erlbaum.
- McNeish, D., & Kelley, K. (2019). Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations. *Psychological Methods*, *24*, 20–35. <http://dx.doi.org/10.1037/met0000182>
- Muller, C. J., & MacLehose, R. F. (2014). Estimating predicted probabilities from logistic regression: Different methods correspond to different target populations. *International Journal of Epidemiology*, *43*, 962–970. <http://dx.doi.org/10.1093/ije/dyu029>
- Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. *Statistical Science*, *5*, 465–472. <http://dx.doi.org/10.1214/ss/1177012031>
- Paluck, E. L., Shepherd, H., & Aronow, P. M. (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 566–571. <http://dx.doi.org/10.1073/pnas.1514483113>
- Paluck, E. L., Shepherd, H., & Aronow, P. (2013). *Changing climates of conflict: A social network experiment in 56 Schools*. New Jersey: Ann Arbor, MI: Inter-University Consortium for Political and Social Research [distributor]. <http://dx.doi.org/10.3886/ICPSR37070.v1>
- Rodriguez, G., & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, *158*, 73–89. <http://dx.doi.org/10.2307/2983404>
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701. <http://dx.doi.org/10.1037/h0037350>
- Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, *2*, 1–26. <http://dx.doi.org/10.2307/1164933>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton, Mifflin and Company.
- Simonsohn, U. (2017). *Interactions in logit regressions: Why positive may mean negative*. Retrieved from <http://datacolada.org/57>
- Thabane, L., Mbuagbaw, L., Zhang, S., Samaan, Z., Marcucci, M., Ye, C., . . . Goldsmith, C. H. (2013). A tutorial on sensitivity analyses in clinical trials: The what, why, when and how. *BMC Medical Research Methodology*, *13*, 92. <http://dx.doi.org/10.1186/1471-2288-13-92>

Wierzbicki, M., & Pekarik, G. (1993). A meta-analysis of psychotherapy dropout. *Professional Psychology: Research and Practice, 24*, 190–195. <http://dx.doi.org/10.1037/0735-7028.24.2.190>

Woolridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psycho-*

logical Science, 12, 1100–1122. <http://dx.doi.org/10.1177/1745691617693393>

Received September 17, 2019

Revision received February 13, 2020

Accepted May 18, 2020 ■