False discovery in the analysis of interactions as a function of the distributional and metric properties of the outcome

Benjamin W. Domingue[1,†], Klint Kanopka[2], Sam Trejo[3], and Elliot M. Tucker-Drob[4]

[1]Graduate School of Education, Stanford University & Center for Population Health Sciences, Stanford Medicine
[2]Graduate School of Education, Stanford University
[3]La Follette School of Public Affairs & Department of Sociology, University of Wisconsin–Madison
[4]Department of Psychology & Population Research Center, University of Texas at Austin
[†]Correspondence about the paper should be sent to ben.domingue@gmail.com.

Abstract

Lived experience suggests substantial heterogeneity in how people react to a common stimuli. Previous work has considered several potential sources of bias and confusion in studying interactions but relatively less attention has been devoted to the nature of the outcome variable in such studies. Here, we consider power and false discovery associated with estimates of interaction parameters as a function of the distributional and metric properties of the outcome. Focusing on a variety of models for non-continuously distributed outcomes (binary, count, and ordinal outcomes), we show that power analyses need to carefully attend to specific details of a given situation and that attempts to use the linear model for recovery can be catastrophic in some settings. Focusing on transformations of normally distributed variables (i.e., censoring and departures from interval scaling) we show that linear models produce spurious interaction effects when such effects are absent from the generating model. We also provide illustrations offering geometric intuition as to why interactions can be a challenge for these incorrectly specified linear models. In light of these findings, we make two specific suggestions. First, a careful consideration of the distributional properties of the outcome variable should be a standard component in interaction studies. Second, hand-tailored power calculations should be provided; to that end, we have written software to help researchers scrutinize their ability to study interactions in given data contexts.

1 Introduction

Lived experience suggests substantial heterogeneity in how people react to a common stimuli.[1] This suggests the hypothesis that some feature (of the stimuli, environment, or person) may explain this heterogeneity. For example, how do demographic characteristics interact with exercise to affect quality of life among people with cancer [2]. Which psychological factors account for variation in the success of smoking cessation programs [3]? Do men and women differentially respond to the addition of chemotherapy to a baseline lung cancer treatment [4]? These few examples are a small portion of the larger literature probing for such heterogeneities in many empirical settings.

The standard tool for analysis of such heterogeneity is the inclusion of interaction terms in statistical models. However, analysis of such models is fraught. We are not the first to note this; we briefly discuss other methodological issues that have been previously discussed in Section 1.1. Here, we focus on issues arising due to specific measurement properties of the outcome variables. Statistical models for interactions rely upon strong assumptions regarding the distribution and measurement properties of the outcome variable. When these assumptions are not met, conventional regression models may produce biased interaction effect estimates and high rates of false discovery (when there is, in fact, no interaction), and properly specified models may have relatively large sample size requirements so as to obtain adequate power.

[1] This is a paraphrase of what we have heard Jeremy Freese call the First Law of Sociology—“Some do, some don’t”—but psychologists also utilize it [1].

Here, we explore the rate of discovery (as indicated via tests of statistical significance) when there is a true interaction effect and when there is not as a function of the nature of the outcome variable and the statistical model used to study it. When there is a true interaction effect, we are effectively asking about statistical power. Considerations of power are crucial given that they are informative for study design and frequently provided in proposals to justify, for example, sample sizes.[2] When there is no interaction, we are effectively studying the false discovery rate (FDR). A certain amount of false discovery is to be expected (the level of false discovery being a function of the α value of the associated test in a properly calibrated model); however, higher levels of false discovery are symptomatic of a biased parameter estimate. As we shall see, we frequently observe such a scenario. Alongside these results, we introduce a new piece of software—an `R` package, `InteractionStudio`—that has many features designed to help improve future studies of interactions.[3] Before discussing our work, we briefly cover the broader literature focused on interactions. We discuss alternative problems (i.e., those not related to the outcome) in study of interactions. We then describe a systematic review of interaction studies in the context of a specific outcome variable to motivate many of the issues we cover in the paper. We then turn to a discussion of the simulation strategy used here.

1.1 Previously expressed concerns regarding analysis of statistical interactions

Below we briefly summarize many of the key methodological points that have been previously made about the study of interactions. We assume that interest is in the interaction of two predictors (x and z) in study of some outcome (y); that is, interest is in estimates of β_3 in a model of the form $y = f(\beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz)$.

- Studies examining interaction may yield false positives if interactions between additional covariates (w) are not also included in the model [5]. If w is a covariate of interest, then analysis of the interaction xz should be based on models that also include xw and zw.
- If the true data generating model is based on x^2, interaction studies focusing on xz may lead to false positives if x and z are correlated [6, 7].
- The substantive implications of findings of interactions may be highly contingent on the nature of x and z. In particular, endogeneity may lead to multiple viable interpretations [8].

We do not aim for a comprehensive survey of previous work but rather highlight these issues for two purposes. First, they help to emphasize that there are numerous challenges to the statistical analysis of interactions. Second, the previous scholarship on study of interactions has not focused as much on characteristics of the outcome variable. As we show below, these characteristics can have major implications for our ability to recover the relevant parameters.

1.2 Key contributions

We focus on describing power and FDR in two key settings. First, how is power affected when outcomes are non-continuous? We show that power can be affected by assumptions (e.g., correlations between predictors) in ways that suggest a need for tailored power analysis. We also show that application of OLS techniques result in either less power or disastrous levels of FDR. Second, how is analysis affected when outcome variables do not meet classical assumptions of OLS? We show that when outcomes are transformed versions of classic linear outcomes there can be undesirable levels of false discovery. Evidence from both settings suggest a need for heightened scrutiny of interaction studies when the outcome does not allow for OLS analysis. Before turning to those scenarios, we start in Section 2 by introducing the simulation strategy that we use throughout.

2 Our simulation strategy

So as to illustrate the problems we describe below, we consider a running example. In this example, we observe some outcome y and interest is in predictors x and z as well as their interaction. We'll assume that

[2] For example, `https://researchmethodsresources.nih.gov/grt-calculator`.

[3] `https://github.com/ben-domingue/InteractionStudio/wiki`

$(x_i, z_i) \sim \text{MVN}(\mu, \Sigma)$ where $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. We consider sample sizes of 250, 1000, and 5000.[4] We define a quantity, Γ, that plays a key role throughout. This quantity is defined as

$$\Gamma_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i. \tag{1}$$

If $\mathbb{E}(y_i|x_i, z_i) = \Gamma_i$, then the linear model is appropriate.

We focus here on more complex cases wherein $\mathbb{E}(y_i|x_i, z_i) \neq \Gamma_i$. In Section 3, we consider scenarios wherein $\mathbb{E}(y_i|x_i, z_i) = g(\Gamma_i)$. Suppose, for example, that we are interested in binary outcomes; in that case, we may assume that $\mathbb{E}(y) = \sigma(\Gamma_i)$ where σ is the logistic sigmoid, $\sigma(x) = 1/(1 + e^{-x})$. In Section 4, we suppose that $\mathbb{E}(y^\star|x, z) = \Gamma_i$ but instead of observing $y^\star$ we observe some transformed version y. In all cases, interest will focus on estimates of β_3; note that power calculations based on given values of this quantity will not necessarily be comparable across the various scenarios we consider so we instead focus on general trends. When recovery models are properly specified (i.e., $\mathbb{E}(y|x, z) = g(\Gamma)$), we focus on statistical power as a function of β_3 and other relevant variables (i.e., ρ). We rely on the conventional statistical significance level ($\alpha = 0.05$). When recovery models are misspecified as linear models (i.e., we assume $\mathbb{E}(y|x, z) = b_0 + b_1 x + b_2 z + b_3 xz$), we will consider the false discovery rate when $\beta_3 = 0$.

3 Transformations of Γ

In this section, we first consider outcomes belonging to the GLM family [9] with non-identity link (binary and count outcomes). We then consider ordinal outcomes that do not fit within the GLM family. In all cases, the observations are dependent on Γ_i and we focus on the discovery rate for both the properly specified model as well as the linear alternative.

3.1 Binary outcomes

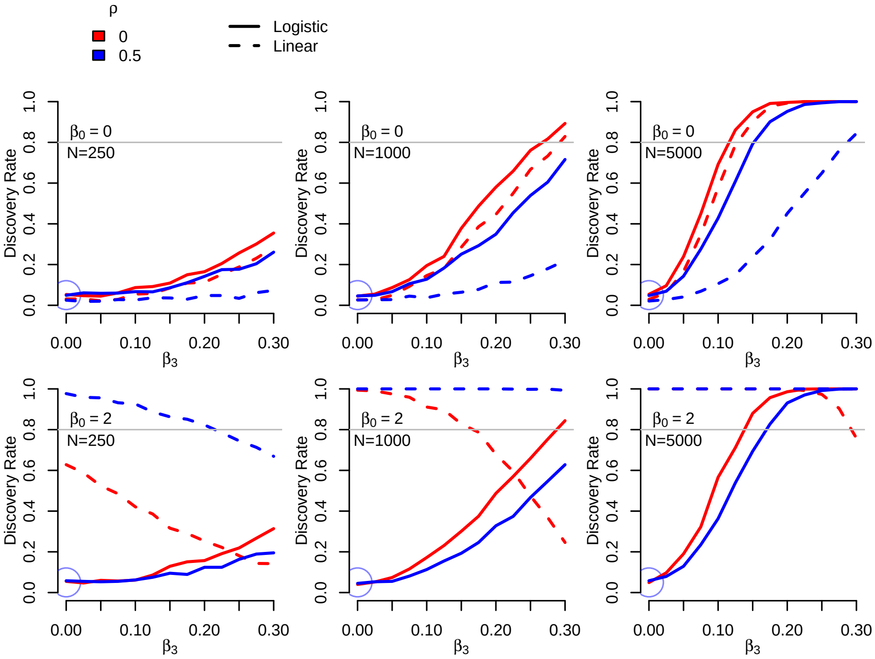
Suppose y_i is a binary outcome; this is, of course, a common analytic scenario [10, 11, 12]. Analysis is based on the assumption that $y_i \sim \text{Bernoulli}(\sigma(\Gamma_i))$; i.e., we consider the logistic regression model (i.e., a GLM with logit link) for estimation. Results are shown in Figure 1. When $\beta_3 = 0$, FDR is appropriate at 0.05 (i.e., the solid lines are in the blue circle). Power behaves as expected with respect to sample size but consider the role of the parameters. Consider first the role of β_0. Note that power is enhanced when $\beta_0 = 0$ relative to $\beta_0 = 2$ (i.e., power is reduced when prevalence is further from 0.5). This suggests a need for attention to the prevalence of binary outcome in considerations of power. Turning to the role of ρ, power is reduced when $\rho > 0$.

We also consider recovery based on the linear probability model (that is, we ignore the fact that y_i is binary and assume that $y_i \sim \text{Normal}(\Gamma_i, \sigma_y^2)$). We have discussed this issue in a specific context elsewhere [13, 14]; interaction analysis of highly prevalent outcomes (or, symmetrically, relatively rare outcomes) with the linear probability model is a flawed approach. To illustrate the underlying problem of geometry, consider Figure 2 which shows $\mathbb{E}(y)$ as a function of x for simulated data when $\beta_3 = 0$ along with the implied logistic and linear fits for two different values of z ($z = \pm 1$). We shade data points with $|z - 1| < 0.5$ in blue and with $|z + 1| < 0.5$ in red (fitted lines for ± 1 are similarly shaded). When $\beta_0 = 0$, the geometry of the $\mathbb{E}(y)$ values is such that both the logistic and linear fits for $z = \pm 1$ are clearly parallel (i.e., suggest $\beta_3 = 0$). However, as β_0 increases, the geometry of the $\mathbb{E}(y)$ values is such that there is an implied interaction. This is due to the fact that the $z = 1$ values in blue are observed at a location where the true underlying sigmoid is quite flat thus leading to an inference that $\beta_3 < 0$.[5]

[4] To inform this choice, we surveyed 150 studies published between January 1, 2016 and October 20, 2020 focused on depression (Based on the Pubmed query: `(depression[Title]) AND (heterogeneity[Title] OR moderation[Title] OR interaction[Title])`) Of the ascertained studies, 30% were empirical studies along the lines of what we consider here. The sample sizes ranged widely (min=22, median=576, IQR=292-2279, max=134357); our choice of 500–5000 is meant to reflect the right-hand side of this distribution.

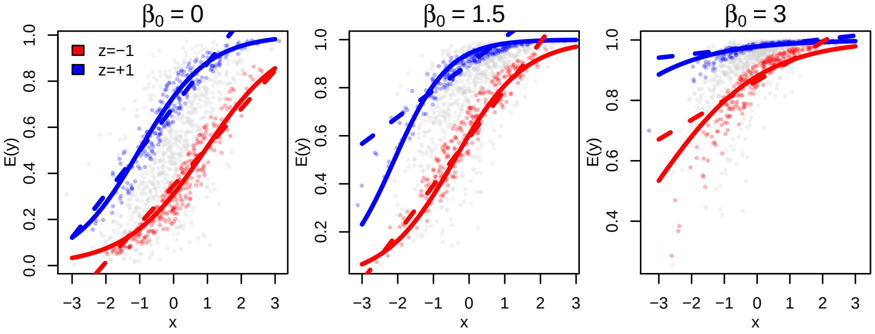
[5] As an alternative illustration, suppose that $\mathbb{E}(y|x, z) = \sigma(\beta_0 + \beta_1 x + \beta_2 z)$, where $\sigma(\cdot)$ is the standard logistic sigmoid. Using a Taylor series expansion about zero, we can write: $\mathbb{E}(y|x, z) = \frac{1}{2} + \frac{1}{4}(\beta_0 + \beta_1 x + \beta_2 z) + \frac{1}{48}(\beta_0 + \beta_1 x + \beta_2 z)^3 + \dots$. When expanding the cubic term, the highest weighted term is an interaction term including x and z: $\frac{1}{8}\beta_0\beta_1\beta_2 xz$. By modeling this expectation as $\mathbb{E}(y|x, z) = b_0 + b_1 x + b_2 z + b_3 xz$, we should expect false discovery of an interaction effect due to this term.

Figure 1: Power analysis when y is a binary outcome as a function of effect size (β_3) and prevalence (β_0). Horizontal gray line represents power of 0.8, a commonly used threshold. Blue circle indicates where there is no interaction and can be used to detect FDR (which should be equal to 0.05 given our choice of α). Line type based on properly specified model or linear alternative.



With this intuition at hand, we return to the question of discovery under the incorrect assumption of the linear model anticipating higher levels of false discovery when $\beta_0 \neq 0$. Consider the dashed lines in Figure 1 which are based on the linear probability model. As anticipated, results are highly sensitive to the value of β_0. When $\beta_0 = 0$, the linear probability models is underpowered relative to the logistic model but, especially with even larger samples than we analyze here, may be viable. However, when $\beta_0 = 2$, false discovery is extremely high.

Figure 2: Illustration of the geometry driving false discovery due to variation in β_0 when the linear model is used ($\beta_1 = \beta_2 = 1, \beta_3 = 0$) for analysis of binary outcomes. Blue and red dots represent those data points within half a unit of their respective z values.

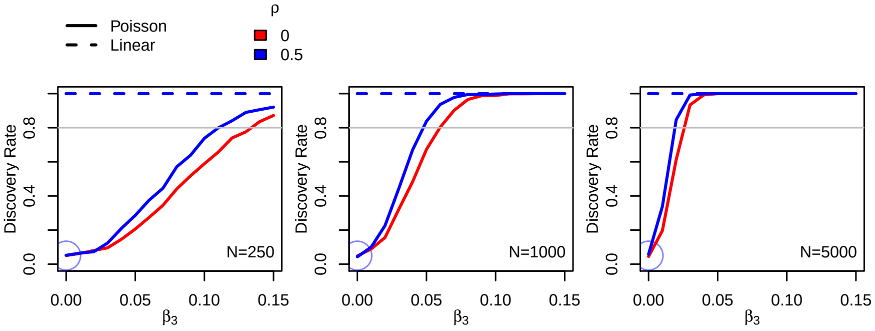


Balanced classes (or $\beta_0 = 0$) improve the situation but do not fully solve this problem, as the cubic term in the Taylor series still expands to contain the higher order interaction terms $\frac{1}{16}\beta_1^2\beta_2 x^2 z$ and $\frac{1}{16}\beta_1\beta_2^2 x z^2$ that a linear approach would likely attribute to an interaction between x and z.

3.2 Count outcomes

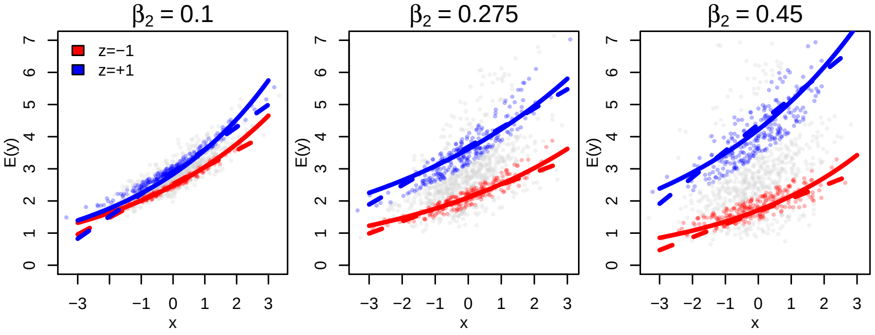
Here we focus on the scenario where y_i is a count outcome [15, 16, 17]. Analysis is based on $y_i \sim \text{Poisson}(\exp(\Gamma_i))$; i.e., we use the Poisson regression model (i.e., GLM with log link) for estimation. Results are in Figure 3; power again behaves as expected. In contrast to the case with binary outcomes, power is improved here when ρ is relatively large.

Figure 3: Power analysis when y is a count outcome as a function of effect size (β_3). Line type based on properly specified model or linear alternative.



False discovery is a catastrophic problem (similar to Figure 1 when $\beta_0 = 2$) if the OLS model is used in this setting. We again start with an illustration of the underlying geometry that leads to false discovery, see Figure 4. Here we focus on β_2 and hold $\beta_1 = 0.2$. We show this via control of variation of β_1 while fixing β_2. As β_2 increases, seeming variation as a function of z is introduced.[6] Turning back to Figure 3, deployment of the linear model leads to false discovery in all cases even when $\beta_3 = 0$. Based on the Taylor series expansion, we can see that this is due to $\beta_1\beta_2 \neq 0$. Use of the linear model for analysis of count outcomes is thus a practice to be discouraged.

Figure 4: Illustration of the geometry driving false discovery due to variation in b_1 when the linear model is used ($b_2 = 0.5, b_0 = 0$) for analysis of count outcomes. Blue and red dots represent those data points within half a unit of their respective z values.



3.3 Ordinal data

Ordinal outcomes are relatively common (e.g., years of education) and yet little has been said about the specific challenges they pose to interaction research. While numerous options exist for analysis of such data, we focus on cumulative link models [18]; while these models are relatively flexible, sensitivity analyses using multiple potential alternatives may be valuable (for an example of how this might be done, see [19]). If we

[6] As an alternative illustration, we can again consider a Taylor series expansion. Suppose that $\mathbb{E}(y|x,z) = \exp(\beta_1 x + \beta_2 z) = \exp(\beta_1 x)\exp(\beta_2 z)$. We can use Taylor series expansion to write this as $\mathbb{E}(y|x,z) = (1+\beta_1 x+(\beta_1 x)^2/2+\ldots)(1+\beta_2 z+(\beta_2 z)^2/2+\ldots) = 1+\beta_1 x+\beta_2 z+\beta_1\beta_2 xz+\ldots$. If we mistakenly assume that $y = b_1 x + b_2 z + b_3 xz + \epsilon$, we would anticipate $b_3 = \beta_1\beta_2$ if we omit higher order terms.

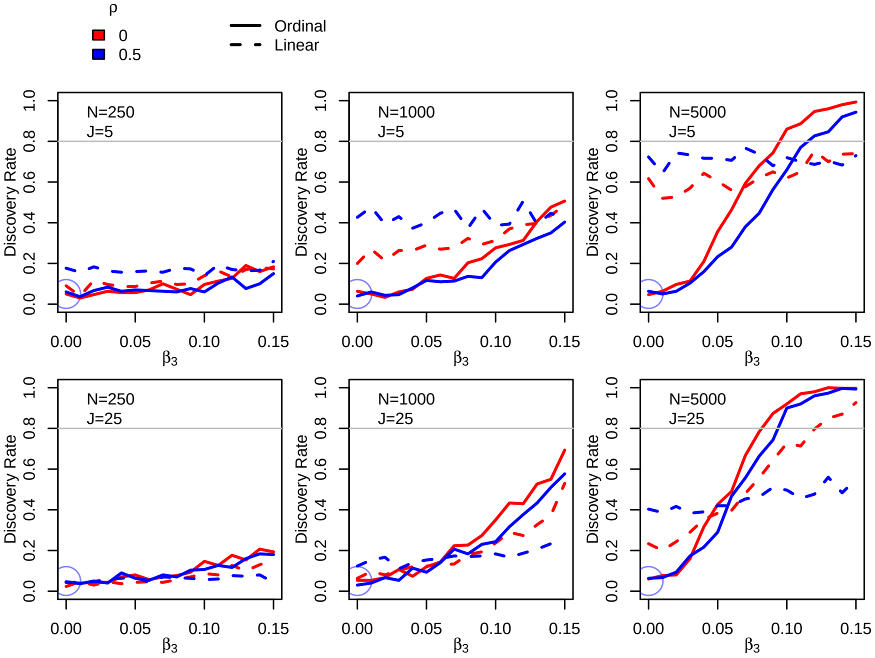
denote the levels of the ordinal variable as $j \in \{1, \ldots, J\}$, cumulative link models posit that

$$\Pr(y_i \leqslant j) = \sigma(\theta_j - \Gamma_i) = \pi_{ij} \tag{2}$$

for $j < J$ (where σ is again the logistic sigmoid and $\Pr(y_i = J) = 1 - \sum_{j<J} \pi_{ij}$).

We focus, see Figure 5, on relatively small and relatively large values of J and where θ_j are ordered independent draws from the standard normal.[7] For each iteration of the simulation we randomly sample θ_j from the standard normal. When we consider the properly specified model, power is generally reduced when $\rho > 0$ but relatively comparable as a function of J.

Figure 5: Power analysis when y is an ordinal outcome (analyzed using cumulative link model) as a function of β_3. Line type based on properly specified model or linear alternative.



Note that recovery with OLS is flawed. False discovery rates increase dramatically with sample size. More distressingly, the ability to detect true positives via OLS is poor even in well-powered samples unless J is large and ρ is small. As with analysis of count variables, application of linear models is likely to lead to spurious findings in this setting. Additional description of the potential benefits of the cumulative link approach can be found elsewhere [20].

3.4 Discussion

We consider analysis of several outcomes—binary, count, ordinal—based on various transformations of Γ. Our finding suggest that power can be sensitive to the configuration of the relevant parameters and that tailored analyses may be required in many cases. As an illustration of the need for carefully tailored power analysis, consider the role of ρ. In some cases, power is higher when $\rho = 0$; in other cases, it is reduced.

Further, our results suggest that naive application of the OLS model will tend to result in either high-level of false discovery (when $\beta_3 = 0$) or reduced power (when $\beta_3 > 0$). In certain cases, inferences should be presumed to be fatally compromised. In fact, false discovery was only a tractable problem when using the linear model in the case of binary outcomes when β_0 was relatively small in magnitude. Although there are other considerations that may lead one away from the GLM approaches we consider here [21, 22], we note that the problems we emphasize are unique to identification of interaction coefficients; analysis of main effects via the linear model is not necessarily compromised in the same way. We would urge heightened scrutiny when attempting to model interaction effects using alternatives that may be incorrectly specified.

[7] An illustration of the implied $\Pr(y = j)$ for a single case can be found in Figure A.1.

4 Transformations of $y^\star$

We now turn attention to outcomes that meet many of the assumptions of the OLS approach but have metric properties that render the standard OLS model insufficient. We assume that the true model is $\mathbb{E}(y^\star|x,z) = \Gamma_i$; however we only observe y where $g(y^\star) = y$ for some transformation g. We focus on transformations g that induce metric limitations in y_i; y_i may have a floor/ceiling or it may be measured on a non-interval scale. For these analyses, we focus on levels of false discovery.

4.1 Ceilings and floors

We first consider transformations g that lead to scales with a ceiling or floor [23]. This phenomenon is also known as censoring. In many cases, censoring can be readily observed (i.e., by looking at a histogram). Note that previous work has emphasized that using a variable with censoring as a predictor may lead to inflated type 1 error rates [24] or other forms of bias [25] but here we focus instead on cases wherein the censored variable is an outcome. We focus on a floor but the same concerns would apply to ceilings. To simulate data, we suppose that

$$y_i^\star \sim \mathrm{N}(\Gamma_i, \sigma^2_{y^\star}). \tag{3}$$

However, we do not observe $y_i^\star$; rather, we observe

$$y_i = g(y^\star) = \begin{cases} y_i^\star & y_i^\star > c \\ c & y_i^\star \leqslant c \end{cases} \tag{4}$$

for some constant c.

Simulations showing the proportion of false positives as a function of c are shown in Figure 6. Outcomes are standardized prior to imposing the floor so that the value of c can be interpreted by reference to the standard normal. We first illustrate recovery when we address the existence of censoring via Tobit regression (via [26]). Note that the level of false discovery is uniformly around 5% irrespective of c. Turning to the results from deployment of the linear model, we again observe large proportions of false positives in certain cases. When the floor is quite low, $c < -2$, there are relatively few false positives in small samples but a hint of false discovery in larger samples. When $c > -2$, false discoveries become pervasive in larger samples. Further, the rate of false positives increases as a function of ρ. Collectively, this suggests that the presence of ceilings or floors may present serious challenges to interaction research even if the censoring does not affect a large proportion of cases.

To illustrate the intuition behind false discovery due to the existence of floors, see Figure 7. We show y points as a function of xz in gray versus $y^\star$ in red. When the floor is sufficiently low, few points are effected and recovery with the censored data yields the correct inference. This is due to the fact that we observe the full "kidney bean" shape of (xz, y). However, when the floor begins to censor a significant number of cases, we no longer observe the bottom half of the kidney bean. When these points are raised towards the y mean, they count less towards the sum in the correlation and lead to the resulting (spurious) positive correlation between xz and y.

4.2 Outcomes with non-interval scales

Many psychological constructs are challenging to measure; one method of dealing with this challenge is the use of latent variable models. For example, rather than simply summing the number of correct responses on a test or the number of symptoms indicated on a depression scale, those responses can be used in item response theory models [27] to produce more complex scale scores. However, neither the raw sum scores nor the outcomes produced by latent variable models necessarily have interval scales. The interval interpretation allows us to suppose that a one unit change across the scale consistently has the same meaning. This interpretation is valid for many physical measures (e.g., a one meter change in length always means a change by a standard amount) but does not necessarily hold for psychological measures [28]. Data may have structure necessary for interval interpretations [29], but in many cases we do not have strong evidence of such structure. A failure to have this equal interval property can have implications for a variety of scale uses [30].

Figure 6: FDR analysis in the presence of a floor. Note that we use the pink background to visually offset these FDR analyses from power analyses. Line type based on properly specified model or linear alternative.

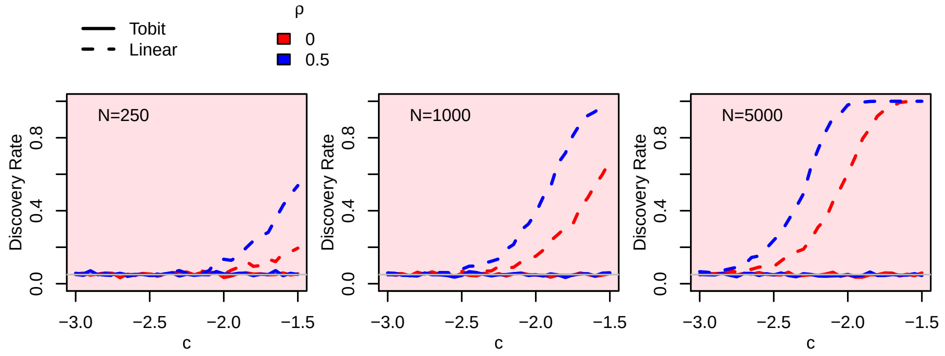
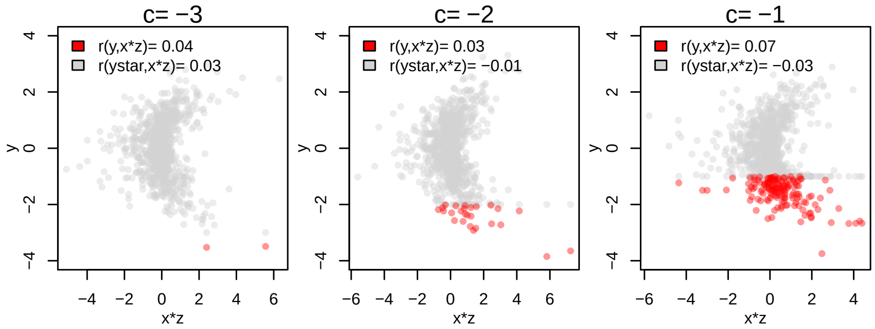


Figure 7: Illustration of the geometry driving false discovery when outcomes are censored (due to a floor). Red dots represent those points that are censored. Correlations are between censored and uncensored variables and the xz product term.



To illustrate the impact that analysis of outcomes mapped to non-interval scales, we consider a monotonic transformation of $y_i^\star$. In particular, motivated by "Lord's transformation" [31], we consider a transformation which stretches the scale for larger values of $y_i^\star$. That is, we suppose that $\mathbb{E}(y^\star|x,z) = \Gamma_i$ but that we observe

$$y_i = g(y^\star) = 1.05^{(y_i^\star-\alpha)/\lambda}. \tag{5}$$

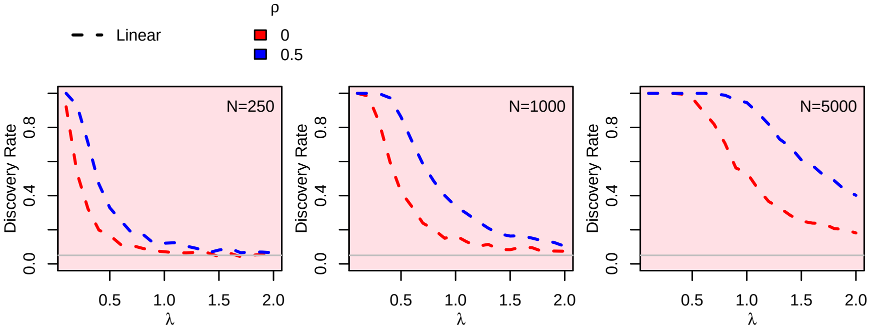
Here we consider $\alpha = 0$ and allow λ to vary. This transformation has the effect of making a one unit difference in y have differential meaning in the $y^\star$ metric, see illustration in Figure A.2. When λ is small, distances between $y^\star$ are compressed while distances between larger values are inflated. Differences between the distances quickly dissipate as λ increases. Note that the transformation's effect has similarities to the effect of the a floor in that, in both cases, differences between small values of $y^\star$ are being minimized.

We consider FDR as a function of λ when $\beta_3 = 0$. Results are shown in Figure 8. False positives are a concern throughout the range of λ. FDR is high here for reasons that are similar to those shown in Figure 7. Note also the role of ρ. When x and z are highly correlated, interactions more nearly mimic the effect of exponentiating y and this leads to elevated false discovery. Note that in this case we do not consider the correctly specified alternative as it is not even apparent in most cases that a given outcome would be measured on such a scale.

4.3 Discussion

Analysis of outcomes that are transformed versions of continuous outcomes can be challenging. In particular, the magnitude of the transformation can have implications for false discovery. When transformations are pronounced, the resulting bias will lead to spurious findings of interactions—i.e., heightened levels of FDR—even when none exist.

Figure 8: FDR analysis for linear model in the presence of a non-interval scaling of the outcome (see Eqn 5 where $\alpha = 0$).



5 General Discussion

The study of interactions is of clear interest in many settings. Yet, we argue here that the specific properties of the outcome variables in question may make accurate inference in such settings challenging. To illustrate this point, we considered two types of outcomes: outcomes based on a transformation of Γ and outcomes based on a transformation of some intermediate variable $y^\star$ where $\mathbb{E}(y_i^\star|x,z) = \Gamma_i$. In both cases, there are a variety of problems that may threaten inference; some of these issues are relatively easy to diagnose and resolve, others less so. We separately discuss implications for analysis of both types of outcomes before concluding with general thoughts on analysis of interactions.

For outcomes based on transformations of Γ, there are clear gains to using the correctly specified alternative. In all cases, the correctly specified approach led to improved power. In most cases, using the OLS approach led to elevated levels of false discovery when in fact $\beta_3 = 0$; in many of these, the level of false discovery was catastrophic. While it is fairly easy to identify the prevalence of binary outcomes, complete knowledge in the other cases may be a challenge. For example, the proportional odds assumption of the cumulative link model may need to be scrutinized. If it does not hold, analysis of interactions will be more challenging given that the answer may depend on the level of the ordinal outcome, j. We would encourage researchers to look for consistency across several models that rely upon different assumptions (e.g., [19]).

For outcomes based on transformations of $y^\star$, our results suggest that common problems will serve to increase FDR when $\beta_3 = 0$. Some of these problems are perhaps fairly easy to identify (i.e., the existence of a floor) and can frequently be addressed using standard solutions (i.e., models for censored data). Other problems (i.e., non-interval scales) may be hard to identify and nevertheless have deleterious effects on statistical inference. We view this as a rationale for a form of statistical humility—in particular, a single study should rarely be viewed as a dispositive piece of evidence regarding heterogeneity absent very large samples[8] and very high-quality measurement—especially when studying outcomes with challenging measurement properties of the kind that abound in social science.

Much of what we discuss is not new in the sense that these findings are anticipated by other studies focusing on power and false discovery in the context of misspecified models or, for example, censored outcomes. We view the issues of interactions as warranting special attention given interest in such studies. In terms of power, our findings suggest the need for carefully constructed power analyses if the goal is to derive appropriate guidance about, for example, sample sizes. To that end, the `InteractionStudio` package can be used to further craft in-house power analyses so that researchers can make informed decisions about their ability to study interactions in the contexts in which they work. We think power calculations that are tailored to the situation at hand are important given that, for example, conclusions about power may be sensitive to conditions such as the correlation between various predictors.

Turning to the issue of false discovery, we make two key points. First, the issues of false discovery raised here cannot be dealt with by, for example, utilizing robust standard errors. The problems illustrated in Figures 2, 4, and 7 show that it is the parameter estimates themselves—not merely the standard errors—that are flawed (for a related argument about the limitations of robust standard errors, see [32]). We have

[8] Although, we would note that increased sample size only serves to exacerbate some problems.

focused on false discovery given that it directly calls attention to the likelihood of high rates of false positives in many interaction studies focused on outcomes that present measurement challenges. The second (closely related) point is that biased parameter estimates of the kind observed here may lead to a large quantity of spurious findings in research focusing on interactions. This is a concern and should lead to increased attention to the nature of the outcome variable in studies of statistical interactions.

If the "some do, some don't" formulation holds true—and, we believe that it does in many cases—interaction studies will clearly be of interest. Yet, poorly-designed interaction studies can lead fields into dead-ends; consider, for example, the era of candidate gene-by-environment studies [33] which is now viewed as a vast literature consisting of almost entirely false positives. A failure to steer clear of dead-ends can lead to wasting large quantities of resources—both in terms of finite research dollars and even scarcer researcher time. Thus, we encourage researchers pursuing questions focused on studies of heterogeneity of the kind that invoke statistical interactions to take a more realistic perspective on the quality of inference likely to result from their particular data context.

References

[1] Julia M Haaf and Jeffrey N Rouder. Some do and some don't? accounting for variability of individual difference structures. *Psychonomic Bulletin & Review*, 26(3):772–789, 2019.

[2] Laurien M Buffart, Joeri Kalter, Maike G Sweegers, Kerry S Courneya, Robert U Newton, Neil K Aaronson, Paul B Jacobsen, Anne M May, Daniel A Galvão, Mai J Chinapaw, et al. Effects and moderators of exercise on quality of life and physical function in patients with cancer: an individual patient data meta-analysis of 34 rcts. *Cancer treatment reviews*, 52:91–104, 2017.

[3] Scott D Halpern, Benjamin French, Dylan S Small, Kathryn Saulsgiver, Michael O Harhay, Janet Audrain-McGovern, George Loewenstein, David A Asch, and Kevin G Volpp. Heterogeneity in the effects of reward-and deposit-based financial incentives on smoking cessation. *American journal of respiratory and critical care medicine*, 194(8):981–988, 2016.

[4] Fabio Conforti, Laura Pala, Vincenzo Bagnardi, Giuseppe Viale, Tommaso De Pas, Eleonora Pagan, Elisabetta Pennacchioli, Emilia Cocorocchio, Pier Francesco Ferrucci, Filippo De Marinis, et al. Sex-based heterogeneity in response to lung cancer immunotherapy: a systematic review and meta-analysis. *JNCI: Journal of the National Cancer Institute*, 111(8):772–781, 2019.

[5] Matthew C Keller. Gene× environment interaction studies have not properly controlled for potential confounders: the problem and the (simple) solution. *Biological psychiatry*, 75(1):18–24, 2014.

[6] David Lubinski and Lloyd G Humphreys. Assessing spurious" moderator effects": Illustrated substantively with the hypothesized (" synergistic") relation between spatial and mathematical ability. *Psychological bulletin*, 107(3):385, 1990.

[7] Robert C MacCallum and Corinne M Mar. Distinguishing between moderator and quadratic effects in multiple regression. *Psychological Bulletin*, 118(3):405, 1995.

[8] Jason M Fletcher and Dalton Conley. The challenge of causal inference in gene–environment interaction research: Leveraging research designs from the social sciences. *American journal of public health*, 103(S1):S42–S45, 2013.

[9] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

[10] Robert C Culverhouse, Nancy L Saccone, Amy C Horton, Yinjiao Ma, Kaarin J Anstey, Tobias Banaschewski, Margit Burmeister, Sarah Cohen-Woods, Bruno Etain, Helen L Fisher, et al. Collaborative meta-analysis finds no evidence of a strong interaction between stress and 5-httlpr genotype contributing to the development of depression. *Molecular psychiatry*, 23(1):133–142, 2018.

[11] Marie-Laure Ancelin, Jacqueline Scali, Joanna Norton, Karen Ritchie, Anne-Marie Dupuy, Isabelle Chaudieu, and Joanne Ryan. Heterogeneity in hpa axis dysregulation and serotonergic vulnerability to depression. *Psychoneuroendocrinology*, 77:90–94, 2017.

[12] Najada Stringa, Yuri Milaneschi, Natasja M van Schoor, Bianca Suanet, Sven van der Lee, Henne Holstege, Marcel JT Reinders, Aartjan TF Beekman, and Martijn Huisman. Genetic liability for depression, social factors and their interaction effect in depressive symptoms and depression over time in older adults. *The American Journal of Geriatric Psychiatry*, 2020.

[13] Ben Domingue, Sam Trejo, Emma Armstrong-Carter, and Elliot Tucker-Drob. Interactions between polygenic scores and environments: Methodological and conceptual challenges. *Sociological Science*, 7:465–486, 2020.

[14] Sam Trejo, Daniel W Belsky, Jason D Boardman, Jeremy Freese, Kathleen Mullan Harris, Pam Herd, Kamil Sicinski, and Benjamin W Domingue. Schools as moderators of genetic associations with life course attainments: evidence from the wls and add heath. *Sociological science*, 5:513–540, 2018.

[15] Chris G Richardson and Pamela A Ratner. Sense of coherence as a moderator of the effects of stressful life events on health. *Journal of Epidemiology & Community Health*, 59(11):979–984, 2005.

[16] Joanne Angosta, Mai-Ly N Steers, Kieran Steers, Jordanna Lembo Riggs, and Clayton Neighbors. Who cares if college and drinking are synonymous? identification with typical students moderates the relationship between college life alcohol salience and drinking outcomes. *Addictive behaviors*, 98:106046, 2019.

[17] Jacquelyn L Meyers, Jessica E Salvatore, Fazil Aliev, Emma C Johnson, Vivia V McCutcheon, Jinni Su, I Sally, Chun Kuo, Dongbing Lai, Leah Wetherill, et al. Psychosomeyers2019psychosocialcial moderation of polygenic risk for cannabis involvement: the role of trauma exposure and frequency of religious service attendance. *Translational psychiatry*, 9(1):1–12, 2019.

[18] Rune Haubo B Christensen. Analysis of ordinal data with cumulative link models—estimation with the r-package ordinal. *R-package version*, 28, 2015.

[19] K Paige Harden, Benjamin W Domingue, Daniel W Belsky, Jason D Boardman, Robert Crosnoe, Margherita Malanchini, Michel Nivard, Elliot M Tucker-Drob, and Kathleen Mullan Harris. Genetic associations with mathematics tracking and persistence in secondary school. *NPJ science of learning*, 5(1):1–8, 2020.

[20] Ben Domingue, Klint Kanopka, Sam Trejo, and Jeremy Freese. An ordinal model for analysis of years of education. 2021.

[21] William Ryan, Ellen Evers, and Don A Moore. False positive poisson. *Available at SSRN 3270063*, 2018.

[22] HS Battey, DR Cox, and MV Jackson. On the linear in probability model for binary data. *Royal Society open science*, 6(5):190067, 2019.

[23] O Garin. Ceiling effect. *Encyclopedia of quality of life and well-being research. Dordrecht: Springer Netherlands*, pages 631–633, 2014.

[24] Peter C Austin and Lawrence J Brunner. Type i error inflation in the presence of a ceiling effect. *The American Statistician*, 57(2):97–104, 2003.

[25] Lijuan Wang, Zhiyong Zhang, John J McArdle, and Timothy A Salthouse. Investigating ceiling effects in longitudinal data analysis. *Multivariate behavioral research*, 43(3):476–496, 2008.

[26] Arne Henningsen. Estimating censored regression models in r using the censreg package. *R package vignettes collection*, 5(2):12, 2010.

[27] Wim J van der Linden. *Handbook of Item Response Theory: Volume 1: Models.* CRC Press, 2016.

[28] Joel Michell. Is psychometrics pathological science? *Measurement*, 6(1-2):7–24, 2008.

[29] Ben Domingue. Evaluating the equal-interval hypothesis with test score scales. *Psychometrika*, 79(1):1–19, 2014.

[30] Dale Ballou. Test scaling and value-added measurement. *Education finance and Policy*, 4(4):351–383, 2009.

[31] Derek Briggs and Damian Betebenner. Is growth in student achievement scale dependent. *Unpublished manuscript*, 2009.

[32] Gary King and Margaret E Roberts. How robust standard errors expose methodological problems they do not fix, and what to do about it. *Political Analysis*, pages 159–179, 2015.

[33] Laramie E Duncan and Matthew C Keller. A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. *American Journal of Psychiatry*, 168(10):1041–1049, 2011.

Supplemental Information

A Supplemental Figures

Figure A.1: $\Pr(y_i = j)$ as a function of x when $\Gamma_i = 1x$ and $\theta = \{-1, 0, 1\}$.

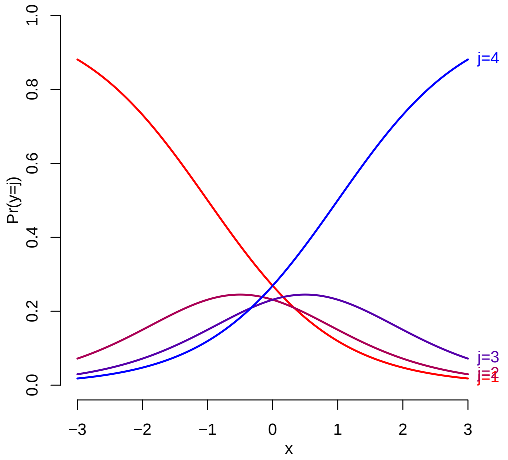
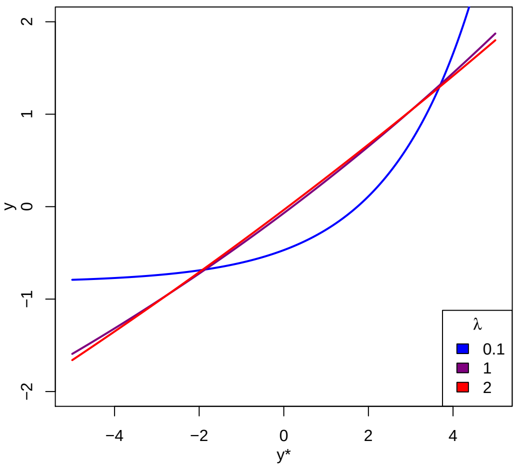


Figure A.2: Effect of transformation when $\alpha = 0$ for various values of λ.



.