Published in final edited form as:
Stat Med. 2011 December 30; 30(30): 3560–3572. doi:10.1002/sim.4377.

The impact of misclassification due to survey response fatigue on estimation and identifiability of treatment effects

Brian L. Egleston[1], **Suzanne M. Miller**[2], and **Neal J. Meropol**[3]

[1]Biostatistics and Bioinformatics Facility, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111

[2]Cancer Prevention and Control Program, Fox Chase Cancer Center, Case Western Reserve University, 11100 Euclid Avenue, Lakeside 1200, Cleveland, OH 44106-5065

[3]Division of Hematology and Oncology, University Hospitals Case Medical Center, Case Comprehensive Cancer Center, Case Western Reserve University, 11100 Euclid Avenue, Lakeside 1200, Cleveland, OH 44106-5065

Abstract

Response fatigue can cause measurement error and misclassification problems in survey research. Questions asked later in a long survey are often prone to more measurement error or misclassification. The response given is a function of both the true response and participant response fatigue. We investigate the identifiability of survey order effects and their impact on estimators of treatment effects. The focus is on fatigue that affects a given answer to a question rather than fatigue that causes non-response and missing data. We consider linear, Gamma, and logistic models of response that incorporate both the true underlying response and the effect of question order. For continuous data, survey order effects have no impact on study power under a Gamma model. However, under a linear model that allows for convergence of responses to a common mean, the impact of fatigue on power will depend on how fatigue affects both the rate of mean convergence and the variance of responses. For binary data and for less than a 50% chance of a positive response, order effects cause study power to increase under a linear probability (risk difference) model, but decrease under a logistic model. The results suggest that measures designed to reduce survey order effects might have unintended consequences. We present a data example that demonstrates the problem of survey order effects.

1 Introduction

Correcting for missing data and measurement error in survey research is an important goal of many researchers. A number of methods have been proposed to account for bias due to nonresponse (missing data), such as multiple imputation [1][2], missing data weights [3][4] [5], and pattern mixture models [6]. Measurement error and misspecification error are other forms of response problems which have been well described in the econometrics and statistics literature.[7][8][9]

A potentially important source of measurement and misclassification error that is often overlooked is participant response fatigue. This occurs when individuals respond to survey questions but do not provide truthful or consistent responses in order to reduce the burden of answering questions. Satisficing theory has been used to conceptualize the cognitive reasons why question order can affect responses.[10] The theory predicts that some individuals will give an acceptable answer, for example, the first or default answer, to minimize the burden of responding to the survey.[11]

There are other ways in which response fatigue can affect responses as well. As an example in the health field, respondents might not respond affirmatively that they have a certain chronic or acute health condition. Mathiowetz and Lair [12] suggest that respondents might learn through the survey experience that they can avoid survey subquestions about a health condition if they respond that they do not have the condition. In this way, study participants might answer in such a way as to reduce the length of the survey for them personally, particularly if they do not perceive the condition to be of high concern. Similarly, Hill and Pylypchuk [13] speculate that failure to acknowledge health conditions due to survey response fatigue could lead to overly optimistic estimates of a population's health status.

Survey response fatigue can have a key role in affecting inferences from randomized clinical trials. To illustrate this issue, we used data collected as part of the CONNECT™ study [14] [15]. We hypothesized that survey response fatigue created baseline differences between study arms due to differing survey lengths between the arms. The CONNECT™ study included an internet-based survey as a component of a cancer patient intervention designed to improve the communication between patients and their oncologists.

In this paper, we highlight the problems related to survey response fatigue using relatively simple models. Here, we consider the impact of survey fatigue on the observed response (i.e. answer) given to a question, rather than the impact of fatigue of non-response and missing data. These findings can guide future methodological work in this area. This paper is organized as follows.

In section 2, we detail the problematic baseline differences between survey arms in the CONNECT™ study. In section 3, we present models of the effects of survey fatigue on responses and discuss how fatigue might affect the variability of treatment effect estimators. In section 4, we discuss the identifiability of such models, and in section 5 we present a simulation study. We present evidence from CONNECT™ in section 6 that supports our choice of models and then end with a discussion.

2 The CONNECT™ study

The CONNECT™ study was a three arm randomized trial of a computer assisted intervention designed to improve patients' communication skills. The primary outcomes of the study were patient satisfaction with communication and decisional conflict. The control group was directed to information available from the National Cancer Institute's website (number randomized [n]=272), while the intervention arms received either 1) computer-based communication skills training (n=242) or 2) the same computer communication aid supplemented by a summary report supplied to the physician (n=229) [14][15]. As part of the CONNECT™ intervention, we measured monitoring and blunting scores (MBBS).[16] [17][18] The monitoring score consists of eight questions measuring patients' attention to potentially threatening health information (higher scores indicate more attention), while the blunting score consists of eight questions measuring patients' avoidance of potentially threatening health information (higher scores indicate more avoidance). Sample questions are shown in an appendix.

Patients in the CONNECT™ control arm were not asked baseline questions related to satisfaction or decisional conflict as study investigators were concerned that such questions could affect patient-physician interaction and hence contaminate the control condition. As a result, the survey administered to the intervention arms was longer than the survey administered to the control arm. Patients reported median times spent completing the survey of 15-30 minutes in the control arm and 31-45 minutes in the intervention arms. Since questions related to monitoring and blunting were placed at the end of both versions of the survey, those in the intervention arms had spent a longer time answering questions by the

time they responded to the monitoring and blunting items than those in the control arm. The monitoring and blunting assessments were asked in such a way that participants had to check off all of the items that might apply to them. Hence, if a respondent did not respond, it was not possible to distinguish nonresponse from a lack of applicability.

When baseline characteristics were compared between the two groups, it was found that the monitoring and blunting scores were higher in the control arm than in the intervention arms (3.9 mean monitor score (standard deviation [SD]=2.0) in the control arm vs. 3.2 (SD 1.7) and 3.4 (SD 1.7) in the intervention arms; 2.6 mean (SD 1.6) blunter score in the control arm vs. 2.2 (SD 1.3) and 2.3 (SD 1.4) in the intervention arms). Pairwise t-tests indicated that the differences in means were statistically significant ($p<0.02$) between the control and intervention arms, but were not statistically significant between the two intervention arms ($p>0.14$).

The differences between the control and intervention arms were unlikely due to participants failing to respond to all of the monitoring and blunting questions, as there were few people with monitoring or blunting scores of zero in any arm (3.7% in the control arm and 3.4% in the intervention arms had monitor scores of 0, for example). Further, despite the fact that both scores are expected to move in opposite directions, meaning that individuals with high monitoring scores should have low blunting scores, both were lower in the intervention arms than the control arm. This suggests that individuals in the intervention arms had not simply stopped answering questions. We posit that two scenarios could have accounted for such baseline differences among the study arms: 1) patients in the intervention arms were primed to answer questions differently by contextual effects of questions not asked in the control arm, and 2) patients might have become less careful in reading and responding to the monitoring and blunting questions.

The possibility that patients might have become less careful in responding caused us to consider how the placement of questions could similarly affect the power to detect effects in our future randomized trials. Much of the literature cited in this paper focuses on the effect of survey response fatigue on estimator bias. While the effect of fatigue on bias is important, the effect of fatigue on power might be of even greater importance in randomized clinical trials. We have found in designing oncologic intervention trials, for example, that the primary focus of clinical trials is often to determine which intervention is associated with the best outcome. This leads to a binary decision regarding which is better. While the magnitude of the effect is important, an experimental intervention that is better will generally become the new standard of care regardless of how much better it is. This motivated us to investigate the impact of survey response fatigue on study power, as the impact of fatigue on making correct decision rules might be of greater importance to researchers than assuring that estimators are unbiased.

3 Misclassification models to assess effects of survey fatigue

Deriving models for a participant's response to a question helps clarify how survey fatigue can affect estimators of treatment effects. For this section let Y_{to} be the response that an individual would have to a question under intervention assignment t if the question was asked in ordered place o in a survey (for example, $o = 4$ indicates the question was the fourth question). Let μ_t be the mean response when the question is asked as the first question. Let n represent the number of participants within each arm of a study; we assume equally sized arms. We present models of response fatigue when Y_{to} is binary (affirmative or yes answer=1, 0 otherwise) and when Y_{to} is continuous. We include the subscripts t and o to simplify notation so that we do not have to condition upon treatment assignment and question order in expectations and probabilities. Thus, it is assumed that $E[Y_{1o}]$ and $E[Y_{0o}]$

refer to expected values of responses in two independent treatment arms (i.e., $E[Y_{to}] = E[Y_{to}|T = t, O = o]$ for treatment indicator T, and order variable O).

We assume in this section that all participants in a study, regardless of arm, are asked questions in the same order. In section 4, we consider identifiability of response fatigue effects when the order is varied among participants. Such identifiability would allow for the correction of response fatigue effects.

These tend to be elementary models of survey response fatigue for ease of presentation. As a reviewer noted, our models do not account for the heterogenous effects that could account for misclassification, such as the number of words in a question, question complexity, or respondent cognitive ability and motivation. Subsequent research could examine how accounting for such effects would alter our findings.

3.1 Binary Y_{to}

3.1.1 Linear Model—A simple model of the effect of survey fatigue would be,

$$P(Y_{to}=1) = \mu_t + \eta(o-1) \quad (1)$$

In the model, η represents the expected effect of question order on the probability of response. The linear probability model is useful for modeling risk differences as seen in the epidemiologic literature.[19] Given our experience with CONNECT™ and the hypothesized effects of survey fatigue reported in the literature [12][13], we would expect $\eta \leq 0$ as some people stop answering questions thoughtfully over time.

From this model, it is easy to see that

$$\begin{aligned} &P(Y_{1o}=1) - P(Y_{0o}=1) \\ &= \mu_1 + \eta(o-1) - \mu_0 - \eta(o-1) \quad (2) \\ &= \mu_1 - \mu_0 \end{aligned}$$

If we introduce the subscript $i = \{1, \ldots, n\}$ for each person within treatment arm $t \in \{0, 1\}$, an estimator for the linear effect would be,

$$\widehat{p}_{1o} - \widehat{p}_{0o} = \frac{\Sigma_{i=1}^{n} Y_{i1o}}{n} - \frac{\Sigma_{i=1}^{n} Y_{i0o}}{n}$$

Under maximum likelihood theory, we have the following approximate variance.

$$var(\widehat{p}_{1o} - \widehat{p}_{0o}) \approx \frac{P(Y_{1o}=1) * P(Y_{1o}=0) + P(Y_{0o}=1) * P(Y_{0o}=0)}{n}$$

Of note is the fact that the variance of the estimator is related to the variance of the responses. $P(Y_{1o} = 1) * P(Y_{1o} = 0)$ for example is the variance for a single binary yes-no response (Bernoulli distribution). Under the binomial distribution, probabilities closer to 50% lead to larger variances than probabilities closer to one or zero.

Impact on study power: To see how fatigue impacts the power to detect effects, notationally define Z to be a standard normal variable (mean=0, variance=1) and z_p represent the quantile such that $P_{hx}(Z < z_p) = p$ under hypothesis h_x. Assume a null hypothesis, h_0, of $P(Y_{1o} = 1) - P(Y_{0o} = 1) = 0$ and an alternative hypothesis, h_1, of $P(Y_{1o} = 1) - P(Y_{0o} = 1) = \delta$. Let β represent the Type II error rate of a study. The approximate

power, $1 - \beta$, to detect an effect using a two sided Wald test with Type I error rate α becomes $P_{h1}(Z < z_{(1-\beta)}) = 1 - \beta$ in which

$$z_{(1-\beta)} = |\delta| / \sqrt{var(\widehat{p}_{1o} - \widehat{p}_{0o})} + z_{\alpha/2}. \quad (3)$$

Under the linear model in equation (2), $P(Y_{t(o+1)} = 1) \leq P(Y_{to} = 1)$. Hence, as o increases such that the probability of response falls below 50%, $var(\widehat{p}_{1o} - \widehat{p}_{0o})$ will decrease. Since the absolute difference, δ, does not change with o (as shown in equation 2), the reduction in the variance as o increases will cause $z_{(1-\beta)}$ to become larger and hence the power to increase. However, if the baseline probability of affirmative response is above 50%, then reducing the probability of affirmative response closer to 50% would increase the variance of the estimated δ and hence reduce power. Of course, as the probability of affirmative response falls very close to zero, then δ cannot remain fixed.

3.1.2 Logistic Model—A linear model for binary variables is not necessarily appropriate, especially if the probability of response is near zero or one. Instead, we might assume a logistic model.

$$\log\left\{\frac{P(Y_{to}=1)}{P(Y_{to}=0)}\right\} = \mu_t + \eta(o-1) \quad (4)$$

Here η is the effect of question order on the log odds of response. A traditional estimand for the intervention effect is the odds ratio (OR = odds $P(Y_{1o} = 1)$/ odds $P(Y_{0o} = 1)$) which when logged would become

$$\log \quad OR = \mu_1 - \mu_0. \quad (5)$$

Assume we estimate the log odds ratio using estimated probabilities, $\widehat{p}_{to} = \Sigma_{i=1}^{n} Y_{ito}/n$. Using maximum likelihood theory combined with the δ-method, we can show that,

$$var\left(\log \quad \widehat{OR}\right) \approx \frac{1}{P(Y_{0o}=1)\times n} + \frac{1}{P(Y_{0o}=0)\times n} + \frac{1}{P(Y_{1o}=1)\times n} + \frac{1}{P(Y_{1o}=0)\times n}$$

Under this formulation, the log odds ratio would remain fixed, but the approximate variance would increase as the probabilities fell further below 50%. Again, this is related to changes in the variability of the actual responses. Thus, we would eventually reduce the power to detect effects, as can be shown analogously to the power discussion in section 3.1.1.

3.2 Continuous Y_{to}

3.2.1 Linear model—A possibility when a measure is continuous is that survey response fatigue causes two groups to converge on a default value, such as 50 on a 0 to 100 scale, and the variability of responses would hence change. A model of this could be,

$$Y_{to} = \mu + \mu_t I(o=1) + o^{\eta}\mu_t I(o>1) + o^{\lambda}\epsilon_{to} \quad (6)$$

Here, μ is the default value and μ_t **represents** the difference in the treatment specific mean when the question is asked as the first question. $I(\cdot)$ is an indicator function which takes 1 if the condition inside the parentheses is true, 0 otherwise. In the model, ε_{to} is a normally distributed error term with mean 0 and variance ψ^2. The variance of Y_{to} hence becomes $o^{2\lambda}\psi^2$.

For $\eta < 0$, o^{η} establishes the rate at which the treatment specific means converge to a common mean. For $\eta > 0$**, the means would diverge.** A likely scenario in equation (6) is that $lim_{o\to\infty} E[Y_{1o}] - E[Y_{0o}] = 0$ because $\eta < 0$. When $\lambda > 0$, then the variance of the response increases with o, while it decreases when $\lambda < 0$. Order has no impact on variability of response when $\lambda = 0$. In a case in which the signs of μ_0 and μ_1 are reversed, we might need to include treatment specific fatigue effects (i.e. η_t instead of η).

The treatment effect of interest for $o > 1$ could be,

$$E[Y_{1o}] - E[Y_{0o}] = o^{\eta}(\mu_1 - \mu_0)$$

The approximate variance of the difference in sample means used to estimate $E[Y_{1o}] - E[Y_{0o}]$ would be $2o^{2\lambda}\psi^2/n$. The impact of survey response fatigue on study power would depend on the rate of change in the variance relative to the rate of change of the treatment effect, $\delta = E[Y_{1o}] - E[Y_{0o}]$**. Under this formulation for** $o > 1$**, equation 3 would become** $z_{(1-\beta)} = \sqrt{n}|o^{(\eta-\lambda)}(\mu_1 - \mu_0)|/\left(\sqrt{2}\psi\right) + z_{\alpha/2}$**. If** $\eta < \lambda$**, then** $z_{(1-\beta)}$ **would decrease as** o **increases, leading to a decrease in power. If** $\eta > \lambda$**, then** $z_{(1-\beta)}$ **would increase as** o increases, leading to an increase in power. Finally, if $\lambda = \eta$, then question order would have no impact on power.

Based on the literature, it is likely that $\lambda < 0$ and hence the variance of Y_{to} decreases. For example, in a study of question order effects, Galesic and Bosnjak [20] found that the variability of responses decreased for questions asked later in a survey. Such reduction is also consistent with the main arguments of Krosnick [21], although he did posit the possibility that variability could increase with o if subjects began to randomly answer questions, implying $\lambda > 0$. Findings by Krosnick and Alwin [11] suggest that $\eta < 0$, since respondents might be more likely to give one of the first options in a list of answers to a question. Over time, subjects in both groups might increase the tendency to pick one of the first options available, leading to convergence of the treatment specific means. As a reviewer suggested, the model could also be modified to account for subject specific convergence to default values. Such subject specific convergence to different individual values regardless of treatment assignment would still be consistent with a convergence of the difference in population treatment means to zero in a randomized trial.

To demonstrate how η **and** λ affect power in equation (6), we present how power changes as η **varies from −0.04 to 0, and** λ **varies from −0.05 to 0.05 in figure 1. We set** n**=250,** $\mu_1 - \mu_0 = 0.25$**,** $\psi^2 = 1$, and the Type I error rate to 5% (two-sided). We also examined power when o=1, 25, and 100. For ease of comparison, these are similar values to those used in the simulations in Section 5. We see that the power in this case is very sensitive to small changes in η **and** λ.

3.2.2 Gamma Model—Assuming a normally distributed error term might not be realistic for many summary score variables. Often, the lower bound of summary measures is zero. Hence, if people stop responding to questions, the score will become closer to zero and the variance of the responses will decrease. In such a case, a more realistic model for the response might be a generalized linear model assuming a log link and a Gamma family distribution. Let γ_t and ν parameterize the Gamma distribution such that $E[Y_{to}] = \exp(\gamma_t)$ and $\text{var}(Y_{to}) = \exp(\gamma_t)^2/\nu$. A model of the effect of response fatigue would be,

$$\log \quad E[Y_{to}] = \mu_t + \eta(o - 1)$$

In this case, η represents the effect of question order on the log of the treatment specific mean. The treatment effect becomes the relative increase or decrease under treatment versus control. That is,

$$\log\left\{\frac{E[Y_{1o}]}{E[Y_{0o}]}\right\}=\mu_1-\mu_0 \quad (7)$$

We can estimate the treatment effect by fitting the model $\log E[Y_{to}] = \gamma_t$, which does not require us to explicitly model response fatigue, and using $\widehat{\delta}=\widehat{\gamma}_1-\widehat{\gamma}_0$ as the estimator. Our model has the following properties.

$$\begin{aligned} E\left[\widehat{\delta}\right] &= \mu_1-\mu_0 \\ \operatorname{var}\left(\widehat{\delta}\right) &\approx 2/(n\nu) \quad \text{by MLE theory and the delta method} \end{aligned}$$

In this formulation, survey fatigue has no effect on the ability to detect effects since the approximate variance of $\widehat{\delta}$ is fixed. An important distinction between the linear and Gamma models is that the treatment effect is modeled as additive in the former while multiplicative in the latter. A multiplicative effect would be more realistic as scores in both groups fall to zero when using instruments that cannot take negative values. Another consequence of the Gamma model is that although the variance in $\widehat{\delta}$ does not change due to survey response fatigue, since we are examining relative differences between treatment groups, the variance of the actual responses, Y_{to} will decrease due to fatigue. This is a result of $\gamma_t = \mu_t + \eta(o-1)$ decreasing as o increases when $\eta < 0$, and hence $\operatorname{var}(Y_{to}) = \exp(\gamma_t)^2/\nu$ also decreasing.

3.3 Extension to summary measures

While we have focused on individual questions in surveys, many psychosocial variables are summary measures of a number of items, as was the case with the monitoring and blunting scales in the CONNECT™ study. For example, let U_t consist of the sum of two continuous questions asked in order o_a and o_b with respective default means of μ and α, and assume the error terms are independent. Under our model in equation (6), for $o_a > 1$ and $o_b > 1$, $E[U_t] = \mu+\alpha+\left(o_a^{\eta}\mu_t+o_b^{\eta}\alpha_t\right)$ and the variance of U_t would become $\psi^2\left(o_a^{2\lambda}+o_b^{2\lambda}\right)$. Our summary measure would still have power and convergence of treatment specific means that are dependent on the relative magnitudes of η and λ.

4 Identifiability

The η parameters in our models are not identifiable unless the order of questions varies among participants. The following hypothetical example can demonstrate this. Let X_{to} represent a baseline survey question. Let π represent the mean value for variable X_{to}; since X_{to} is a baseline variable, the mean does not depend on treatment assignment. If the questions were asked in the same order for each participant, we represent the full and hypothetical observed data in table 1.

Assuming extensions of the linear probability models for two binary variables, we have:

$$P(X_{to}=1)=\pi+\eta(o-1) \quad (8)$$

$$P(Y_{to}=1)=\mu_t+\eta(o-1) \quad (9)$$

for $t \in \{0, 1\}$. Note that these models are not identifiable from the observed data in Table 1 Panel b since our data provides us with estimates of $P(X_{t1} = 1)$ and $P(Y_{t2} = 1)$, but we would need to estimate more than two parameters in equations (8) and (9): μ_t, π and η. The number of parameters would grow if we more realistically allow for different ηs in models (8) and (9).

We could make stronger assumptions that identify η. In the case of the CONNECT™ study, for example, patients were randomized to three arms. Let T be an indicator of treatment assignment ($T = 1$ if assigned to a specific intervention arm, $T = 0$ if assigned to the control arm). By randomization, we know that $T \perp \{X_{1o}, X_{0o'}\}$ for all o and o'. In words, this means that T is independent (with $\perp$ denoting independence) of the potential question responses. Note that the observed response is X_{1o} if $T = 1$ and X_{0o} if $T = 0$. Hence, we can show that,

$$\eta = \frac{P(X_{1o}=1) - P\left(X_{0o'}=1\right)}{o - o'} = \frac{P(X_{1o}=1 \mid T=1, O=o) - P\left(X_{0o'}=1 \mid T=0, O=o'\right)}{o - o'}$$

Since $P(X_{1o} = 1 \mid T = 1, O = o)$ and $P(X_{0o'} = 1 \mid T = 0, O = o')$ are identified in the observed data for $o \neq o'$, we can estimate η. In studies in which survey length does not differ between study arms, we can randomly assign O as well as T such that we can identify η by using data between study arms or within one study arm.

5 Simulations

We performed a simulation study to examine how survey fatigue affects the power to detect effects. For each simulation, we generated data 5,000 times under the conditions described below. For relevant link function $g(\cdot)$, we tested the null hypothesis that $g(E[Y_{1o}]) - g(E[Y_{0o}]) = \delta = 0$ versus the alternative hypothesis that $\delta \neq 0$. We then estimated the power under various conditions by calculating the proportion of times that we rejected the null hypothesis under a 5% Type I error rate (two-sided) when the alternative hypothesis was true. We set the simulations such that the expected power would be in the vicinity of 80% to 90% for some of the assumptions. The parameterization of the models was varied within each simulation to investigate the sensitivity of the findings to changes in the parameterization. We used the rbinom(·), rnorm(·), and rgamma(·) functions in R (The R Foundation for Statistical Computing, Vienna, Austria) to generate variables used in simulations. In the simulations, we examined models under two assumptions about the η parameters: the baseline assumptions listed below and after multiplying η by five. We also estimated power when $o \in \{1, 25, 100\}$. The assumptions used to generate the data are as follows.

5.1 Specification of Simulations

Linear Model: Binary Case—For the linear probability (risk difference) model, we assumed that $\mu_1 \in \{40\%, 60\%, 90\%\}$ and that $\mu_1 - \mu_0 = .1$. **We assumed that** $\eta = -.001$. We set the sample size per arm at 250.

Logistic Model—For the logistic model, we assumed that $\mu_1 \in \{-.5, .5, 1.5\}$ and that $\mu_1 - \mu_0 = .5$. **We assumed that** $\eta = -.007$. The parameterization set $\{P[Y_{0o} = 1 | o = 1], P[Y_{1o} = 1 | o = 1]\}$ to {27%, 38%}, {50%, 62%}, and {73%, 82%}. We set the sample size per arm to 250.

Linear Model: Normal Case—For the linear model with normally distributed data, we assumed that, $\mu = 0$, $\lambda = -.025$, $\psi^2 = 1$, $\mu_1 \in \{-5, 0, 10\}$ and that $\mu_1 - \mu_0 = .25$. **We assumed that** $\eta = -.025$. We set the sample size per arm to 250.

Gamma Model—For the Gamma model, we assumed that $\mu_1 \in \{3, 4, 4.5\}$ and that $\mu_1 - \mu_0 = .25$. **We assumed that** $\eta = -.025$. We set ν equal to 1.5. The parameterization set $\{E[Y_{0o} = 1|o = 1], E[Y_{1o} = 1|o = 1]\}$ to {15.6, 20.1}, {42.5, 54.6}, and {70.1, 90.0}. We set the sample size per arm to 200.

5.2 Simulation results

In table 2, we show the estimated power from the simulations. In table 2a, we present the results for the linear model of binary data. As expected, if survey fatigue causes probabilities of affirmative answers, $P(Y_{to} = 1)$, to get closer to 0, power goes up. However, as survey fatigue reduces probabilities from above 50% to closer to 50%, the power goes down. When η **is multiplied by five and** $\mu_1 = 40\%$**, then model probabilities are negative at** $o = 100$. This shows the limitation of the linear binary model for small probabilities. When $\mu_1 = 60\%$ the power is consistently 62% or 63% except in the extreme case. This is due to $\{P(Y_{1o} = 1), P(Y_{0o} = 1)\}$ **being close to** {60%, 50%} **or** {50%, 40%}**, which are similar distances from** 50%.

In table 2b, we present results from the logistic model. As expected, if survey response fatigue causes probabilities of affirmative answers to get closer to 0, power goes down. However, as survey fatigue reduces probabilities from above 50% to closer to 50%, the power goes up. Under $5 \times \eta$, the power falls substantially as the probabilities approach 0. When $\mu_1 = -.5$**,** $o = 100$**, and** η multiplied by five, both the probabilities of affirmative answers and power fall to low levels: $P(Y_{1o}|o = 100) = 1.8\%$ **and** $P(Y_{0o}|o = 100) = 1.1\%$**, and power=2%.**

In table 2c, we present the results of the linear model in which both the means and variances converge as o **increases since both** λ **and** η **are less than zero. In the baseline case,** $\eta = \lambda$**, so there is no change in power as** o **increases. However, since** $5 \times \eta < \lambda$**, power decreases as** o **increases.**

In table 2d, we present results for the Gamma model. Fatigue has no impact on overall power in the Gamma model. This is due to the relative difference between means remaining stable even as the absolute difference decreases.

6 Applicability of models to CONNECT™

Since the questions in CONNECT™ were asked in the same order within survey arms, we could not test the applicability of the models above directly using the CONNECT™ data. However, we could investigate the models indirectly by examining how the relative (intervention divided by control) means and variances of different subsets of questions changed between the control and intervention arms. To do this, we examined 10 similarly scored questions related to the SF-12 (5 point ordinal scores [22]) which were asked at the same location in both the control and intervention arms, and hence we would not expect differences between the groups in the expected means or variances of responses. We next examined relative differences in the Revised Impact of Events Scale (RIES [23] ordinal responses of 0,1,3,5 possible) questions which were asked 16 questions later in the intervention arm than the control arm. Finally, we examined the monitoring and blunting scores (MBBS) which were asked 48 questions later in the intervention arm than the control arm (binary scores).

In Figures 2 and 3, we plot the relative means and variances and present the slopes and 95% confidence intervals (CIs) from simple linear regressions of the ratios. These slopes have different interpretations from the slope parameters in Section 3 models. We see that relative means and variances of SF-12 scores did not vary much between intervention and control groups, as we would expect since questions were asked in the same order between groups. In the figures, ratios of 1 indicate no differences in means or variances between control and intervention groups, ratios above 1 indicate larger values under intervention, and ratios below 1 indicate smaller values under intervention. Although the 95% confidence interval for the slope of SF-12 relative means did not cross zero, the magnitude of the slope was small. For the RIES scales, the ratios of the means and variances showed slight downward slopes whose upper bounds on the 95% confidence intervals were close to zero. This is consistent with the responses to the RIES questions converging to a common response as in the linear model, equation 6.

For the MBBS questions that had mean responses of below 50% in the control group, the average value of the binary answers fell in the intervention group and the variance of the answers also fell, as would be predicted under a linear probability (risk difference) model. The averages did not fall as much for the MBBS scores in which the baseline probability of responding affirmatively was above 50%, but the variance increased, again consistent with the linear probability model. For the MBBS figures, the 95% confidence interval is close to zero only for the slope of the relative variance line for questions with larger probabilities of answering affirmatively.

While it is possible that the patterns could be driven by contextual differences due to different questions asked between the two surveys, the fact that we see trends within the subsets of questions that are predicted by our models suggests that survey response fatigue was present. If the context of the questions was driving the results, we would not necessarily expect to see the downward slopes within subsets of contiguous questions since priming by previous questions should affect all of the responses similarly, on average. While the 95% confidence intervals were not well bounded away from zero, the small number of questions within each group limited the power to make definitive conclusions.

7 Discussion

Our theoretical considerations of the effects of survey response suggest that survey response fatigue can both increase and decrease study power depending on the model used. For continuous data, survey response fatigue would have no impact on study power under a Gamma model in which the data are assumed to be bounded by 0, which is applicable to many psychosocial summary measures with lower bounds. However, under our linear model which converges to a common value regardless of treatment arm, the impact of response fatigue on power is dependent on the degree to which fatigue affects the variance of responses relative to the rate of mean convergence. If the response variance decreases more rapidly than the convergence in means, then response fatigue could increase power. However, if the convergence in means is fast relative to changes in response variance, then response fatigue could decrease power under the convergent linear model. For binary data, survey response fatigue would cause power to increase under a linear probability model, but eventually decrease under a logistic model. Linear probability models can be used to model absolute risk differences, while logistic models are more useful for modeling relative differences between groups.[24]

As discussed in Section 6, some of the models we investigated were consistent with differences in mean and variance patterns between the control and intervention arms in the CONNECT™ study. The linear probability (risk difference) and both continuous models

were also consistent with the findings of Galesic and Bosnjak [20] in that the models allow for a reduction in the variability of responses for questions asked later in a survey.

A number of researchers have conducted studies investigating the impact of question and response order on survey response rates. Malhotra [25], for example, found that the order of response options within questions on a webpage could affect responses, particularly among those with less education who answered a survey more quickly. Benton and Daly [26] similarly found an education effect in the effect of question order on response. Such demographic effects would suggest that the order effect parameters described above, such as the η and λ terms, could be subgroup specific.

In this paper, we have considered how survey response fatigue affects power rather than why response fatigue can cause question order effects. The importance of knowing how fatigue will affect the decision making aspect of a randomized clinical trial (e.g. determining which of two treatments is superior) can be as important as knowing why there might be an impact. Our findings that survey response fatigue does not always reduce the power to detect effects suggests that costly interventions to overcome such fatigue might not be necessary. When collecting binary data, addressing survey response fatigue would be most necessary when relative measures of association, such as the odds ratios, are of interest. When absolute differences are of interest with respect to binary outcomes, measures to correct for response fatigue might be counterproductive.

We have focused on the effect of response fatigue and question order on study power. We have not focused on other issues that could affect the relationship of question order with response. For example, we did not consider cases in which individuals' responses might be influenced by other questions in a survey. Responses to questions about self-rated health might be influenced by whether the questions are asked before or after objective health questions.[27] Such priming effects on response would presumably have other effects on study power. Also, the complexity of a survey could exacerbate survey response fatigue. In our models, this could impact the magnitude of the η parameters; greater complexity would be associated with larger η parameters. Further, it is likely that there is a non-linear relationship between question order and response fatigue. This would argue for order dependent effects which could be incorporated into the models above by entering o via a polynomial or spline transformation with additional model parameters.

Our binary and Gamma models often assumed that we were always able to obtain unbiased effect estimates. This is a reasonable assumption under the Gamma and logistic models as we are interested in relative rather than absolute differences between the two groups. This might be less realistic for modeling risk differences using a linear probability model as the probability of response is bounded by zero from below. Hence, a 10% absolute difference (for example, 40% rates **of "yes" responses** in one arm versus 50% in another arm) cannot be maintained as the probability of **answering yes** falls below 10% in both arms. Our findings concerning the effects of survey fatigue may not therefore be generalizable to extreme cases of response fatigue.

Issues concerning response fatigue and question order effects in general have not received much attention in the statistics literature. This is in contrast to other common problems found with survey research. For example, an ISI Web of Knowledge topic search on missing data found over 2,000 papers in the statistics and probability literature addressing this issue. Many of the most cited papers develop methods for handling data that is completely missing (for example, Rubin [28]). However, few if any methods have been developed to correct for response fatigue in which data is observed, but answers are biased towards a presumed default value.

While response fatigue might not always impact the bias of treatment effect estimators, it could still bias estimators such as within treatment arm means. For example, a two armed trial finding the proportion stating that their quality of life is good to a binary quality of life question (1=good quality, 0 otherwise) might have baseline proportions of 50% in the control arm and 60% in the treatment arm. If survey response fatigue reduced the proportion responding “good” to 20% and 30%, the treatment effect remains unchanged (60%-50%=30%-20%=10%). However, the difference of proportions within treatment arms (e.g. 50% versus 20% in the control arm) is clinically meaningful. Thus, the goal of a study will dictate the degree to which measures should be undertaken to reduce survey response fatigue.

Survey response fatigue could also impact longitudinal data if questions are asked in different orders at different time points. In a longitudinal dataset, a change in the responses to a survey question over time could be linked to true temporal changes or to changes resulting from survey fatigue. It is hence important for researchers to also consider how survey question order over time can impact inferences.

In conclusion, we have have found that the relationship between survey response fatigue and study power is not always intuitive. The measures that researchers take to reduce survey response fatigue should be commensurate with the goals of the study.

Acknowledgments

This work was supported in part by NIH grants R01 CA82085, P30 CA 06927, and an appropriation from the Commonwealth of Pennsylvania. We thank Eric Ross and Fang Zhu for their comments.

9 Appendix

Sample questions used to create the monitoring and blunting scores.

1. Vividly imagine that you are going to the dentist and have to get some dental work done. Which of the following would you do? Check all of the statements that might apply to you.
    - **1a)** I would ask the dentist exactly what work was going to be done.
    - **1b)** I would take a tranquilizer or have a drink before going.
    - **1c)** I would try to think about pleasant memories.
    - **1d)** I would want the dentist to tell me when I would feel pain.
    - **1e)** I would try to sleep.
    - **1f)** I would watch all the dentist’s movements and listen for the sound of the drill.
    - **1g)** I would watch the flow of water from my mouth to see if it contained blood.
    - **1h)** I would do mental puzzles in my mind.

References

[1]. Rubin DB. Multiple imputation after 18+ years. Journal of the American Statistical Association. 1996; 91(434):473–489.

[2]. Schafer, JL. Analysis of incomplete missing data. Chapman & Hall/CRC; 1997.

[3]. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association. 1994; 89(427):846–866.

[4]. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association. 1995; 90(429):106–121.

[5]. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric non-response models. Journal of the American Statistical Association. 1999; 94(448):1096–1120.

[6]. Little RJA. Pattern-mixture models for multivariate incomplete data. Journal of the American Statistical Association. 1993; 88(421):125–134.

[7]. Bound J, Brown C, Mathiowetz N. Measurement error in survey data. Handbook of Econometrics. 2001; 5:3705–3843.

[8]. Goldberg JD. The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. Journal of the American Statistical Association. 1975; 70(351):561–567.

[9]. Wang CY, Huang Y, Chao EC, Jeffcoat MK. Expected estimating equations for missing data, measurement error, and misclassification, with application to longitudinal nonignorable missing data. Biometrics. 2008; 64(1):85–95. DOI: 10.1111/j.1541-0420.2007.00839.x. [PubMed: 17608787]

[10]. Holbrook AL, Krosnick JA, Moore D, Tourangeau R. Response order effects in dichotomous categorical questions presented orally: The impact of question and respondent attributes. Public Opinion Quarterly. 2007; 71(3):325–348. DOI: 10.1093/poq/nfm024.

[11]. Krosnick JA, Alwin DF. An evaluation of a cognitive theory of response-order effects in survey measurement. Public Opinion Quarterly. 1987; 51(2):201–219. DOI: 10.1086/269029.

[12]. Mathiowetz NA, Lair TJ. Getting Better? Change or error in the measurement of functional limitations. Journal of Economic and Social Measurement. 1994; 20(3):237–262.

[13]. Hill SC, Pylypchuk Y. Reports of fewer activity limitations: Recovery, survey fatigue, or switching respondent? Medical Care. 2006; 44(5):I73–I81. DOI: 10.1097/01.mlr.0000208199.13219.8b. [PubMed: 16625067]

[14]. Fleisher L, Buzaglo J, Collins M, Millard J, Miller SM, Egleston BL, Solarino N, Trinastic J, Cegala DJ, Benson AB 3rd, Schulman KA, Weinfurt KP, Sulmasy D, Diefenbach MA, Meropol NJ. Using health communication best practices to develop a web-based provider-patient communication aid: the CONNECT study. Patient Education and Counseling. 2008; 71(3):378–387. DOI:10.1016/j.pec.2008.02.017. [PubMed: 18417312]

[15]. Meropol NJ, Egleston BL, Buzaglo JS, Benson AB 3rd, Cegala DJ, Diefenbach MA, Fleisher L, Miller SM, Sulmasy DP, Weinfurt KP, CONNECT Study Research Group. Cancer patient preferences for quality and length of life. Cancer. 2008; 113(12):3459–3466. DOI: 10.1002/cncr.23968. [PubMed: 18988231]

[16]. Miller SM. Monitoring versus blunting styles of coping with cancer influence the information patients want and need about their disease. Implications for cancer screening and management. Cancer. 1995; 76(2):167–177. DOI: 10.1002/1097-0142(19950715)76:2¡iexcl;167::AID-CNCR2820760203¿iquest;3.0.CO;2-K. [PubMed: 8625088]

[17]. Miller, SM.; Fang, CY.; Diefenbach, MA.; Bales, CB. Tailoring psychosocial interventions to the individual's health information-processing style: The influence of monitoring versus blunting in cancer risk and disease. Psychosocial interventions for cancer. In: Baum, A.; Andersen, BL., editors. American Psychological Association; Washington, DC, US: 2001. p. 343-362.

[18]. Miller, SM.; Bowen, DJ.; Croyle, RT.; Rowland, JH. Handbook of cancer control and behavioral science: A resource for researchers, practitioners, and policymakers. American Psychological Association; Washington, DC, US: 2009. Handbook of cancer control and behavioral science: A resource for researchers, practitioners, and policymakers.

[19]. Cheung YB. A modified least-squares regression approach to the estimation of risk difference. American Journal of Epidemiology. 2007; 166(11):1337–1344. DOI: 10.1093/aje/kwm223. [PubMed: 18000021]

[20]. Galesic M, Bosnjak M. Effects of questionnaire length on participation and indicators of response quality in a web survey. Public Opinion Quarterly. 2009; 73(2):349–360. DOI: 10.1093/poq/nfp031.

[21]. Krosnick JA. Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied Cognitive Psychology. 1991; 5(3):213–236. DOI: 10.1002/acp.2350050305.

[22]. Ware JE, Kosinski M, Keller SD. A 12-item short form health survey: Construction of scales and preliminary tests of reliability and validity. Medical Care. 1996; 34(3):220–223. [PubMed: 8628042]

[23]. Horowitz M, Wilner N, Alvarez W. Impact of Event Scale: A measure of subjective distress. Psychosomatic Medicine. 1979; 41(3):209–218. [PubMed: 472086]

[24]. Ukoumunne OC, Forbes AB, Carlin JB, Gulliford MC. Comparison of the risk difference, risk ratio, and odds ratio scales for quantifying the unadjusted intervention effect in cluster randomized trials. Statistics in Medicine. 2008; 27(25):5143–5155. DOI: 10.1002/sim.3359. [PubMed: 18613226]

[25]. Malhotra N. Completion time and response order effects in web surveys. Public Opinion Quarterly. 2008; 72(5):914–934. DOI: 10.1093/poq/nfn050.

[26]. Benton JE, Daly JL. A question order effect in a local government survey. Public Opinion Quarterly. 1991; 55(4):640–642. DOI: 10.1086/269285.

[27]. Lee S, Grant D. The effect of question order on self-rated health status in a multilingual survey context. American Journal of Epidemiology. 2009; 169(12):1525–1530. DOI: 10.1093/aje/kwp070. [PubMed: 19363097]

[28]. Rubin DB. Inference and missing data. Biometrika. 1976; 63(3):581–592. DOI: 10.1093/biomet/63.3.581.

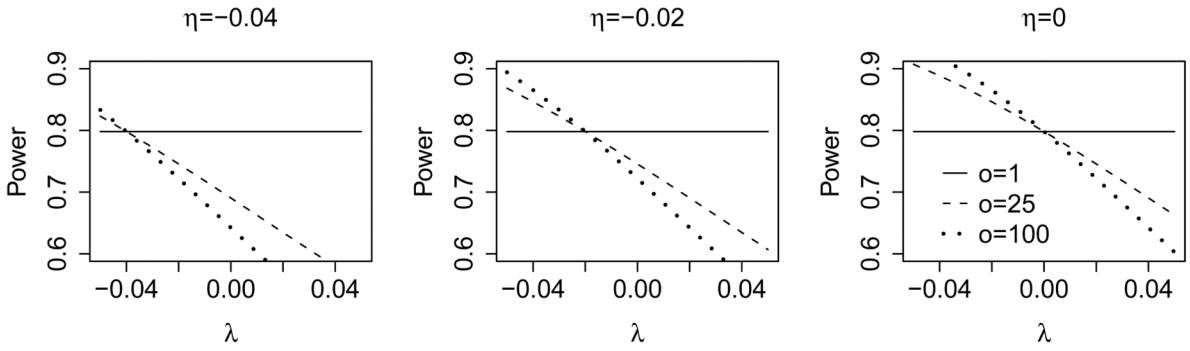

Figure 1.

Approximate power to detect effects for various values of η and λ in the linear model presented in equation (6). Here, n=250, $\mu_1 - \mu_0 = 0.25$, $\psi^2 = 1$, and the Type I error rate is set to 5% (two-sided).

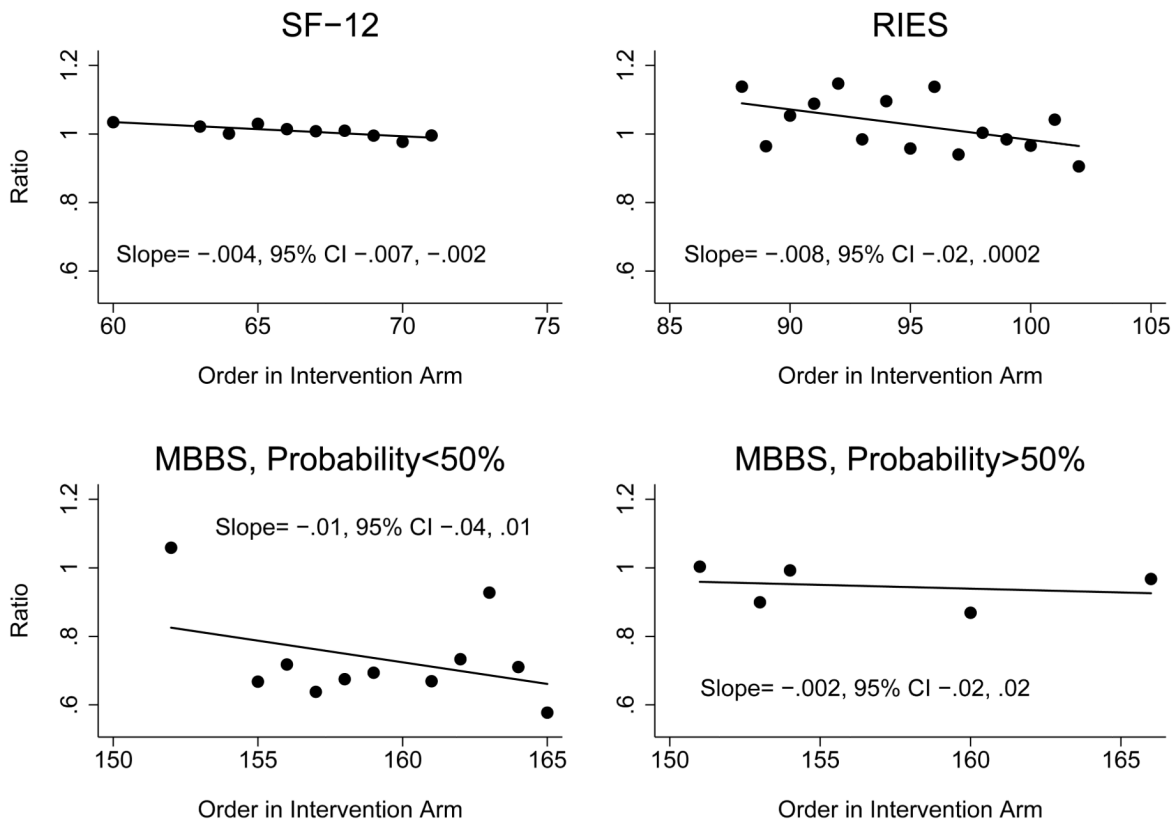

Figure 2.
Relative means, intervention divided by control, among subsets of CONNECT™ questions. The slope and 95% confidence intervals are given for each line.

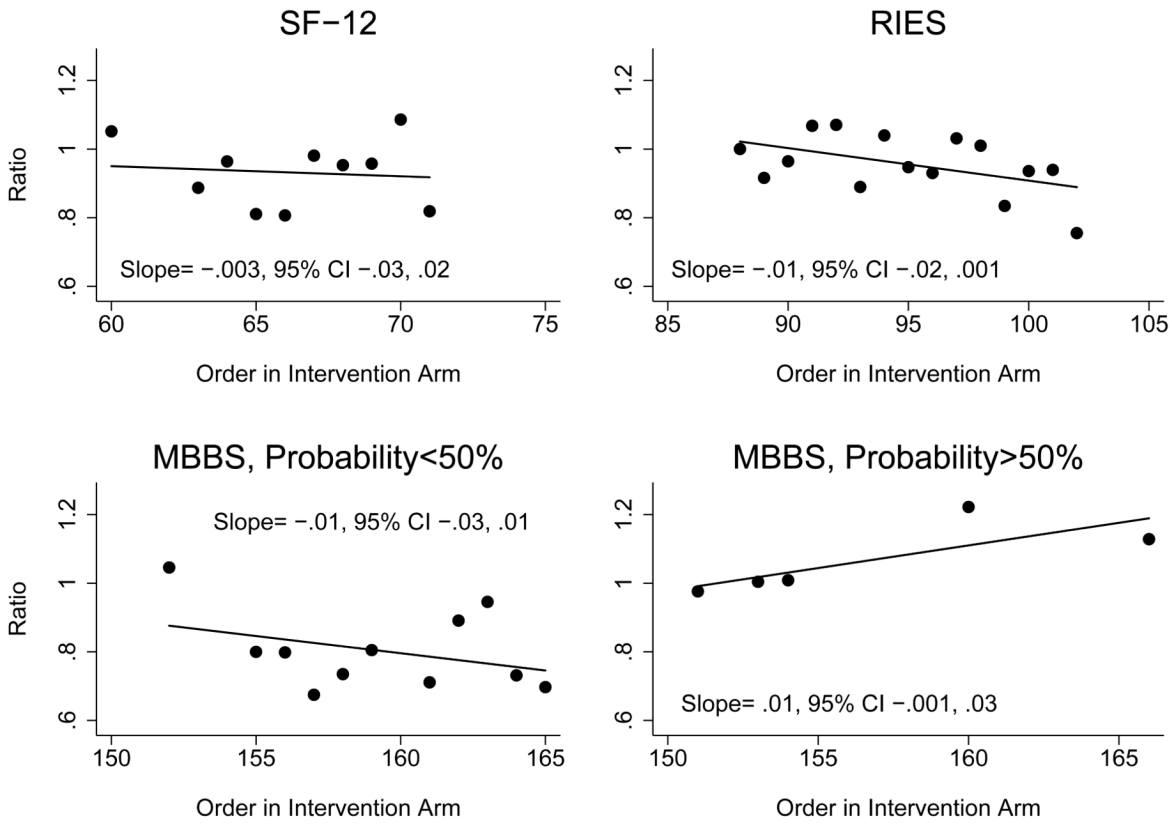

Figure 3.
Relative variances, intervention divided by control, among subsets of CONNECT™ questions. The slope and 95% confidence intervals are given for each line.

Table 1

Typical pattern of observed data in surveys in which question order is not randomly assigned to participants

	a: Full Data		b: Observed Data	
Question	Order=1	Order=2	Order=1	Order=2
X_{to}	X_{t1}	X_{t2}	X_{t1}	Missing
Y_{to}	Y_{t1}	Y_{t2}	Missing	Y_{t2}

Table 2

Empirical power to detect effects from simulations.

	Baseline η			5×η		
a: Linear Model: Binary case						
	Order (o)			Order (o)		
$\mu_{.1}$	1	25	100	1	25	100
40%	66%	67%	74%	65%	77%	NA
60%	63%	62%	63%	63%	62%	100%
90%	88%	84%	74%	88%	72%	64%
b: Logistic Model						
−0.5	74%	71%	59%	74%	55%	2%
0.5	80%	80%	77%	80%	75%	16%
1.5	64%	67%	76%	64%	78%	38%
c: Linear Model: Normal case						
−5	79%	79%	80%	79%	55%	41%
0	80%	80%	79%	80%	52%	42%
10	79%	81%	79%	80%	52%	42%
d: Gamma Model						
3	86%	86%	86%	86%	86%	86%
4	86%	87%	86%	87%	86%	86%
4.5	86%	86%	86%	86%	86%	87%

NA= Not applicable as probabilities become negative: $P(Y_{1O}|o = 20) = 40\% - .005 \times (100 - 1) = -9.5\%$