Sample Size Considerations for Micro-Randomized Trials with Binary Proximal Outcome

(to be finalized) Eric Cohn, Tianchen Qian, Susan A. Murphy

November 1, 2020

[Confidential; do not share.]

1 Introduction

Mobile health interventions refer to the healthy behavior change interventions delivered through mobile devices such as smartphones and wearable trackers, usually in the form of a push notification, a text message, or an audible ping. They have the potential to be delivered to each individual at the time and in the context they are most likely to benefit. To realize this potential, it is important to gather empirical evidence to inform when and under what context the interventions are the most beneficial, in order to improve and optimize the interventions.

The micro-randomized trial (MRT) is an optimization trial design that provides data to answer such questions (Liao et al., 2016; Dempsey et al., 2015; Klasnja et al., 2015). In an MRT, each participant is repeatedly randomized among multiple options of an intervention (no intervention is usually one of the options), usually hundreds or thousands of times throughout the trial. The prefix "micro-" reflects the frequent randomization on each participant. After each of such times (called a decision point), a near-term, proximal outcome is measured, which is typically an outcome that the intervention is directly targeting. For example, in a study where the intervention is a push notification suggesting a near-term physical exercise, the proximal outcome is the step count in the subsequent 30

minutes (Klasnja et al., 2015); in another study where the intervention is a message reminder for completing a daily in-app self-report, the proximal outcome is whether the participant completes the self-report in the next few hours Rabbi et al. (2018). Primary analyses of an MRT usually concern the treatment effect of the intervention on the proximal outcome and effect modification by baseline and/or time-varying covariates. Unbiased inference for such effects is made possible thanks to the micro-randomization. Such analyses suggest directions for improving the existing mobile health intervention.

Many MRTs involve binary proximal outcomes, because mobile health interventions target behavior change and a natural measure for the impact of such interventions is participant adherence. Besides the self-report completion example mentioned above, additional examples include BariFit (Klasnja et al., 2020), where one of the interventions is a message reminder for completing daily food log and the corresponding proximal outcome is whether the participant completes the food log on that day; JOOLHealth (Bidargaddi et al., 2018), where the intervention is a push notification to increase participant engagement and the proximal outcome is whether the participant self-monitors behaviors and feelings via the app during the subsequent 24 hours; Drink Less (Bell et al., 2020), where the intervention is a push notification to engage participant and the proximal outcome is whether the participant opens the app in the subsequent hour. Because there is no off-the-shelf software for determining sample size for MRTs with binary proximal outcome, to size such studies researchers currently rely on simulation, which can be time consuming, or rules of thumb combined with the sample size calculator for continuous outcome MRTs (Liao et al., 2016), which can be imprecise.

We propose a formula for determining the sample size of an MRT with binary proximal outcome using a test statistic constructed based on Qian et al. (2019). Input to the sample size formula includes the desired power and type I error, effect size, as well as characteristics of the participants (including the success probability of the outcome under no intervention and availability) and characteristics of the trial (including randomization probability and duration of the trial). We prove that under certain working assumptions the sample size formula results in the desired power to detect a prespecified marginal causal effect with desired type I error control. We show via simulation that the sample size formula works well

under various violations to the working assumptions. We also provide practical guidelines on applying the sample size formula and point out settings where the formula might fail. The sample size calculator is implemented in R package [TQ: package name and citation pending: currently two UCI students are working on the package] and an R Shiny web-app can be accessed at https://tqian.shinyapps.io/mrt_ss_binary/.

The rest of the paper is organized as follows. [TQ: To fill.]

2 Preliminaries

2.1 Notation

We describe the data structure collected from an MRT. We focus on the setting where each participant is in the MRT for the same number of decision points, the treatment option at each decision point is binary (e.g., deliver/not deliver a message), the timing of decision points are pre-determined (e.g., at fixed calendar times determined at the beginning of the study), and the randomization probability at each decision point can depend on the decision point index but not on other time-varying information such as the outcomes at previous decision points, nor can it vary between individuals. We will use the terms "intervention" and "treatment" interchangeably.

Consider an MRT with n participants. Denote by m the total number of decision points for a participant. For the i-th participant, at each decision point they are randomized among the two treatment options. We use $A_{it} \in \{0,1\}$ to denote their randomized treatment assignment at decision point t. Let $p_t = P(A_{it} = 1)$ denote the randomization probability at t; p_t can depend on t but not on i. Let $Y_{i,t+1}$ denote the binary proximal outcome measured after t. We assume that $Y_{i,t+1} = 1$ is desired over $Y_{i,t+1} = 0$.

There are times when it is inappropriate/unethical to deliver a treatment. For example, for safety reasons a mobile health app is usually designed so that a treatment is never delivered when the individual is detected to be driving. This is captured by the notion of availability: an individual is considered unavailable for treatment at such times. Let I_{it} be an indicator denoting whether participant i is available at t. Randomization only occurs

when $I_{it}=1$; if $I_{it}=0$, no randomization occurs and $A_{it}=0$ with probability 1.

The longitudinal data observed for participant i is $O_i=(I_{i1},A_{i1},Y_{i2},\ldots,I_{im},A_{im},Y_{i,m+1})$. We assume that O_i, $i=1,\ldots,n$ are independent and identically distributed (i.i.d.) draws from an unknown distribution P_0. In the following, we use letters without subscript i to denote variables of a generic participant. Data for a participant may also include baseline and time-varying covariates and they may help increase estimation precision in the primary analysis. We do not incorporate them in the calculation to ensure that the sample size calculated is conservative. We use $\mathbb{R}^p$ to denote the p-dimensional Euclidean space.

2.2 Marginal Excursion Effect

We use potential outcomes notation (Rubin, 1974; Robins, 1986) to define the causal effect of interest. Let $\bar{A}_t=(A_1,A_2,\ldots,A_t)$ denote the vector of (stochastic) treatment assignment up to t, and $\bar{a}_t=(a_1,a_2,\ldots,a_t)$ a realization of $\bar{A}_t$. Let $Y_{t+1}(\bar{a}_t)$ denote the potential proximal outcome that would have been observed if the individual is assigned treatment $\bar{a}_t$. Similarly, $I_t(\bar{a}_{t-1})$ denotes the potential availability at time t under treatment history $\bar{a}_{t-1}$.

Because mobile health interventions are usually designed to have an immediate influence on near-term behavior, as seen in the examples in Section 1, accessing whether A_t has the intended proximal impact on Y_{t+1} is of prominent interest in primary analyses. To quantify this impact we use the so-called marginal excursion effect (MEE), which is a fully marginal version of causal excursion effects (Boruvka et al., 2018; Qian et al., 2019). For binary proximal outcome, MEE is defined as

$$\text{MEE}(t)=\log\frac{P\{Y_{t+1}(\bar{A}_{t-1},1)=1\mid I_t(\bar{A}_{t-1})=1\}}{P\{Y_{t+1}(\bar{A}_{t-1},0)=1\mid I_t(\bar{A}_{t-1})=1\}},\ \text{for } t=1,\ldots,m. \tag{1}$$

The probabilities in (1) are integrated over the distributions of $\bar{A}_{t-1}$ and Y_{t+1} (conditional on $I_t(\bar{A}_{t-1})=1$). We refer to the distribution of $\bar{A}_{t-1}$ as the treatment protocol, because it captures how the treatment is (sequentially) randomly assigned in the MRT. MEE(t) quantifies the log relative risk in the proximal outcome at t between two excursions from the treatment protocol. One excursion is to follow the treatment protocol till $t-1$ and assign treatment ($A_t=1$) at t; the other is to follow the treatment protocol till $t-1$ and assign no treatment ($A_t=0$) at t. For MEE(t), a value greater than 0 indicates that the treatment is

effective. MEE(t) is defined conditional on being available at t, because a treatment may only be delivered when the individual is available and thus only the treatment effect at available moments are of interest.

It is possible that treatments assigned at earlier decision points, $\bar{A}_{t-1}$ will also have an impact on the current proximal outcome, Y_{t+1}. Such delayed effects may be attributed to habit formation (a positive delayed effect) or user-burden/habituation (a negative delayed effect). It is also likely that the near-term impact of the treatment would depend on certain baseline or time-varying covariate values (i.e., effect modification). MEE(t) is marginal over the past treatment assignment as well as the covariate distribution, and the primary focus on such marginal quantities is consistent with other optimization trials such as the factorial design (e.g., Collins et al., 2014). Delayed effects and effect modification may be further explored in secondary and exploratory analyses (Qian et al., 2019).

Because we focus on the setting where the randomization probability p_t may only depend on the decision point index but not other history information, MEE(t) can be expressed in terms of observed data distribution:

$$\text{MEE}(t) = \log \frac{P(Y_{t+1}=1 \mid A_t=1, I_t=1)}{P(Y_{t+1}=1 \mid A_t=0, I_t=1)}. \tag{2}$$

The proof is included in Appendix [TQ: to fill].

3 Test Statistic

The sample size formula we develop is based on a test statistic for testing the null hypothesis

$$H_0 : \text{MEE}(t) = 0 \text{ for all } t = 1,2,\ldots,m$$

against the alternative hypothesis

$$H_1' : \text{MEE}(t) \neq 0 \text{ for some } t \in \{1,2,\ldots,m\}.$$

An omnibus test that aims to detect every possible MEE(t) under H_1' will have low power for any particular alternative (Salkind, 2010). Instead, we propose a test statistic that will have high power against a targeted alternative MEE(t). That is, the test statistic trades-off

bias and variance so as to achieve high power against MEE alternatives close to the targeted alternative. Here variance is captured coarsely by the degrees of freedom in a t-statistic and bias is how far the true MEE function is from the targeted alternative.

We consider the hypothesis test where MEE(t) in the targeted alternative can be expressed as linear in a vector parameter $\beta_0 \neq 0$:

$$H_1 : \text{MEE}(t) = f(t)^T \beta_0 \text{ for all } t = 1,2,\ldots,m.$$

$f(t)$ is a pre-specified p-dimensional vector-valued function of t and $\beta_0 \in \mathbb{R}^p$. The choice of this targeted alternative is usually determined through conversation with the scientific team. For example, the scientific team might conjecture that a likely alternative MEE(t) function would be roughly close to zero early in the study, gradually increase with time and then possibly decrease to below zero by the end of the study. In this case the test statistic might target a quadratic alternative with $f(t) = (1,t,t^2)$, which involves 3 parameters. If the scientific team does not think that the alternative MEE(t) function would decrease below zero near the end of the study, it might be better to target a constant-in-time MEE(t) with $f(t) = 1$—this alternative involves only 1 parameter, thus, as will be seen below, saving 2 degrees of freedom.

Qian et al. (2019) proposed an estimator for $\beta \in \mathbb{R}^p$ when $\text{MEE}(t) = f(t)^T \beta$ for all $t = 1,\ldots,m$. Note that $\beta = 0$ corresponds to H_0 and $\beta = \beta_0$ corresponds to H_1. We construct the test statistic based on a modified version of their estimator as follows. Let $g(t)^T \alpha$ be a working model for $\log E(Y_{t+1} \mid A_t = 0, I_t = 1)$, where $g(t)$ is a q-dimensional feature vector and $\alpha \in \mathbb{R}^q$. We require that the linear span of $g(t)$ contains $p_t f(t)$. Let $(\hat{\alpha}, \hat{\beta})$ denote the solution (α, β) to the following estimating equation:

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{m} I_{it} \left\{ e^{-(A_{it} - p_t) f(t)^T \beta} Y_{i,t+1} - e^{g(t)^T \alpha} \right\} \begin{bmatrix} g(t) \\ (A_{it} - p_t) f(t) \end{bmatrix} = 0. \tag{3}$$

$\hat{\beta}$ is the estimator for β.

We show in Appendix [TQ: to fill] that when $\text{MEE}(t) = f(t)^T \beta$ for all $t = 1,\ldots,m$, under regularity conditions there exists $\alpha' \in \mathbb{R}^q$ such that $\hat{\alpha} \xrightarrow{P} \alpha'$ and that

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, M^{-1} \Sigma M^{-1,T}) \quad \text{as } n \to \infty, \tag{4}$$

where

$$M = \sum_{t=1}^{m} E\left\{ I_t e^{-(A_t - p_t) f(t)^T \beta} Y_{t+1} (A_t - p_t)^2 f(t) f(t)^T \right\},$$
$$\Sigma = \sum_{t=1}^{m} \sum_{s=1}^{m} E\left\{ I_t I_s r_t(\alpha', \beta) r_s(\alpha', \beta) (A_t - p_t)(A_s - p_s) f(t) f(s)^T \right\}, \tag{5}$$

and $r_t(\alpha, \beta)$ is defined as

$$r_t(\alpha, \beta) = e^{-(A_t - p_t) f(t)^T \beta} Y_{t+1} - e^{g(t)^T \alpha}. \tag{6}$$

The variance-covariance matrix $M^{-1} \Sigma M^{-1,T}$ can be consistently estimated by the sandwich estimator $\hat{M}^{-1} \hat{\Sigma} \hat{M}^{-1,T}$, where $\hat{M}$ and $\hat{\Sigma}$ are M and Σ with (α_0, β) replaced by $(\hat{\alpha}, \hat{\beta})$ and expectation replaced by sample average over all n participants.

We consider the following Wald-type test statistic

$$T = n \hat{\beta}^T (\hat{M}^{-1} \hat{\Sigma} \hat{M}^{-1,T})^{-1} \hat{\beta}. \tag{7}$$

Setting $\beta = 0$ in (4) implies that under H_0 the large sample distribution of T is χ^2_p, a chi-squared distribution with p degrees of freedom. Thus, a hypothesis test that uses critical value of the chi-squared distribution will have nominal type I error control asymptotically. To correct the downward bias of the sandwich estimator $\hat{M}^{-1} \hat{\Sigma} \hat{M}^{-1,T}$ when the sample size n is small, we use the critical value from a scaled F-distribution because $\frac{n-q-p}{p(n-q-1)} T$ approximately follows $F_{p, n-q-p}$, an F-distribution with degrees of freedom $(p, n-q-p)$ (Pan and Wall, 2002; Liao et al., 2016). In particular, the rejection region of a γ-level test for H_0 is

$$\left\{ T : T > \frac{p(n-q-1)}{n-q-p} F^{-1}_{p, n-q-p}(1-\gamma) \right\}, \tag{8}$$

where $F^{-1}_{p, n-q-p}$ is the quantile function of $F_{p, n-q-p}$. One can also incorporate the small sample correction in Mancl and DeRouen (2001) by adjusting $\hat{\Sigma}$ in (7) using the "hat" matrix; see Appendix [TQ: to fill] for the form of the adjusted $\hat{\Sigma}$.

Remark 1. To ensure the consistency and asymptotic normality of $\hat{\beta}$ (and thus the type I error control), $g(t)^T \alpha$ does not need to be a correct model for $\log E(Y_{t+1} \mid A_t = 0, I_t = 1)$ as long as the linear span of $g(t)$ contains $p_t f(t)$. A working model that approximates the true $\log E(Y_{t+1} \mid A_t = 0, I_t = 1)$ well would typically result in greater efficiency of $\hat{\beta}$ (and thus

higher power). When the working model is correctly specified, α', the probability limit of $\hat{\alpha}$, satisfies

$$e^{g(t)^T\alpha'} = e^{p_t f(t)^T\beta} E(Y_{t+1} \mid A_t = 0, I_t = 1). \tag{9}$$

This result will be useful for the sample size formula derivation in Section 4.

4 Sample Size Formula

Under $H_1 : \text{MEE}(t) = f(t)^T\beta_0$ for all t with $\beta_0 \neq 0$, the test statistic T approximately follows $\chi^2_p(\lambda_n)$, a non-central chi-squared distribution with p degrees of freedom and non-centrality parameter λ_n (Tu et al., 2004), where

$$\lambda_n = n\beta_0^T (M^{-1}\Sigma M^{-1,T})^{-1}\beta_0. \tag{10}$$

To improve small sample performance, we use the F-distribution approximation as in the previous section. In particular, $\frac{n-q-p}{p(n-q-1)}T$ approximately follows $F_{p,n-q-p;\lambda_n}$, a non-central F-distribution with degrees of freedom $(p, n-q-p)$ and non-centrality parameter λ_n (Pan and Wall, 2002; Liao et al., 2016). In order to have at least $1-b$ power under H_1, we require

$$P\left\{T > \frac{p(n-q-1)}{n-q-p} F^{-1}_{p,n-q-p}(1-\gamma)\right\} \geq 1-b \quad \text{when} \quad \frac{n-q-p}{p(n-q-1)}T \sim F_{p,n-q-p;\lambda_n}.$$

Therefore, the required sample size is the smallest integer n such that

$$1 - F_{p,n-q-p;\lambda_n}\left\{F^{-1}_{p,n-q-p}(1-\gamma)\right\} \geq 1-b. \tag{11}$$

The sample size formula (11) relies on λ_n, which depends on the data generating distribution that is typically unknown during trial planning. In the following we simply M and Σ under a few working assumptions so that the formula can be directly used to determine the sample size n. Simulation studies to assess the performance of the sample size formula when the working assumptions are violated are reported in Section 5.

We make the following working assumptions.

(a) $\text{MEE}(t) = f(t)^T\beta_0$ for all $t = 1,\ldots,m$, where both $f(t)$ and $\beta_0 \in \mathbb{R}^p$ are known.

(b) Suppose $E(Y_{t+1} \mid A_t = 0, I_t = 1) = e^{g(t)^T\alpha_0}$ for all $t = 1,\ldots,m$, where both $g(t)$ and $\alpha_0 \in \mathbb{R}^q$ are known.

(c) Suppose $E\{r_t(\alpha',\beta_0)r_s(\alpha',\beta_0) \mid I_t=1,I_s=1,A_t,A_s\}$ is constant in A_t,A_s for all $1\le s<t\le m$. (Recall that α' is the probability limit of $\hat{\alpha}$.)

(d) Suppose $E(I_t)=\tau(t)$ for all $t=1,\ldots,m$, where $\tau(t)$ is known.

We show in Appendix [TQ: To Fill] that under working assumptions (a)-(c), M and Σ defined in (5) become

$$M=\sum_{t=1}^{m}\tau(t)e^{p_t f(t)^T\beta_0+g(t)^T\alpha_0}(1-p_t)p_t f(t)f(t)^T,$$
$$\Sigma=\sum_{t=1}^{m}\tau(t)e^{2p_t f(t)^T\beta_0+g(t)^T\alpha_0}(1-p_t)p_t\left[(1-p_t)e^{-f(t)^T\beta_0}+p_t-e^{g(t)^T\alpha_0}\right]f(t)f(t)^T. \quad (12)$$

Therefore, 1under these working assumptions, the required sample size n is calculated using (11), with λ_n defined in (10) and M,Σ calculated using (12). Inputs needed from the researcher to calculate the sample size are listed in Table 1.

Input	Interpretation
$1-b$	desired power
γ	desired type I error
m	total number of decision points per participant
p_t $(1\le t\le m)$	randomization probability at each decision point
$\tau(t)$ $(1\le t\le m)$	average availability at each decision point
β_0, $f(t)$ $(1\le t\le m)$	marginal excursion effect MEE(t) under H_1
α_0, $g(t)$ $(1\le t\le m)$	success probability null curve $E(Y_{t+1}\mid A_t=0,I_t=1)$

Table 1: Input to the sample size formula.

Working assumption (a) specifies the target alternative MEE curve under H_1. We include it as one of the working assumptions so as to assess in the simulation studies the performance of the sample size formula (in terms of, e.g., power) in settings where the true MEE curve is different than what is assumed under H_1.

Working assumption (b) states that the researcher knows $E(Y_{t+1}\mid A_t=0,I_t=1)$, the success probability of the binary outcome under no treatment and how the success probability changes over time. We will refer to $E(Y_{t+1}\mid A_t=0,I_t=1)$ as the success probability null curve. This may appear as a strong assumption, yet it is comparable to working assumption (c) in Liao et al. (2016) on homoscedasticity when the outcome is continuous. In our

binary outcome setting, homoscedasticity implies constant success probability, so one could argue that knowing the success probability null curve while allowing it to change over time is perhaps less restrictive. In the simulation studies we will see that correctly specifying the success probability null curve is critical for adequate power, and we will provide practical suggestions when the null curve is unknown.

Working assumption (c) concerns the delayed impact of prior treatments. $r_t(\alpha', \beta_0)$ is closely related to the blipping-down technique (Robins, 1994) and can be intuitively thought of as the outcome Y_{t+1} with the impact of the most immediate treatment (A_t) removed (also known as the blipped-down outcome). Therefore, one implication of working assumption (c) is that the blipped-down outcome at t is not impacted the prior treatment A_s. We show in Appendix [TQ: to fill] that a sufficient condition for working assumption (c) to hold is that (i) for $t > s$, $E(Y_{t+1} \mid I_t = 1, I_s = 1, A_t, A_s, Y_{s+1})$ only depends on A_t (no delayed effects and no serial dependence in the outcomes), and (ii) I_t is exogenous (does not depend on prior treatments or prior outcomes). In most mobile health applications, delayed effects and serial dependence are expected, so working assumption (c) is likely violated. We will assess the impact of such violations on the sample size formula performance use the simulation studies.

Working assumption (d) states that the researcher knows the average availability at each decision point. In the simulation studies, we will see the impact of misspecifying the average availability on power, and we provide practical suggestions for specifying $\tau(t)$.

5 Simulation

[TQ: work in progress]

6 Application

[TQ: work in progress]

7 Discussion

[TQ: work in progress]

References

Bell, L., Garnett, C., Qian, T., Perski, O., Potts, H. W., and Williamson, E. (2020). Notifications to improve engagement with an alcohol reduction app: Protocol for a micro-randomized trial. *JMIR research protocols*, 9(8):e18690.

Bidargaddi, N., Almirall, D., Murphy, S., Nahum-Shani, I., Kovalcik, M., Pituch, T., Maaieh, H., and Strecher, V. (2018). To prompt or not to prompt? a microrandomized trial of time-varying push notifications to increase proximal engagement with a mobile health app. *JMIR mHealth and uHealth*, 6(11).

Boruvka, A., Almirall, D., Witkiewitz, K., and Murphy, S. A. (2018). Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association*, 113(523):1112–1121.

Collins, L. M., Dziak, J. J., Kugler, K. C., and Trail, J. B. (2014). Factorial experiments: efficient tools for evaluation of intervention components. *American journal of preventive medicine*, 47(4):498–504.

Dempsey, W., Liao, P., Klasnja, P., Nahum-Shani, I., and Murphy, S. A. (2015). Randomised trials for the fitbit generation. *Significance*, 12(6):20–23.

Klasnja, P., Hekler, E. B., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., and Murphy, S. A. (2015). Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34(S):1220.

Klasnja, P., Rosenberg, D. E., Zhou, J., Anau, J., Gupta, A., and Arterburn, D. E. (2020). A quality-improvement optimization pilot of barifit, a mobile health intervention to promote physical activity after bariatric surgery. *Translational Behavioral Medicine*.

Liao, P., Klasnja, P., Tewari, A., and Murphy, S. A. (2016). Sample size calculations for micro-randomized trials in mhealth. *Statistics in medicine*, 35(12):1944–1971.

Mancl, L. A. and DeRouen, T. A. (2001). A covariance estimator for gee with improved small-sample properties. *Biometrics*, 57(1):126–134.

Pan, W. and Wall, M. M. (2002). Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in medicine*, 21(10):1429–1441.

Qian, T., Yoo, H., Klasnja, P., Almirall, D., and Murphy, S. A. (2019). Estimating time-varying causal excursion effect in mobile health with binary outcomes. *arXiv preprint arXiv:1906.00528*.

Rabbi, M., Kotov, M. P., Cunningham, R., Bonar, E. E., Nahum-Shani, I., Klasnja, P., Walton, M., and Murphy, S. (2018). Toward increasing engagement in substance use data collection: development of the substance abuse research assistant app and protocol for a microrandomized trial using adolescents and emerging adults. *JMIR research protocols*, 7(7).

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and methods*, 23(8):2379–2412.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Salkind, N. (2010). *Encyclopedia of Research Design*. SAGE Publications, Inc.

Tu, X. M., Kowalski, J., Zhang, J., Lynch, K., and Crits-Christoph, P. (2004). Power analyses for longitudinal trials and other clustered designs. *Statistics in medicine*, 23(18):2799–2815.